Optimizing Online Marketing Efficiency by Analyzing the Mutual Influence of Online Marketing Channels with Respect to Different Devices

Business Management Department
Universitat Politèchnica de València

A thesis for the degree of PhD in Business Administration

València, February 2019

Author

Ole Nass

Supervisors

José Albors Garrigós
Hermenegildo Gil Gómez
Klaus-Peter Schoeneberg

Abstract

What does attribution in an omni-channel environment look like? A major distinction can be determined in contrast to attribution in a multi-channel environment. Besides providing the Marketing Analytics Process, a specification of the Cross-industry standard process for data mining (CRISP-DM), a sequential mixed method approach is utilized to analyze the main research question.

Within the first step of this presented research characteristics, and requirements of efficient attribution in an omni-channel environment are analyzed. Based on semi-structured expert interviews and a holistic structured literature research process, the lack of an omni-channel attribution approach is clearly identified. Existing attribution approaches are identified by conducting the structured literature review process. Those identified approaches are evaluated by applying the results of the semi-structured expert interviews – the requirements and characteristics of efficient omni-channel attribution. None of the identified attribution approaches fulfill a majority of the analyzed omni-channel requirements.

By having the research gap – the lack of an omni-channel attribution approach – clearly identifed, an omni-channel attribution approach is developed in the second part of this presented research. Utilizing the MAP methodology, the main research gap is filled by providing the Holistic Customer Journey (HCJ): an omni-channel ready data foundation and a corresponding omni-channel attribution approach. Among other things the developed attribution approach consists of a machine learning classification. This presented research is the first to utilize information from almost 240.000.000 interaction data sets, containing cross-device and cross-platform information. All underlying data sources are provided by one of Germany's largest real-estate platforms.

Resumen

¿Cómo es la atribución en un entorno de omnicanal? Se puede determinar una distinción importante en contraste con la atribución en un entorno multicanal. Además de proporcionar el proceso de análisis de marketing, una especificación del proceso estándar intersectorial para la minería de datos (CRISP-DM), se utiliza un enfoque de método mixto secuencial para analizar la cuestión principal de la investigación.

En el primer paso de esta investigación se analizan las características y los requisitos de atribución eficiente en un entorno omnicanal. A partir de entrevistas semiestructuradas con expertos y de un proceso de investigación bibliográfica holística estructurada, se identifica claramente la falta de un enfoque de atribución omnicanal. Los enfoques de atribución existentes se identifican mediante la realización de un proceso estructurado de revisión de la literatura. Estos enfoques identificados se evalúan aplicando los resultados de las entrevistas semiestructuradas con expertos, es decir, los requisitos y características de una atribución omnicanal eficiente. Ninguno de los enfoques de atribución identificados cumple con la mayoría de los requisitos de omnicanal analizados.

Al tener la brecha de investigación - la falta de un enfoque de atribución de omnicanales - claramente identificada, se desarrolla un enfoque de atribución de omnicanales en la segunda parte de esta investigación presentada. Utilizando la metodología MAP, la principal laguna de investigación se llena proporcionando el Holistic Customer Journey (HCJ): una base de datos lista para el omni-canal y un enfoque de atribución de omni-canal correspondiente. Entre otras cosas, el enfoque de atribución desarrollado consiste en una clasificación de aprendizaje automático. Esta investigación presentada es la primera en utilizar información de casi 240.000.000 de conjuntos de datos de interacción, que contienen información entre dispositivos y entre plataformas. Todas las fuentes de datos subyacentes son proporcionadas por una de las plataformas inmobiliarias más grandes de Alemania.

Resum

Com és l'atribució en un entorn de omnicanal? Es pot determinar una distinció important en contrast amb l'atribució en un entorn multicanal. A més de proporcionar el procés d'anàlisi de màrqueting, una especificació del procés estàndard intersectorial per a la mineria de dades (CRISP-DM), s'utilitza un enfocament de mètode mixt seqüencial per analitzar la qüestió principal de la investigació.

En el primer pas d'aquesta investigació s'analitzen les característiques i els requisits d'atribució eficient en un entorn omnicanal. A partir d'entrevistes semiestructurades amb experts i d'un procés de recerca bibliogràfica holística estructurada, s'identifica clarament la falta d'un enfocament d'atribució omnicanal. Els enfocaments d'atribució existents s'identifiquen mitjançant la realització d'un procés estructurat de revisió de la literatura. Aquests enfocaments identificats s'avaluen aplicant els resultats de les entrevistes semiestructurades amb experts, és a dir, els requisits i característiques d'una atribució omnicanal eficient. Cap dels enfocaments d'atribució identificats compleix amb la majoria dels requisits de omnicanal analitzats.

En tenir la bretxa de recerca - la manca d'un enfocament d'atribució de omnicanales - clarament identificada, es desenvolupa un enfocament d'atribució de omnicanales a la segona part d'aquesta investigació presentada. Utilitzant la metodologia MAP, la principal llacuna de recerca s'omple proporcionant el Holistic Customer Journey (HCJ): una base de dades a punt per al omni-canal i un enfocament d'atribució de omni-canal corresponent. Entre altres coses, l'enfocament d'atribució desenvolupat consisteix en una classificació d'aprenentatge automàtic. Aquesta investigació presentada és la primera a utilitzar informació de gairebé 240.000.000 de conjunts de dades d'interacció, que contenen informació entre dispositius i entre plataformes. Totes les fonts de dades subjacents són proporcionades per una de les plataformes immobiliàries més grans d'Alemanya.

Table of Contents

Abstract ... ii

Resumen ... iii

Resum ... iv

Table of Contents ... v

List of Tables ... x

List of Figures ... xii

List of Abbreviations ... xiii

List of Appendices ... xvi

Acknowledgements ... xviii

1 Introduction ... 1

1.1 Leading to the Topic ... 1

1.2 Objectives and Contributions to the Scientific Community ... 3

1.3 Organization of the Thesis ... 4

2 Methodology and Research Framework ... 5

2.1 Justifying the Relevance of the Main Research Question for Science and Practice ... 5

2.2 Research Framework - Course of the Examination ... 6

2.2.1 Structured Literature Review ... 9

2.2.2 Exploratory Sequential Mixed-Method Approach ... 9

2.2.2.1 Qualitative Analysis ... 10

2.2.2.2 Quantitative Analysis ... 12

2.3 Justifying the Hypotheses ... 16

2.3.1 Hypothesis 1 ... 17

2.3.2 Hypothesis 2 ... 17

2.3.3 Hypothesis 3 ... 17

2.3.4 Hypothesis 4 ... 18

3 Summary of the Marketing Analysis Process (MAP) Methodology ... 19

4 **Publication 1** MAP - Marketing Analytics Process ... 22

Einleitung ... 23

Marketing-Analytics-Process-(MAP)-Modell ... 24

Phase I: Problemidentifikation/Zieldefinition 25
Exkurs: Analysen im (Online-)Marketing-Kontext 26
Praxis Use-Case 27
Phase II: Datenquellenauswahl 29
Praxis Use-Case 30
Phase III: Datenaufbereitung 34
Praxis Use-Case 35
Phase IV: Modellierung 40
Praxis Use-Case 41
Phase V: Modellevaluierung 42
Praxis Use-Case 43
Phase VI: Handlungsempfehlungen 44
Praxis Use-Case 44
Implementierung 45
Fazit 46
Literatur 47
5 **Publication 2** Attribution Modelling in an Omni-Channel Environment - New Requirements and Specifications from a Practical Perspective 50
Introduction 51
Theoretical Background 51
Deriving Towards Omni-Channel Marketing 52
State of the Art 55
Channel Performance 55
Customer Satisfaction 56
Dynamic Attribution Approaches 56
Hypotheses 60
Research Methodology 61
Expert Sampling 62
Interview Guideline 63
Conducting the Semi-Structured Expert Interviews 64
Evaluation Methodology 64

Results 66

Identify Dynamic Attribution Approaches 66

Evaluation Criteria for Attribution Approaches in an Omni-Channel Environment 69

Evaluation of Existing Dynamic Attribution Models Towards their Applicability in an Omni-Channel Environment 73

Discussion 76

Critical Examination of the Results 78

Future Research Fields and Research Questions from a Practical Perspective 79

Conclusion 82

References 83

6 Developing an Omni-Channel Ready Data Foundation 91

6.1 Identifier Matching 91

6.2 Data Transformation Process 92

6.2.1 Workflow [01]: GOOGLE_ANALYTICS_STAGE_2_CORE 95

6.2.2 Workflow [01A]: GOOGLE_ANALYTICS_HITS_CUSTOMDIM_CORE 96

6.2.3 Workflow [01B]: GOOGLE_ANALYTICS_HITS_CORE 97

6.2.4 Workflow [02]: EVENTSTORE_STAGE_2_CORE 97

6.2.5 Workflow [03]: PRICING_DATA_STAGE_2_CORE 98

6.2.6 Workflow [04]: INTELLIAD_CLICK_REPORT_STAGE_2_CORE 99

6.2.7 Workflow [04A]: INTELLIAD_CLICK_REPORT_PRICE_CORE 99

6.2.8 Workflow [05]: PRODUCT_PRICES_STAGE_2_CORE 100

6.2.9 Workflow [06]: IMPORT_ONE_DAY_2_CORE 100

6.2.10 Workflow [07]: CREATE_HOLISTIC_CUSTOMER_JOURNEY_OF_ONE_DAY 101

6.3 Feature Generation 101

6.4 Descriptive Statistics of the HCJ 106

7 **Publication 3** Ready for Omni-Channel: Cross Device and Cross Platform Machine Learning Attribution Approach – A Field Experiment 111

Introduction and Objectives 112

Theoretical Background 112

Towards Attribution in an Omni-Channel Environment 113

Importance of Machine Learning for Marketing in Businesses 114

Define Objective 115

Research Methodology 116
Data Foundation 119
Collecting Data 122
Cross-Device / Cross-Platform 122
The HCJ Data Foundation in Numbers 122
HCJ Summarized 123
Feature Generation 124
Target Specification 124
Attribution Model 126
Model Requirements 126
Prepare Features and Target 128
Dimension Reduction – Increasing Performance 129
Define Classification Metric 130
Data Split 130
Identify Best Classifying Algorithm 131
Identify Best Hyperparameter Configuration 132
Model Accuracy 134
Results and Discussion 134
Recommendations for Practitioners 138
Attribution Model Extension and Integration 138
Conclusion 139
References 140
8 Results 148
8.1 Marketing Analytics Process (MAP) 148
8.2 What does Efficient Attribution in an Omni-Channel Environment Look Like? 148
8.2.1 Verifying Hypotheses H1 and H2 148
8.2.2 Verifying Hypotheses H3 and H4 149
9 Conclusion 150
9.1 Summary of the Investigation 150
9.2 What does Efficient Attribution in an Omni-Channel Environment Look Like? 150
9.3 Critical Appraisal, Limitations and Opportunities for Further Research 151

9.4 Contribution to Knowledge 153
9.4.1 Implications for Theory 154
9.4.2 Implications for Practitioners 155
10 Publication bibliography 156
Appendices I

List of Tables

Table 1: Contribution of the current research towards the science community 3

Table 2: Dimensions of a mixed method design based on Creswell et al. (2003) 10

Table 3: Hypothesis and applied methodologies 16

Table 4: Phases of the Marketing Analytics Process 20

Table 5: Projektplan inkl. Zeiteinschätzung 29

Table 6: Selektierte Datenfelder aus Google BigQuery 33

Table 7: Selektierte Datenfelder aus der Objekt-Datenbank.. 33

Table 8: Selektion aller Neukunden des Monats März 36

Table 9: Datenfelder der verknüpften Tracking- und Objektdaten 39

Table 10: Structured literature process: identified publications based on concept 1 and concept 2 59

Table 11: Formulated hypotheses for the current research 60

Table 12: List of interviews consisting of an Identifier, the assigned expert group and the field of work of the interviewee. 62

Table 13: Dynamic Attribution Approaches in Science: Results of the structured literature research process 66

Table 14: Results: Evaluation criteria for attribution models in an omni-channel environment from a practical perspective. Each category is sorted in descending order of importance. 70

Table 15: Evaluation of the identified attribution models towards identified requirements and specifications from a practical perspective... 74

Table 16: Proposed research agenda for further research in the field of dynamic attribution modelling from a practical perspective 79

Table 17: Generated features for the machine learning approach 102

Table 18: Description of the transformation process towards the HCJ............................... 108

Table 19: Descriptive statistic values and distribution of selected features 110

Table 20: Formulated hypotheses for the current research.. 115

Table 21: Conceived data requirements for the current analysis.. 120

Table 22: Used data source, loading frequency and access type 121

Table 23: Conversion probability cases.. 125

Table 24: Conceived model requirements for the current analysis..................................... 127

Table 25: Accuracy and trainings duration of the applied algorithms Random Forest, Extra Tree, Ada Boost and Gradient Boost 132

Table 26: Hyperparameters to tune with value ranges 133

Table 27: Verification of the implementation of derived data requirements for the current analysis 135

List of Figures

Figure 1: Amount of publications listed in the Web of Science on August 30th, 2018. 2

Figure 2: Research Framework of the investigation 7

Figure 3: Research Framework of the investigation (cont.).. 8

Figure 4: Marketing-Analytics-Process Modell (MAP). 25

Figure 5: Marketing-Mix-Framework im Rahmen des Big Data-Managements. 27

Figure 6: Technische Implementierung des Trackings auf immonet.de, inkl. Datenflusses.... 31

Figure 7: Ergebnis der vier Segmente. 42

Figure 8: Ergebnis der vier Segmente mit bereinigter Datengrundlage. 43

Figure 9: Data flow in a multi-channel and an omni-channel environment. Based on Paccard (2017).. 52

Figure 10: Delimitation: Static, Simplistic, Rule-based, Dynamic and Algorithmic attribution models.. 54

Figure 11: Guideline for the expert interviews: Main- and sub-categories 64

Figure 12: Top-down and bottom-up approach - import setup for all data sources divided up into workflows across all stages... 93

Figure 13: HCJ: Top-down and bottom-up approach - import setup for all data sources divided up into workflows across all stages (cont.) 94

Figure 14: Linkage of all utilized data sources 122

Figure 15: Plot of the first two principal components (PC). The color distinguishes between the two categories: invest / not invest. 128

Figure 16: Cumulative variance of all principal components (PC) 129

Figure 17: Data split into training set, validation set and test set 130

Figure 18: Accuracy of the applied algorithms Random Forest, Extra Tree, Ada Boost, and Gradient Boost 131

List of Abbreviations

Ad	Advertisement
AI	Artificial Intelligence
(mobile) App	(mobile) Application
BI	Business Intelligence
Cat.	Category
CLS	Cleanse
CLT	Customer-Life-Time
CO	Core
Cont.	Continue
cpc	Costs Per Click
CRISP-DM	Cross-Industry Standard Process for Data Mining
CRM	Customer Relationship Management
Csv	Comma Separated Value
DF	Data Flow
DMP	Data Management Platform
DOI	Digital Object Identifier
DR	Data requirement
DWH	Data Warehouse
e.g.	exempli gratia
Eng.	English
et al.	et alii
ELT	extract, load, transform (process)
ETL	extract, transform, load (process)
etc.	et cetera

GUA	Google Universal Analytics
H#	Hypothesis #
HCJ	Holisitc Customer Journey
HMM	Hidden Markov model
ICD	International Data Corporation
Id	Identifier
Inc.	Incorporated
IPO	input-process-output
ipynb	IPython Notebook
IT	Internet Technology
JSON	Javascript Object Notation
MAP	Marketing Analytics Process
Max	Maximum
MF/S	Model Feature / Specification
Min	Minimum
ML	Machine Learning
MSI	Marketing Science Institute
n	Sample Size
OR	Other requirement
p	Page
pp	Pages
qual	Qualitative
QUANT	Quantitative
ROI	Return of Investment
SEA	Search Engine Advertisement

SEO	Search Engine Optimization
sql	structured query language
STD	standard deviation
STG	stage
TB	Terabyte (1 TB = 1024 Gigabyte)
WF	Workflow
WoS	Web of Science

List of Appendices

Appendix 1: Structured literature review process I

Appendix 2: Interview Coding Categories XIII

Appendix 3: Questionnaire Guideline XXII

Appendix 4: Java-Script: ID-Matching XXIII

Appendix 5: Transformation performance XXV

Appendix 6: WF[01]_stage_run XXVII

Appendix 7: WF[01]_cleanse_run XXVIII

Appendix 8: WF[01]_core_create XL

Appendix 9: WF[01]_core_run XLV

Appendix 10: WF[01A]_core_create L

Appendix 11: WF[01A]_core_run LI

Appendix 12: WF[01B]_core_create LII

Appendix 13: WF[01B]_core_run LIII

Appendix 14: WF[02]_staging_run LIV

Appendix 15: WF[02]_cleanse_run LV

Appendix 16: WF[02]_core_run01 LVI

Appendix 17: WF[02]_core_run02 LVII

Appendix 18: WF[03]_stage_create LVIII

Appendix 19: WF[03]_stage_run LIX

Appendix 20: WF[03]_cleanse_run01 LX

Appendix 21: WF[03]_cleanse_run02 LXI

Appendix 22: WF[03]_core_run LXII

Appendix 23: WF[04]_stage_run LXIII

Appendix 24: WF[04]_cleanse_run LXIV

Appendix 25: WF[04]_core_run LXV

Appendix 26: WF[04A]_core_run.. LXVI

Appendix 27: WF[05]_stage_run LXVII

Appendix 28: WF[05]_cleanse_run LXVIII

Appendix 29: WF[05]_core_run LXIX

Appendix 30: WF[07]_core_run01 LXX

Appendix 31: WF[07]_core_run02 LXXII

Appendix 32: Feature generation sql script LXXIV

Appendix 33: Descriptive statistic values and distribution of features LXXXIII

Appendix 34: Python: Jupyter notebook LXXXVI

Appendix 35: Performance of different hyperparameter combinations............................XCVIII

Acknowledgements

There are a couple of people whom I would like to thank:

Thank you to my awesome supervisors who guided, challenged, and supported me during the whole research process: Prof. Jose Albors Garrigos Ph.D. and Prof. Hermenegildo Gil Gómez Ph.D. of the Universitat Politècnica de València, Spain. Thank you for your guidance, comments and inspirational dialogues. I especially thank Prof. Dr. Klaus-Peter Schoeneberg. Klaus-Peter, thank you for offering me the opportunity to do this research. Your support, feedback and comments were invaluable for me. I'm so grateful for this inspiring time and our resulting friendship.

Special thanks go to Dr. Katrin Blankenburg and the team of the Promotionszentrum of the HAW Hamburg. Their valuable feedback, advice, critique, as well as the well-organized and conceived workshops offered were immensely helpful.

I also thank my employer and my working colleagues for their inspiring input and support.

I would like to express my gratitude to my family. I thank my wife Isabel and my children Nele and Jorne for their support, love and time. I also thank my parents Volker and Dörte for their time and support. Without all of you, this research would not have been possible.

I dedicate this work to my grandmother Käthe, who at the proud age of 97 has supported and inspired me since the beginning of this work. Thank you, you have convinced me during our many phone calls and personal conversations to do this research and encouraged me again and again. Thank you!

1 Introduction

1.1 Leading to the Topic

New technologies enable companies or institutions to track a user on his way to purchase in a very granular and detailed way. Important trace data is not generated on a company's website alone. Trace data can be generated in different mobile applications (apps) or other online or offline marketing channels provided by a company. Online marketing channels such as social media, mailings, paid search or display advertisement generate channel-specific usage data separately. Any other third-party vendor connected to the provided eco-system of a company or institution individually produces data as well. Furthermore, data is generated if a customer interacts with a helpdesk or while using the hotline. Offering different independent so-called *channels* to communicate and interact with one's customer is a widely applied strategic approach (Econsultancy 2015). This strategy is termed a multi-channel approach. In such a multi-channel environment, responsible channel marketers act as independent departments mainly using their own, within the channel generated data (Neslin et al. 2006). Information about the user (data) is not shared across different departments or channels. Acting as independent channels, important information about a user gathered by other departments is neglected.

Having some information exchanged between departments and channels, e.g. sending a coupon via email to buy a product in an online store is called a cross-channel strategy. A pursued user action such as purchase, or a newsletter signup is achievable a*cross* different channels.

The next more advanced strategic approach to communicating with one's customers is termed omni-channel strategy (Camiade 2013). The term 'omni'-channel (lat. omnis), can be translated as 'all'. An omni-channel strategy, opposed to a multi-channel strategy, enables a seamless user experience across different channels (Lazaris and Vrechopoulos 2014; Levy et al. 2014). From a data-driven perspective a major change can be determined regarding an omni-channel strategy. A centralized data hub containing or connecting decentralized data sources which are generated from different (marketing-) channels needs to present. Figure 9 inspired by Paccard (2017) illustrates the different setups.

Today, there is a widely spread shift towards an omni-channel setup. Verhoef et al. (2015) describe the necessary shift towards an omni-channel setup within a retailing context. To achieve the pursued seamless user experience within an omni-channel environment enforces data management and data-driven decisions within the marketing department(s) and other departments interacting with the customer.

Facing new challenges in an omni-channel environment is relatively new. The Web of Science is utilized by applying the two search terms “multi-channel marketing” and “omni-channel marketing” for a comparison on the amount of available publications. Figure 1 illustrates the amount of publications for each year. The first research on “omni-channel marketing” is available in 2014. There is a positive trend since 2014. In August 2018 there are already more publications on “omni-channel marketing” than on “multi-channel marketing”.

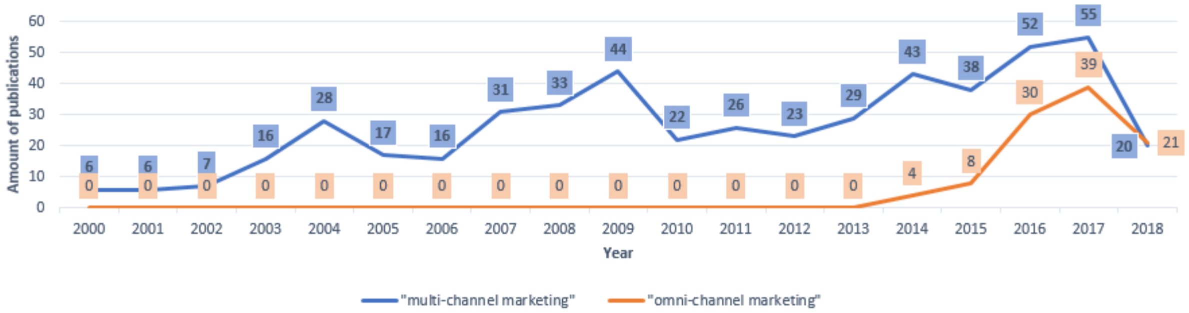


Figure 1: Amount of publications listed in the Web of Science on August 30th, 2018.

New challenges, questions and research areas arise in this field. The Marketing Science Institute analyzed and identified different research priorities for the years 2016 to 2018 (MSI 2016). Most of the identified research priorities arise because the shift towards an omni-channel strategy brings along new challenges requiring modern solutions and new approaches.

The presented research focuses on the general question of what attribution in an omni-channel environment should look like.

Defining Attribution Modelling

Generally speaking, attribution modelling is the definition of how much impact or value a touch-point within a provided channel consists of onto a predefined action. (Nottorf 2014; Li and Kannan 2014). Exemplary of such actions are: a purchase, the generating of a lead, or any other event providing a company income. The company tries to encourage the user to perform such actions.

In the following, a definition from a scientific publication and a definition with a practical perspective are presented to sharpen the term *attribution modelling*.

“Attribution modeling is the practice of mapping touchpoints to monetarily relevant events within a customer journey, which is directly related to the return on investment of e-commerce websites, campaigns, or rankings in the organic search results.” (Ryte 2016)

"The attribution problem […] measures the relative effectiveness of channels in a given setting [(Li and Kannan 2014)], so the results are conditional upon a number of management decisions such as channels used or budget limits per channel. Therefore, optimizing the budget allocation remains an iterative process. Correct attribution, however, is a necessary prerequisite for managers to optimize their budget decisions." (Anderl et al. 2016a)

1.2 Objectives and Contributions to the Scientific Community

The objective of the current research is to identify and formulate what efficient attribution in an omni-channel environment looks like.

The presented research consists of three independent publications contributing to the academic science community. Table 1 lists the scientific contributions of each publication.

Table 1: Contribution of the current research towards the science community

Publication		**Contribution**
1	MAP- Marketing Analytics Process	• A specification of the Cross-Industry Standard for Data Mining (CRISP-DM) process for (online-) marketing specific bigdata problems, the MAP methodology.
2	Attribution modelling in an omni-channel environment new requirements and specifications from a practical perspective	• Identification of existing dynamic attribution models in the science community based on a structured literature research process. • Present criteria (requirements and specifications) for dynamic attribution models in an omni-channel environment based on expert interviews. • Evaluation of the identified models based on the resulting criteria from the expert interviews. • Formulate and define exigencies, research fields and research questions for further research.
3	Ready for Omni-Channel: Cross Device and Cross Platform Machine Learning Attribution Approach – A Field Experiment	• An omni-channel ready attribution approach trained onto a cross-device and cross-platform data foundation. • Proof of practicability of the implantation of the pre-identified requirements and specification for efficient attribution in an omni-channel environment.

1.3 Organization of the Thesis

The presented research is enclosed in a cumulative thesis consisting of three individual publications. The entire research is structured in the following way:

The introduction outlines the relevance and a dedication towards the general research field of omni-channel marketing. Furthermore, a definition of the research topic *attribution modelling* is presented, followed by the introduction of the main objective, the contributions of the current research towards the scientific community and the organization of the thesis.

In the second chapter, the applied methodology and the research framework are presented. The first publication, presented in chapter 4, is originally published in German. The third chapter consists of an English summary of the first publication. Within the first publication the Marketing Analytic Process (MAP) methodology is presented. The MAP methodology is utilized in this research.

The first and the second publication directly follow each other. A connecting chapter between the two publications is not needed.

The presented research utilizes a mixed-method approach. The first qualitative analysis enclosed in publication two is presented in chapter 5. Within the qualitative analysis, requirements and specifications for attribution modelling in an omni-channel environment are identified.

Based on the identified requirements, the quantitative analysis is conducted. Chapter 6 connects the research between publication two and publication three. In this chapter the first part of the quantitative analysis – the development of an omni-channel ready data foundation is described. The following chapter 7 contains publication three. In publication three the development of the attribution approach including the machine learning approach is described.

In chapter 8 the results are presented and later discussed in chapter 9. Also, the limitations of the current research and the implications for theory and practitioners are presented in chapter 9.

All references including the references of the three publications are listed in chapter 10 followed by the appendices.

2 Methodology and Research Framework

Empirical research is defined by Früh (2015) as a systematic, intersubjective verifiable collection control and criticism of experiences. According to Früh, an idea or a research question needs to be formulated at the beginning of the research.

For the current research the following main research question is formulated:

“What does efficient attribution in an omni-channel environment look like?”

2.1 Justifying the Relevance of the Main Research Question for Science and Practice

A research question must describe a theoretical knowledge gap. “[…] eine Forschungsfrage [muss] also eine theoretische Wissenslücke beschreiben.” (Gläser and Laudel 2010). The research gap needs to be identified and described.

According to Ulrich (1995) the applied research designs begins with practical problems which are unresolved. Those problems are analyzed by utilizing available literature and theories. His scientific approach conceives business studies as part of the application-orientated social science, which does not act in a static way, but considers change and alteration as instruments for creating design concepts of the future social reality (Ulrich 1981). Ulrich claims that business studies understood as applied science should be adjusted by problems related to the practice of corporation management (Ulrich et al. 1976; Ulrich 1981, 1985). By also considering the aforementioned research priority defined by the MSI (MSI 2016), the attribution problem within the presented research is both, a scientific problem as well as a practice relevant problem which enables the course of the current examination.

The main research question is placed in the research areas of marketing, bigdata analytics and computer science. *Attribution* itself belongs to the number one research priority identified by the Marketing Science Institute for the years 2016 to 2018 (MSI 2016). The most important research priority is defined as “Quantitative models to understand causality, leavers, and influence in a complex word” (MSI 2016). The relevance of the main research question for science is outlined. To identify the research gap, the main research question needs to be analyzed further.

The question of how much a customer is currently worth to a company is complex and difficult to determine. The answer to this question is of great interest for companies, firms or other institutions (Nottorf 2014; Anderl et al. 2016a). Therefore, the main research question is relevant for both, science and practice.

2.2 Research Framework - Course of the Examination

After having the main research question justified, the course of the examination must be defined prior to the investigation (Kuckartz 2014; Creswell 2014). This includes the formulation of the hypotheses guiding the research process. Figure 2 and Figure 3 illustrate the applied research framework for the investigation which is explained afterwards in chapter 2.2.2. The notation for mixed-method approaches developed by Morse (1991) is applied.

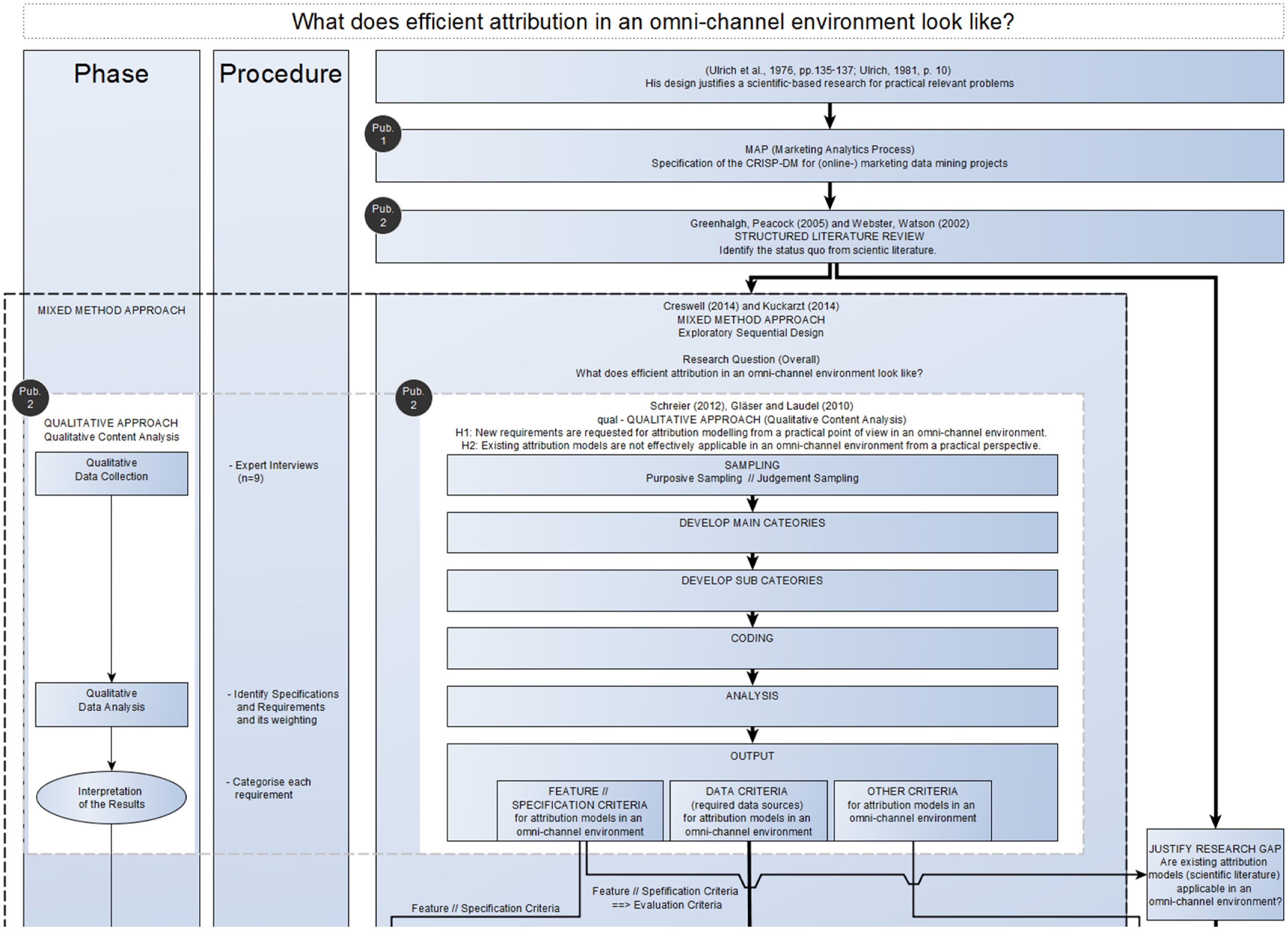

What does efficient attribution in an omni-channel environment look like?
Phase
Procedure
(Ulrich et al., 1976, pp.135-137; Ulrich, 1981, p. 10)
His design justifies a scientific-based research for practical relevant problems
Pub. 1
MAP (Marketing Analytics Process)
Specification of the CRISP-DM for (online-) marketing data mining projects
Pub. 2
Greenhalgh, Peacock (2005) and Webster, Watson (2002)
STRUCTURED LITERATURE REVIEW
Identify the status quo from scientic literature.
MIXED METHOD APPROACH
Creswell (2014) and Kuckarzt (2014)
MIXED METHOD APPROACH
Exploratory Sequential Design
Research Question (Overall)
What does efficient attribution in an omni-channel environment look like?
Pub. 2
QUALITATIVE APPROACH
Qualitative Content Analysis
Qualitative Data Collection
- Expert Interviews (n=9)
Pub. 2
Schreier (2012), Gläser and Laudel (2010)
qual - QUALITATIVE APPROACH (Qualitative Content Analysis)
H1: New requirements are requested for attribution modelling from a practical point of view in an omni-channel environment.
H2: Existing attribution models are not effectively applicable in an omni-channel environment from a practical perspective.
SAMPLING
Purposive Sampling // Judgement Sampling
DEVELOP MAIN CATEORIES
DEVELOP SUB CATEORIES
CODING
Qualitative Data Analysis
- Identify Specifications and Requirements and its weighting
ANALYSIS
OUTPUT
Interpretation of the Results
- Categorise each requirement
FEATURE // SPECIFICATION CRITERIA for attribution models in an omni-channel environment
DATA CRITERIA (required data sources) for attribution models in an omni-channel environment
OTHER CRITERIA for attribution models in an omni-channel environment
JUSTIFY RESEARCH GAP
Are existing attribution models (scientific literature) applicable in an omni-channel environment?
Feature // Spefification Criteria
==> Evaluation Criteria
Feature // Specification Criteria

Figure 2: Research Framework of the investigation

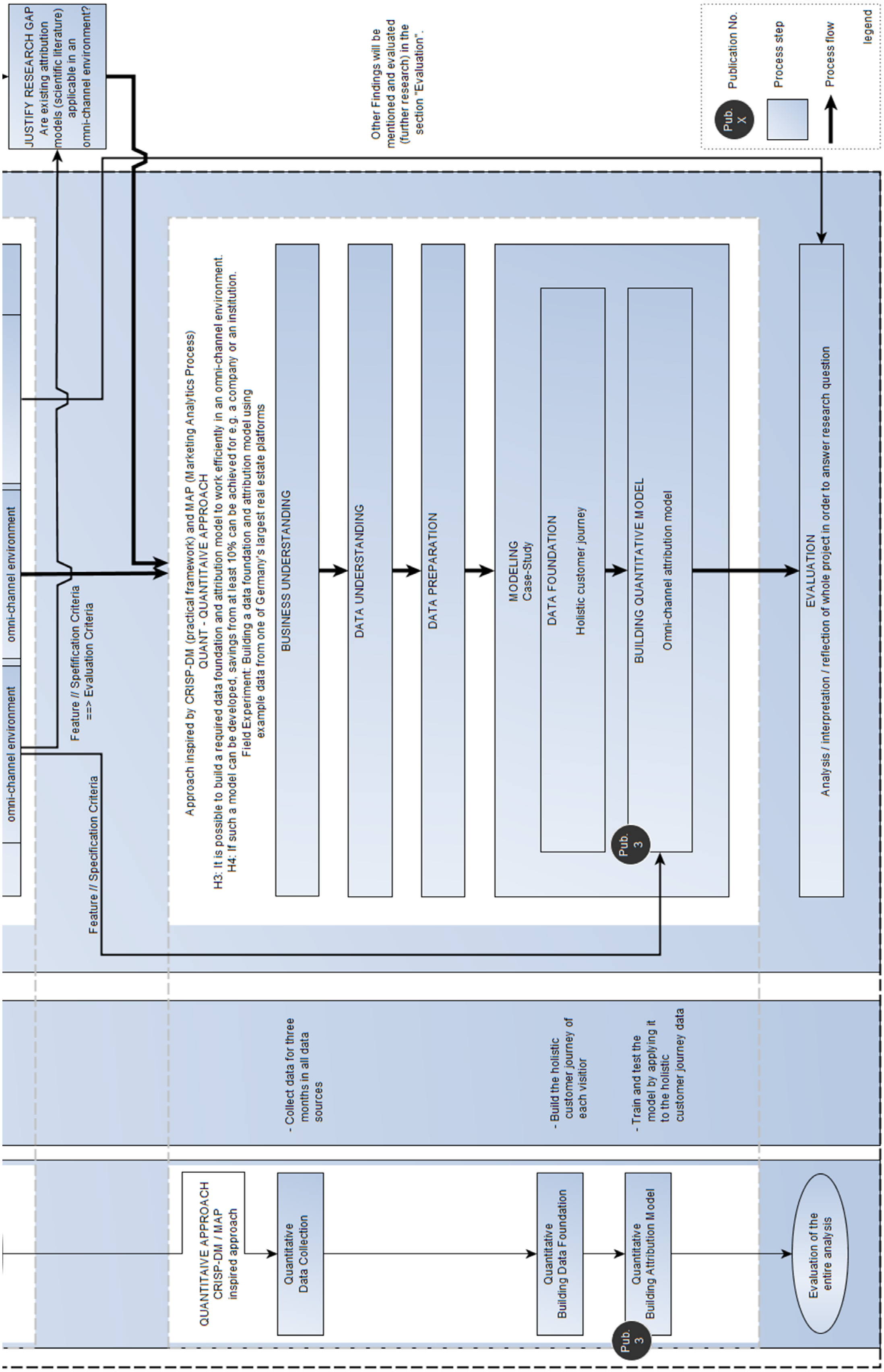

omni-channel environment
omni-channel environment
JUSTIFY RESEARCH GAP
Are existing attribution models (scientific literature) applicable in an omni-channel environment?
Feature // Speffication Criteria
==> Evaluation Criteria
Feature // Specification Criteria
Other Findings will be mentioned and evaluated (further research) in the section "Evaluation".
Approach inspired by CRISP-DM (practical framework) and MAP (Marketing Analytics Process)
QUANT - QUANTITAIVE APPROACH
H3: It is possible to build a required data foundation and attribution model to work efficiently in an omni-channel environment.
H4: If such a model can be developed, savings from at least 10% can be achieved for e.g. a company or an institution.
Field Experiment: Building a data foundation and attribution model using example data from one of Germany's largest real estate platforms
BUSINESS UNDERSTANDING
DATA UNDERSTANDING
DATA PREPARATION
MODELING
Case-Study
DATA FOUNDATION
Holistic customer journey
Pub. 3
BUILDING QUANTITATIVE MODEL
Omni-channel attribution model
EVALUATION
Analysis / interpretation / reflection of whole project in order to answer research question
- Collect data for three months in all data sources
- Build the holistic customer journey of each visitior
- Train and test the model by applying it to the holistic customer journey data
QUANTITAIVE APPROACH
CRISP-DM / MAP
inspired approach
Quantitative
Data Collection
Quantitative
Building Data Foundation
Pub. 3
Quantitative
Building Attribution Model
Evaluation of the entire analysis
Pub. X
Publication No.
Process step
Process flow
legend

Figure 3: Research Framework of the investigation (cont.)

Prior to the main analysis, the *MAP – Marketing Analytics-Process*, an extension of the Cross Industry Standard Process for Data Mining (CRISP-DM) (Shearer 2000) is developed as a research methodology for practical problems. As described above, the methodology of Ulrich (1995) enables consideration of a practical problem to solve it with scientific approaches. This methodology is constructed as a guide for practical problems to find solutions to (online-) marketing bigdata problems. Next to the CRISP-DM, the MAP methodology is applied in the QUANT analysis of the mixed-method design to guide the investigation.

2.2.1 Structured Literature Review

Each research must be placed in a scientific context by filling a pre-identified research gap. (Creswell 2014; Kuckartz 2014).

At the beginning of the presented research a structured literature research is conducted inspired by Greenhalgh and Peacock (2005) and Webster and Watson (2002), identifying existing dynamic attribution approaches to state the status quo of research in the field of attribution.

To the best of the author's knowledge, there is no publication dealing either with the topic of comparing dynamic attribution models or evaluating them with respect to omni-channel requirements. There is only one related paper in which the authors classify dynamic attribution models from a statistical perspective (Jayawardane et al. 2015). A structured and comprehensive analysis is not within the scope of their article. Based on seven identifiable model features, a classification of the statistical approach within a model has been analyzed. Their paper concentrates on the mathematical and statistical approach.

The process or the structured literature research process is attached in Appendix 1. All identified attribution approaches are verified towards their applicability in an omni-channel environment within the first publication (see chapter 5). By analyzing the applicability, the research gap – the lack of an omni-channel attribution approach – is clearly identified as shown in the research framework (see Figure 2 and Figure 3).

2.2.2 Exploratory Sequential Mixed-Method Approach

The main investigation is inspired by the exploratory sequential mixed method design (Creswell 2014; Kuckartz 2014). This design is utilized to analyze the main research question.

Creswell (2014) requires considering the four criteria when deciding on a mixed method design. These criteria are listed in Table 2. Within this table the attributes of the presented research are highlighted in bold.

Table 2: Dimensions of a mixed method design based on Creswell et al. (2003)

Implementation	Priority	Integration	Theoretical Perspective
• **No order** • **Sequential: qualitative first** • Sequential: quantitative first	• Equivalent • qualitative • **quantitative**	• during data collection • during data analysis • **during data interpretation** • multiple times	• **explicit** • implicit

The highlighted criteria characterize a "qualitativ-vertiefendes Design" (Kuckartz 2014) or "Vertiefungsdesign" (Mayring 2001). In contrast to the definition of Creswell et al. (2003), the current design prioritizes the quantitative analysis and not the qualitative analysis. Firstly, this presented research identifies criteria for attribution in an omni-channel environment by conducting semi-structured expert interviews. The research focusses on the subsequent analysis in the second step. In this step, in a field experiment, the identified criteria are analyzed according to their feasibility in the *real-world* utilizing the provided data sources by one of Germany's largest real-estate platforms.

A parallel mixed method design is not applicable because the results of the first analysis are required for the second analysis (Creswell 2014).

According to Kuckartz (2014), a mixed-method design is chosen because the research question is too complex to be answered with only a qualitative approach or a quantitative approach. A mixed-method methodology enables a better understanding of such complex research questions (Kuckartz 2014). With an only qualitative or only quantitative methodological approach the current investigation could not be implemented.

The qualitative analysis is following a deductive approach, whereas the quantitative analysis is executed in an inductive way.

2.2.2.1 Qualitative Analysis

The first analysis, the exploratory sequential inspired design, consists of a qualitative analysis (qual) identifying requirements and specifications towards attribution modelling in an omni-channel environment. Such requirements are identified by conducting semi-structured expert interviews (Gläser and Laudel 2010) and applying the *Qualitative Content Analysis* by Schreier (2012) and Gläser and Laudel (2010) for the evaluation process. This qual analysis is guided by the first two hypothesis H1 and H2 listed in Table 3. All interviews are guided by the guideline attached in Appendix 3. The justification of the hypotheses is described within the corresponding publication. H1 and H2 are discussed and analyzed in publication two in chapter 5. The latter hypotheses H3 and H4 are analyzed in publication three in chapter 7.

To ensure an appropriate degree of quality of the research, the main criteria – objectivity, reliability and validity (Amelang et al. 2004) – are utilized for both the qual analysis and the QUANT analysis.

A qualitative approach is chosen since the topic of investigation is relatively new and the access to experts in this field is limited. All experts are selected based on pre-defined criteria to ensure high quality input and reproducibility. The semi-structured interviews are executed based on a guideline (see Appendix 3). Since it is the target to identify criteria for attribution in an omni-channel context, feelings and attitudes of the experts are not relevant. Only requirements and specifications are relevant for the presented research. No subjective selection of the data has occurred. Interviewing the same experts with the same guideline results in the same collected data. This ensures the objectivity of the qual analysis.

The process of coding, selecting and prioritizing the evaluation criteria is inspired by the *Qualitative Content Analysis* by Schreier (2012) and Gläser and Laudel (2010). Since the selection and prioritization are based on the amount of mentions and the evaluation of the experts, the results are reliable.

External stimuli were kept low during the interviews in order not to influence the focus on the topic. During the qualitative analysis, the expected information was collected to ensure the validity of the research.

The results of the qual analysis and the structured literature research are included in the second publication of this research (see chapter 5).

2.2.2.1.1 Results of the Qualitative Analysis

As required in a sequential mixed-method approach, the results for the first analysis (qual) need to be utilized by the following analysis (QUANT). The qual analysis results in the following three categories of omni-channel attribution criteria.

1. *Feature / Specification criteria* comprising attribution model specific requirements
2. *Data criteria* comprising data requirements for attribution modelling
3. *Other criteria* containing other requirements identified during the research

2.2.2.1.2 Research Ethics

Conducting expert interviews raises ethical concerns. Although this investigation does not collect personal information or habits it is important to the author to meet research ethical standards (Warwick 1982). To ensure compliance with ethical guidelines, the following restrictions and concerns were part of this investigation:

- introducing the research project and its objective to the interview participants
- obtaining the interviewee's consent to voluntary participation and the use of collected information in the resulting publication
- information on how the data is being used as part of a doctoral study
- anonymization of personal data to ensure that no inference to the participants is possible
- an accompanying responsible treatment of all personal data
- transfer of the research results back to the participants

Informing the interviewees about the assumptions in advance was a challenge in complying with these guidelines because this can influence the research results (Diener and Crandall 1978). In keeping with Gläser and Laudel (2010) recommendations, the abstract description of the research goal was communicated to the interviewees for this analysis.

2.2.2.1.3 Identifying the Research Gap – the Lack of an Omni-Channel Ready Attribution Approach

By applying the identified requirements and specifications onto the identified attribution models, the research gap is clearly identified as required by Kuckartz (2014). No attribution model meets a majority of the identified requirements. Having the research gap identified justifies the second QUANT analysis.

2.2.2.2 Quantitative Analysis

For the QUANT analysis based on the qual analysis, the CRISP-DM (Shearer 2000) and the MAP (see chapter 4) are utilized to develop a data basis meeting the data requirements and a corresponding attribution approach meeting the prior identified model requirements. For the research different data sources are utilized, provided by one of the largest online real-estate platforms in Germany.

The in 1996 developed CRISP-DM is still the mostly applied methodology for analytics, data mining, or data science projects (Piatetsky 2014). Since 1996 the CRISP-DM has not been significantly further developed. Piatetsky (2014) points out the need for problem-specific, more detailed approaches. The developed MAP methodology (see chapter 3 and 4) offers such an approach for marketing bigdata problems. In contrast to the more general CRISP-DM approach, the MAP approach specifies six tangible phases. The following six phases serve as orientation for the presented research.

1. Problem statement and goal definition
2. Selection of data sources
3. Data preparation
4. Modeling
5. Model evaluation
6. Recommendation for action

The QUANT analysis utilizes the CRISP-DM and the MAP methodology. The research is guided by the hypotheses H3 and H4 listed in Table 3.

For the quantitative analysis of the presented research, real user interaction data from one of Germany's largest real-estate platforms is utilized. The utilized raw data sets consist of over 240.000.000 hits/touchpoints from which over 225.000.000 hits are placed in more than 9.700.000 journeys. This large amount of data ensures significance (*The Law of Large Numbers* (Mlodinow 2008)) within the data to support the results. Data quality and objectivity depend on existing tracking issues such as ad-blockers or clients with deactivated javascript. It is the objective to assemble a holistic data set of the customers to understand the usage behavior.

All data is structured and raised by machine. Considering tracking issues, by utilizing a well-tested and longstanding tracking approach maintained by the data providing company, any research would result in the same raw data. The *objectivity of execution* is ensured.

The *objectivity of analysis* is ensured as well. All executed transformations are necessary transformations for the process and are well documented. This includes data selection, data joins and data neglection.

The feature engineering process is based on *domain knowledge*. The same set of features will be engineered if the same degree of domain knowledge is available.

The execution of a Principal Component Analysis (PCA) is objective by default.

The setup of the machine learning approach including the described data splits ensure the objectivity of analysis.

Finally, the last step of the research, the hyperparameter optimization, ensures the objectivity as well.

The objectivity of interpretation is ensured since the interpretation of the results is based on facts contained in the data or resulting from the process.

The objectivity of the qualitative research is assured, since the *objectivity of execution,* the *objectivity of analysis* and the *objectivity of interpretation* are given.

In as much as the objectivity of the investigation is ensured, the investigation needs to be analyzed regarding its *reliability* – e.g., the accuracy (Krauth 1995). No random errors exist which falsify the result. As described within the process, outliers representing users with uncommon behavior, were analyzed and rated as users with realistic behavior.

The whole data set is split into a training, validation and test portion. The test portion has not been utilized for any training or modeling. This setup ensures the reliability of the investigation. Utilizing the same data foundation results in the same outcome. The reliability of the investigation is given.

The third quality factor is the *validity* of the investigation. The developed model has been tested regarding the test split of the collected data. Since the intended results were achieved, the process can be considered valid.

2.2.2.2.1 Developing a Data Foundation for Omni-Channel Attribution

The first part of the QUANT analysis consists of the development of an omni-channel ready data basis. The generated data basis is analyzed with respect to the applicability for attribution in an omni-channel environment. The applicability is assured by evaluating the data foundation by applying the pre-identified data requirements. By utilizing the results of the first qual analysis, the sequential mixed-method approach is correctly implemented (Kuckartz 2014). This ETL (Extract, Transform, Load) process is described in chapter 6.

Based on the generated data basis, so-called *features* need to be selected and/or extracted (Meyer and Whateley Brendon 2004; Menkov et al. 2006). Such features represent the input variables for the downstream machine learning model, which is part of the presented attribution approach. The process of feature selection consists of identifying relevant features by selecting them from the given feature set (Meyer and Whateley Brendon 2004). In contrast to the feature selection process the feature extraction process involves the development or derivation of new features based on existing features (Menkov et al. 2006). Combining the development of the holistic customer journey (HCJ) data foundation and the feature generation process, the whole data preparation process is structured as an ELT (Extract, Load, Transform) process.

The process of feature generation (selection / extraction) is a very complex task (Meyer and Whateley Brendon 2004). There is no holistic methodology which can be applied to guide this process to identify the optimal set of features. For the feature generation process there are two different approaches which can be utilized (Menkov et al. 2006). The first one applies *domain knowledge* as key driver to select or extract features. This approach enforces outstanding knowledge about the data in general, the described processes represented within the data and an excellent understanding of the data context. For the presented research an understanding of the company’s customers is indispensable.

Alternatively, as a second option, an automated feature generating approach could be utilized. There are different libraries supporting the feature engineering process such as the *featuretools* Python library (Feature Tools Development Group 2018). Considering the main research question, it needs to be analyzed what attribution looks like in an omni-channel environment. The current research is not focusing on developing the optimal attribution approach for the provided data set. Applying an automatic feature generation approach is very time consuming and the quality of the results are often not outstanding and purposeful (Domingos 2012). For the presented research existing domain knowledge is chosen for the feature generation process.

2.2.2.2.2 Developing the Omni-Channel Ready Attribution Approach

As already mentioned, the generated features represent the input data for the development of the machine learning (ML) model approach. Based on the semi-structured expert interviews during the prior qual analysis, it has been identified that a ML or artificial learning approach is stated as a model requirement. Therefore, a ML approach utilizing different tree-based algorithm is chosen for the presented research. The attribution approach is evaluated the same way as the data foundation. All identified model requirements are used to analyze the applicability of the attribution approach in an omni-channel environment.

The attribution problem is treated as a data mining problem, trying “[to mine] knowledge […] from data” (Han et al. 2012) and to solve it with ML algorithms. The methodology for the development of the machine learning is described in detail in publication 3 (see chapter 7).

As identified as model requirement (see chapter 5), in an omni-channel environment an attribution on a channel basis alone is no longer purposeful. An attribution on an audience level or user level is required. Based on the results of the expert interviews and in cooperation with the data providing company, two user-level attributes or indicators are identified to be the targeted output of the attribution approach. Firstly, a *customer value* representing the value of a customer at its current state and secondly, a classification indicating the *conversion probability*, the likeliness of the user to perform another conversion are chosen.

The QUANT analysis results in an omni-channel ready data basis and an omni-channel ready attribution approach. The data basis and the attribution approach are evaluated to answer the main research question.

2.3 Justifying the Hypotheses

The presented research is guided by the following four hypotheses H1, H2, H3 and H4. Each hypothesis is described in the enclosed corresponding publication. This includes the detailed derivation from the main research questions, the research gap and the theoretical background. In Table 3 all four hypotheses are listed with the corresponding applied methodology.

The first two hypotheses "*New requirements are requested for attribution modelling from a practical point of view in an omni-channel environment*" and "*Existing attribution models are not effectively applicable in an omni-channel environment from a practical perspective*" aim at analyzing requirements for an attribution approach in an omni-channel environment and identify the research gap, that no efficient attribution approach for an omni-channel environment exists. Hypotheses three and four guide the research on building an omni-channel ready attribution approach and its performance.

Table 3: Hypothesis and applied methodologies

Hypothesis		**Applied methodology**
H1	New requirements are requested for attribution modelling from a practical point of view in an omni-channel environment.	Structured Literature Review (Greenhalgh and Peacock 2005; Webster and Watson 2002) Qualitative Approach • Qualitative Content Analysis: Semi-structured expert interviews (Schreier 2012)
H2	Existing attribution models are not effectively applicable in an omni-channel environment from a practical perspective.	
H3	It is possible to build a required data foundation and attribution model to work efficiently in an omni-channel environment.	Quantitative Approach • CRISP-DM (Shearer 2000) • MAP (see chapter 4)
H4	If such a model can be developed, savings from at least 10% can be achieved for e.g. a company or an institution.	

2.3.1 Hypothesis 1

New requirements are requested for attribution modelling from a practical point of view in an omni-channel environment.

The identified shift towards omni-channel marketing and the research priorities of the Marketing Science Institute (MSI 2016) identify the relevance for research in an omni-channel environment. Since there is no research identified during the structured literature review process, presenting evaluation criteria for attribution approaches in a multi-channel environment, omni-channel environment or in general, there is a lack of attribution requirements which needs be analyzed. It has not been determined what attribution in an omni-channel environment looks like. The analyzing of the applicability of the identified attribution models in an omni-channel environment clearly indicates that prior research cannot be applied in an omni-channel environment. Assuming, existing attribution approaches meet *requirements* in a multi-channel environment indicates the necessity of new requirements within an omni-channel environment (H1).

2.3.2 Hypothesis 2

Existing attribution models are not effectively applicable in an omni-channel environment from a practical perspective.

Static attribution approaches are still widely applied (eMarketer 2016). Already in a multi-channel environment those static attribution approaches are identified as inaccurate (Petersen et al. 2009). Those models probably remain inaccurate in an omni-channel environment as well. This circumstance has not been analyzed yet and justifies the second hypothesis.

2.3.3 Hypothesis 3

It is possible to build a required data foundation and attribution model to work efficiently in an omni-channel environment.

Since the identified attribution modelling requirements have not been applied by any research, the third hypothesis, analyzing whether those criteria are implementable or not, is justified.

2.3.4 Hypothesis 4

If such a model can be developed,
savings from at least 10% can be achieved for e.g. a company or an institution.

As in hypothesis three, the modelling requirements and the derived attribution approach have not been analyzed by any other researcher. Analyzing the results in terms of savings and optimization potential justifies the last hypothesis of the research.

3 Summary of the Marketing Analysis Process (MAP) Methodology

The first publication of the current research is titled “Marketing-Analytics-Process (MAP) – Data-Driven-Marketing-Projekte erfolgreich durchführen“, Eng. *Marketing-Analytics-Process (MAP) – Successful implementation of data-driven marketing projects*. This publication is written in German as it is a book chapter in a German marketing controlling guide. Since the current research is written in English, a summary of the publication is included in this chapter. The first publication, the original book chapter is included in chapter 4.

The developed Marketing-Analytics-Process (MAP), a specification of the CRISP-DM (Shearer 2000), is described within the first publication from both a theoretical perspective and a use-case perspective. The MAP is a framework for the procedural approach in the implementation of data-driven marketing projects based on bigdata.

Due to the increasing dynamization of markets and the available technical possibilities to use bigdata to generate competitive advantages, the extent of external and internal complexity has increased significantly (Schoeneberg et al. 2016). Today, however, only a limited number of companies have the necessary corporate strategies and business models to exploit these competitive advantages strategically and operationally successfully. However, the possession of large amounts of data offers no added value in the absence of suitable analytical models or methods (Malgara 2014). Implemented in the specialist department, bigdata projects are executed in six phases (see Figure 4) in which several iterations and returns are possible.

The Marketing Analytics Process (MAP) model was developed specifically for the challenges of operational and strategic marketing and adapts the world’s most widely used data analysis process, the CRISP-DM. It should be noted that the CRISP-DM, originally developed in 1996, has not been significantly further developed. New questions and challenges - such as the transfer of technical know-how to the specialist departments and the handling of the challenges posed by bigdata - are no longer sufficiently considered. Fan et al. already point out in 2015 that bigdata acts as a driver for decision-making processes in (online-) marketing and the alignment of processes is necessary (Fan et al. 2015). This is where the MAP comes in. The six phases of the MAP are shortly outlined.

Table 4: Phases of the Marketing Analytics Process

MAP	Description of the phase / Actions
Phase 1	**Problem statement and goal definition** The central phase of any MAP analysis project is the first phase. It identifies the problems and defines the goals. The focus is on understanding the project goals from a professional perspective. These problems must be specifically translated into technical problems and a preliminary project plan must be drawn up.
Phase 2	**Selection of data source** The data source selection phase can be divided into five steps according to Shearer (2000). First, a meaningful sample is drawn from each data source (I), the structure of the data sources is described (II) and the data sifted (III). The data can be viewed by means of specific queries or visualization. This generates initial results and hypotheses. The review of the data provides insights into data quality, which must be ensured for each data source (IV). If the data quality meets the requirements or has been subsequently transformed into the desired form, a data schema including sample data must be created (V), which is utilized for modeling. The live system is not burdened by queries and is protected against unintentional manipulation.
Phase 3	**Data preparation** In the data preparation phase, the data is transformed to be utilized for further processing using modelling/analysis software. Three steps - data set selection (I), data linking (II) and data cleansing (III) - are performed for this purpose. As already described in Phase I, data preparation with a share of between 50 and 80 % takes up most of the time required for an analysis project (Granville 2015; Dasu and Johnson 2003). The objectives of data preparation and the previous data source selection are to ensure the quality criteria objectivity, reliability and validity of the data.
Phase 4	**Modeling** The modeling phase is divided up into four steps. The selection of the modelling technique (I) is based on the objective or business problem and the data properties (Liao et al. 2012). Test procedures are then created (II). At least one model is created in step three (III). According to Ngai, the most common methods of analysis are association, classification, cluster and regression analysis (Ngai et al. 2009), but descriptive analyses are also relevant according to Haumer (Haumer 2015). Finally, the models are evaluated in this step (IV). Here, intensive cooperation between the Data Scientist and the specialist department is highly recommended (Shearer 2000).

Phase 5	**Model evaluation** In the model evaluation phase, the models resulting from the previous phase are reviewed (Shearer 2000). The technically flawless models - from the analysts' point of view - must now be inspected with regards to their technical objectives. Furthermore, the procedure that led to the creation of the models must be validated. The data sources and data preparation are therefore also checked once again. It is therefore validated whether everything relevant has been considered so far in order to achieve the defined goal with the models.
Phase 6	**Recommendation for action** The final phase includes the formulation of recommendations for action. A frequent recommendation for action is the implementation or automation of the models. A recommendation for action is to formulate a report that documents the process carried out, describes and visualizes results - such as models, findings or artifacts created - and makes a recommendation for further action. The actions are recommended with a focus on Return of Investment (ROI). Alternatives are described and evaluated and possible effects on the existing business - both strategic and operational - are explained.

4 Publication 1 MAP - Marketing Analytics Process

MAP – Marketing Analytics Process	
DOI	10.1007/978-3-662-50406-2_2
Format	Book chapter
Book	Handbuch Marketing-Controlling 4th Ed.
Language	German
Status	Published
Summary	Aufgrund der zunehmenden Dynamisierung der Märkte und den vorhandenen technischen Möglichkeiten, Big Data zur Generierung von Wettbewerbsvorteilen nutzen zu können, hat der Umfang unternehmensexterner wie -interner Komplexität stark zugenommen. Dieser Beitrag beschreibt den von den Autoren entwickelten Marketing-Analytics-Process, der ein Framework zur prozessualen Vorgehensweise bei der Umsetzung von Data-Driven-Marketing-Projekten auf Basis von Big Data darstellt. In der Fachabteilung implementiert, werden Big-Data-Projekte dazu in sechs Phasen umgesetzt, bei denen Iterationen und Rücksprünge möglich sind. Zum besseren Verständnis wird jede Phase detailliert beschrieben und durch einen fortlaufenden Use-Case, basierend auf einem After-Sales-Projekt des Unternehmens Immonet, ergänzt. Der Marketing-Analytics-Process wird detailliert aus wissenschaftlicher, praktischer und technischer Sicht beschrieben.

Einleitung

Aufgrund der zunehmenden Dynamisierung der Märkte und den vorhandenen technischen Möglichkeiten, Big Data zur Generierung von Wettbewerbsvorteilen nutzen zu können, hat der Umfang unternehmensexterner wie -interner Komplexität stark zugenommen (Schoeneberg et al. 2016). Bisher verfügen jedoch nur eine begrenzte Anzahl von Unternehmen über die notwendigen Unternehmensstrategien und Geschäftsmodelle, um diese Wettbewerbsvorteile strategisch wie operativ erfolgreich für sich nutzen zu können. Der Besitz großer Datenmengen bietet jedoch keinen Mehrwert, sofern geeignete Analysemodelle oder -methoden fehlen (Malgara 2014).

Im Hinblick auf Big Data müssen sich Organisationen vorab die Fragen stellen, welche Ziele sie mit der Nutzung verfolgen wollen, welche Quellen sie nutzen können und welche pragmatischen und ergebnisbezogenen Vorgehensweisen für sie zielführend sind. Dafür ist es erforderlich, dass die gesamte Organisation das Potenzial von Big Data erkennt. Im Marketing haben sich hier Buzzwords, wie Marketing-Intelligence, Marketing-Analytics, Smart Data oder Data-Driven-Marketing, bereits etabliert. Das datenbasierte Marketing bietet Organisationen heute vielfältige neue Möglichkeiten, mehr über die wahren Bedürfnisse der Kunden zu erfahren und diese gezielt ansprechen zu können. Mittels prädiktiver Analysen ist die Kommunikation mit bestehenden und potenziellen Kunden zum richtigen Zeitpunkt crossmedial realisierbar. Dieser Beitrag beschreibt den von den Autoren entwickelten Marketing-Analytics-Process, der ein Framework zur prozessualen Vorgehensweise bei der Umsetzung von Data-Driven-Marketing-Projekten auf Basis von Big Data darstellt. In der Fachabteilung implementiert, werden Big-Data-Projekte in sechs Phasen umgesetzt, bei denen mehrere Iterationen und Rücksprünge möglich sind. Zum besseren Verständnis wird nachfolgend jede Phase detailliert beschrieben und durch einen fortlaufenden Use-Case zur Implementierung ergänzt. Der jeweils hervorgehobene Use-Case dient als Best-Practice-Beispiel und basiert auf einem After-Sales-Projekt des Unternehmens Immonet. Über die Beschreibung des Projektverlaufs hinaus werden hier auch die technischen Implementierungen

anhand von Programmcodes skizziert, um einem Analysten einer Fachabteilung eine entsprechende Guideline zu bieten.

Der Marketing-Analytics-Process wird anschließend detailliert aus wissenschaftlicher, praktischer und technischer Sicht dargestellt.

Marketing-Analytics-Process-(MAP)-Modell

Das *Marketing-Analytics-Process-(MAP)-Modell* wurde speziell für die Herausforderungen im operativen und strategischen Marketing entwickelt und adaptiert den meist genutzten Datenanalyseprozess der Welt, den CRISP-DM (Piatetsky 2014). Kritisch anzumerken ist hierzu, dass der ursprünglich bereits 1996 entwickelte CRISP-DM nicht mehr wesentlich weiterentwickelt wurde. Neue Fragestellungen und Herausforderungen – wie die Verschiebung

von technischem Know-how in die Fachabteilungen sowie der Umgang mit den Herausforderungen durch Big Data – werden nicht mehr hinreichend berücksichtigt. Fan et al. (2015) weisen bereits 2015 darauf hin, dass für Entscheidungsprozesse im (Online-)Marketing Big Data als Treiber fungiert und das Angleichen von Prozessen notwendig ist (Fan et al. 2015). An dieser Stelle setzt der MAP an.

Der Prozess ist so konstruiert, dass er in der (Online-)Marketing-Fachabteilung implementiert wird. Der MAP bildet eine explorierende Basis mit dem Ziel der Wissens- und Erfahrungsgenerierung, also *Insights*, aus Daten für eine spätere mögliche Automatisierung oder Implementierung durch Analysten, deren Fokus dabei auf der fachlichen Problemlösung liegt.

Der MAP-Modell besteht aus sechs Phasen (vgl. Figure 4), welche zumeist in einer Implementierung des Modells münden. Der Anstoß zur Umsetzung von strategischen Zielen durch den MAP kommt in der Regel von außerhalb, von operativen Zielen innerhalb der (Online-) Marketing-Abteilung. Zunächst werden Probleme identifiziert und Ziele definiert (*Phase I*). Anschließend werden die benötigten Daten aus internen und externen Datenquellen gesammelt (*Phase II*). Liegen die zu analysierenden Daten vor, werden diese durch Datenverknüpfung, Datensatzselektion und Datenbereinigung aufbereitet (*Phase III*), um den Ansprüchen der Modellierung (*Phase IV*) und den einzusetzenden Analyseverfahren zu genügen. Bei der Modellierung ist besonderer Fokus auf eine spätere Flexibilität, d. h. Erhöhung der Modellpräzision durch Erweitern oder Verändern der Parameter, zu legen. Zielführend ist das Erstellen mehrerer Modelle, deren Güte anschließend evaluiert wird (*Phase V*). Ergibt die Validierung der Modelle, dass durch diese die zuvor identifizierten Probleme nicht wie gewünscht gelöst werden können, so sind weitere Iterationsschritte notwendig. Hierbei ist das Augenmerk aber nicht auf eine perfekte Modellgüte zu legen, da durch jede weitere Iteration wiederholt Ressourcen gebunden werden. Daher kann innerhalb des MAP-Modells aus jeder Phase in eine der vorherigen Phasen zurückgesprungen werden, um kurzfristig Anpassungen vorzunehmen und Fehler frühzeitig zu kompensieren. Die aus modelltheoretischer wie praktischer Sicht gängigen Feedback-Sprünge sind in Figure 4 dargestellt.

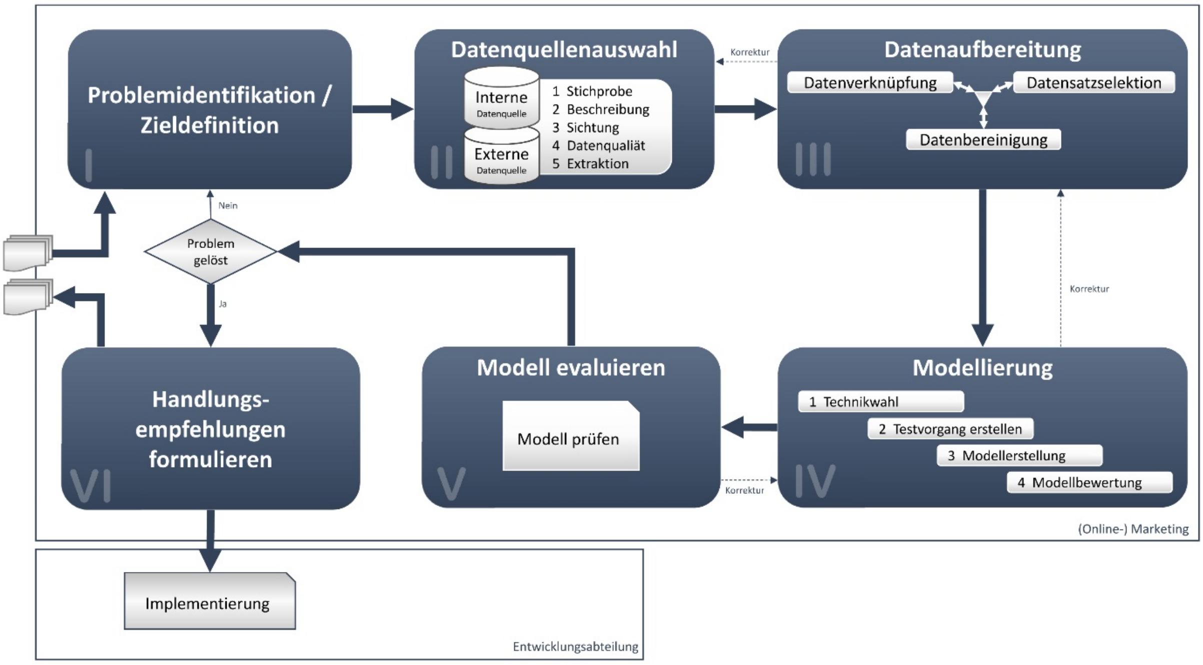


Figure 4: Marketing-Analytics-Process Modell (MAP). *(Quelle: eigene Darstellung)*

Anschließend wird eine Handlungsempfehlung an den Auftraggeber ausgesprochen, welche immer auch in einem resultierenden Bericht stattfindet (*Phase VI*). Eine der häufigsten Handlungsempfehlungen stellt die Implementierung einer automatisierten Modellausführung durch eine Entwicklungsabteilung dar. Durch eine solche Implementierung können die Ergebnisse der automatisierten Analyse regelmäßig und zeitnah abgerufen oder verteilt werden.

Auf die Phasen des MAP-Modells wird in den folgenden Kapiteln detailliert eingegangen.

Zum besseren Verständnis wird jede Phase anhand eines fortlaufenden Use-Case praxisnah beschrieben.

Phase I: Problemidentifikation/Zieldefinition

Die zentrale Phase jedes MAP-Analyseprojektes ist die erste Phase. In dieser werden die Probleme identifiziert und die Ziele definiert. Der Fokus wird auf das Verständnis der Projektziele aus der fachlichen Perspektive gesetzt. Diese fachlichen Problemstellungen sind speziell in technische Problemstellungen zu überführen und ein vorläufiger Projektplan ist zu erstellen.

Da von dieser Phase aus der weitere Prozessablauf gebildet wird, haben die getroffenen Entscheidungen eine große Hebelwirkung auf den Ressourceneinsatz, die Ergebnisqualität sowie den Projekterfolg. Um sicherzustellen, dass nicht die richtigen Antworten auf die falschen Fragen gefunden werden, ist die Analyse der wichtigsten fachlichen Ziele und die damit verbundenen Fragestellungen unabdingbar. Wenn das fachliche nicht richtig in ein technisches Ziel überführt werden kann, dann ist eine erneute Definition in Betracht zu ziehen. Abschließend ist ein Projektplan zu erstellen, welcher eine Planung für das Erreichen der technischen Ziele, eine Zeitplanung, Betrachtung potenzieller Risiken und eine Auswahl an Tools und Techniken beinhaltet. In der Praxis haben sich die Phasen der Datenauswahl sowie der Datenaufbereitung als besonders ressourcenintensiv herausgestellt. Während die Datenauswahl und das entsprechende Verständnis dafür erfahrungsgemäß ein Viertel der zeitlichen Projektaufwände binden, nimmt die Datenaufbereitung zwischen 50 und 80 % der Projektressourcen in Anspruch (Granville 2015; Dasu and Johnson 2003).

Bei neuen Projekten ist darauf zu achten, dass in der Phase der Zieldefinition und Problemidentifikation die Einarbeitung des Analysten in die fachliche Problemstellung nicht ausschließlich im Zentrum steht. Der Fokus liegt ebenso in der Schaffung einer realistischen Erwartungshaltung in der Fachabteilung sowie der Anpassung der fachlichen Prozesse an die Arbeit mit den resultierenden Analysemodellen. Um diese Diskrepanzen möglichst gering zu halten, sind sowohl der MAP als auch mindestens ein Analyseexperte direkt in der Fachabteilung bzw. in einem Kompetenzcenter zu implementieren.

Des Weiteren zeigt sich in der Praxis, dass die vom Management vorgegebenen Projektziele oft nicht mit den Anforderungen der Fachabteilung übereinstimmen. Aufgabe der Fachabteilung ist es daher, aus den vom Management vorgegebenen globalen Zielvorgaben, konkrete, praktisch umsetzbare Projektziele abzuleiten.

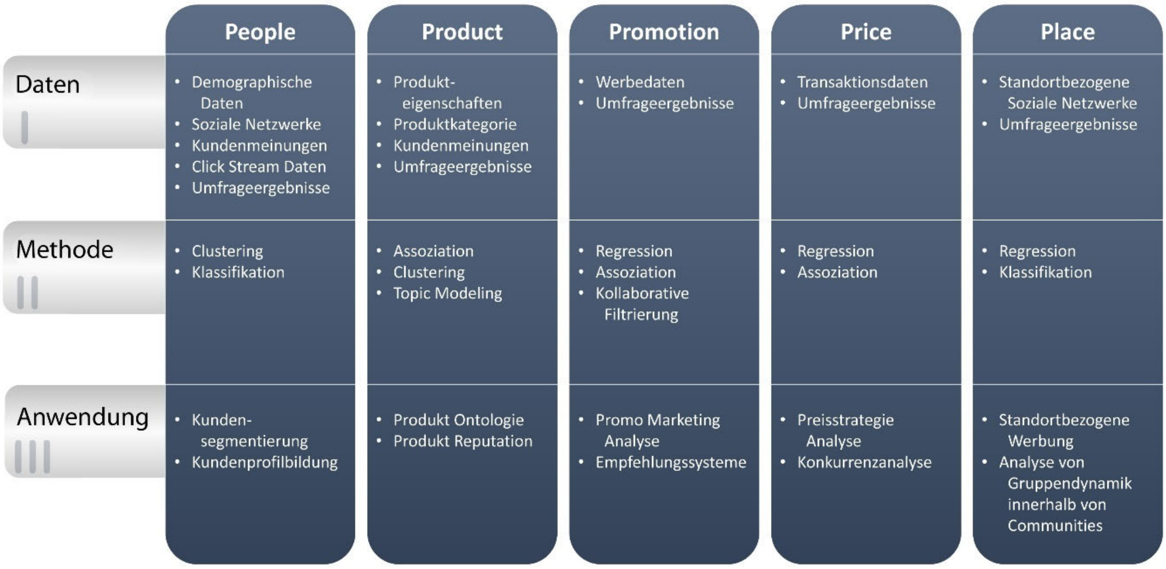
Exkurs: Analysen im (Online-)Marketing-Kontext

Herausfordernd für eine Organisation sind heute die wachsenden Ansprüche des Kunden, welche sich in den letzten Dekaden stark gewandelt haben. Aus Marketingsicht beinhaltet dies, eine Relevanz beim Kunden zu erzeugen bzw. zu erhalten. Der Endkunde im digitalen Zeitalter erwartet „maßgeschneiderte“ Angebote und eine individuelle Ansprache. Die Relevanz eines individuellen Marketing ist bereits seit langem bekannt und stellt trotzdem immer noch eine große Herausforderung dar.

Während Borden 1964 den Begriff des „Marketing Mix“ mit seinen 12 Elementen prägte (Borden 1964), gruppierte McCarthy diese Elemente in vier Gruppen, die als 4Ps (für Product, Price, Promotion und Place) bekannt sind (McCarthy 1964). Laut Goi wird das 4P-Modell als höchst relevant für das Consumer Marketing angesehen (Goi 2009). Damit wurde das Marketing, laut Judd, jedoch als zu produktionsorientiert definiert. Daher integrierte dieser ein fünftes P, für People (Judd 1987).

Fan et al. veröffentlichten 2015 ein Framework auf Basis der 5Ps, welches einen Leitfaden für den Entscheidungsprozess im Marketing im Kontext von Big Data bereitstellt (Fan et al. 2015).

Die 5Ps werden in Relation zu den Daten, den möglichen Analysemethoden und Anwendungsfällen gesetzt, wodurch eine strukturierte Datenanalyse ermöglicht wird.



	People	Product	Promotion	Price	Place
Daten I	• Demographische Daten • Soziale Netzwerke • Kundenmeinungen • Click Stream Daten • Umfrageergebnisse	• Produkt-eigenschaften • Produktkategorie • Kundenmeinungen • Umfrageergebnisse	• Werbedaten • Umfrageergebnisse	• Transaktionsdaten • Umfrageergebnisse	• Standortbezogene Soziale Netzwerke • Umfrageergebnisse
Methode II	• Clustering • Klassifikation	• Assoziation • Clustering • Topic Modeling	• Regression • Assoziation • Kollaborative Filtrierung	• Regression • Assoziation	• Regression • Klassifikation
Anwendung III	• Kunden-segmentierung • Kundenprofilbildung	• Produkt Ontologie • Produkt Reputation	• Promo Marketing Analyse • Empfehlungssysteme	• Preisstrategie Analyse • Konkurrenzanalyse	• Standortbezogene Werbung • Analyse von Gruppendynamik innerhalb von Communities

Figure 5: Marketing-Mix-Framework im Rahmen des Big Data-Managements. *(Quelle: eigene Darstellung in Anlehnung an Fat et al. 2015, S.29)*

Praxis Use-Case

Ziel des (Online-)Marketing ist es, die eigenen Produkte oder Dienstleistungen mit wenig Aufwand und Budget bestmöglich dem Kunden zu präsentieren, sodass dieser die vom Marketing angestrebte Handlung durchführt. In diesem Use-Case wird zunächst der Geschäftsbereich der After-Sales-Produkte von Immonet.de skizziert. Darüber hinaus wird detailliert beschrieben, wie der gezielte Einsatz von Big-Data-Quellen und -Analysen zu einer erheblichen Absatzsteigerung durch Einsatz des MAP geführt hat.

Immonet.de ist eines der führenden Immobilien-Portale in Deutschland. Mit über 20 Millionen Visits pro Monat ist Immonet.de ein seit Jahren etabliertes Portal, auf dem unterschiedliche Objekte (*Objekt wird als Oberbegriff für unterschiedliche Immobilientypen verwendet. Darunterfallen u. a. Häuser, Wohnungen, Appartements, Garagen, Stellplätze, Gewerbe- und Anlage-Immobilien*.) von Immobilienanbietern angeboten und durchsucht werden können.

Neben dem eigentlichen Kerngeschäft von Immonet.de, dem Zusammenführen von Objekt-Anbietern (Makler oder auch private Anbieter) und Objekt-Nachfragern (Kunden), verkauft Immonet.de Leads an verschiedene Partner. In diesem Kontext bezeichnen Leads einen Kontakt, der über eine (Online-)Marketingmaßnahme gewonnen wurde. Diese Leads werden als After-Sales-Produkte bezeichnet.

Bereiche, in denen Leads generiert werden (alphabetisch sortiert):

- Finanzierung (Fipa)
- Hausbaukatalog (Katalog-Hausbau)
- Umzug (UA)

Bisher wird der Kunde durch interne Werbung auf Immonet.de und über Newsletter auf die verschiedenen After-Sales-Angebote aufmerksam gemacht. Darüber hinaus werden SEO- (Searchengine Optimization) und SEA- (Searchengine Advertisement) Maßnahmen durchgeführt, um die Einnahmen im Lead-Verkauf zu steigern.

Zielsetzung

Auf der Managementebene wird die Anforderung formuliert, dass die Performance der After-Sales-Produkte gesteigert werden soll. Hier wird ein großes, in der Vergangenheit ungenutztes, Potenzial gesehen. Eine konkrete, messbare Zielvorgabe beinhaltet die Vorgabe des Managements nicht.

Wie bereits erwähnt, werden die After-Sales-Produkte bisher zumeist ohne übergeordnete Strategie mithilfe verschiedener Online-Marketing-Kampagnen beworben. Der MAP soll genutzt werden, um die Effizienz der Leadgenerierung zu optimieren.

Folgende Ziele werden definiert:

1. Erstellung einer Customer-Life-Time (CLT) für die vier Kernprodukte Haus/kaufen, Haus/mieten, Wohnung/kaufen und Wohnung/mieten. Als Datengrundlage sind die bereits erhobenen Informationen des Kundennutzungsverhaltens (Sessiondaten) anzuwenden.
2. Erstellung eines Forecast-Modells zur Vorhersage, zu welchen Zeitpunkten im Kundenlebenszyklus ein Kunde Bedarf an After-Sales-Produkten hat.
3. Integration des Forecast-Modells zur automatisierten Kundenansprache.

Projektplan

Folgende Meilensteine mit jeweiliger Zeiteinschätzung werden definiert und in Table 5 dargestellt.

Table 5: Projektplan inkl. Zeiteinschätzung

	Beschreibung	**Zeiteinschätzung**
M1	Datenquellen wählen und ein Verständnis der Daten erhalten	4PT
M2	Daten-Aufbereitung, inkl. Daten-Qualitätsprüfung, -Transformationen, -Verknüpfbarkeit und -Auswahl	8PT
M3	Erstellung des Modells, inkl. Anwendung	2PT
M4	Auswertung der Modellergebnisse	1PT
M5	Handlungsempfehlungen entwickeln	1PT

Legende: M = Meilenstein; PT = Personentag

Die in diesem Projektplan definierten Zeiteinschätzungen beinhalten sowohl die eigentliche Arbeitszeit wie auch die benötigte Zeit für Meetings und andere Abstimmungsprozesse.

Phase II: Datenquellenauswahl

Die Phase der Datenquellenauswahl kann in Anlehnung an Shearer (2000) in fünf Schritte unterteilt werden. Zunächst wird eine aussagekräftige Stichprobe aus jeder Datenquelle gezogen (I), die Struktur der Daten beschrieben (II) und die Daten gesichtet (III). Die Sichtung erfolgt durch gezieltes Abfragen oder eine Visualisierung der Daten. Hierdurch werden erste Ergebnisse generiert sowie Hypothesen aufgestellt. Aus der Sichtung der Daten ergeben sich Erkenntnisse hinsichtlich der Datenqualität, welche für jede Datenquelle sicherzustellen ist (IV). Entspricht die Datenqualität den Anforderungen oder ist diese durch nachträgliche Transformationen in die gewünschte Form gebracht worden, ist ein Abbild der Daten zu erstellen (V), welches für die Modellierung genutzt wird. Das Live-System wird dadurch nicht aufgrund von Abfragen belastet und ist vor oft unbeabsichtigter Manipulation geschützt.

Eine Datenquelle wird entweder als intern oder als extern bezeichnet. Historisch gesehen wird dies durch die physikalische Lokalität definiert. In Zeiten des Cloud-Computings werden Daten immer öfter nicht auf unternehmenseigenen Servern gespeichert. Zur aktuellen Unterscheidung bedarf es einer Abgrenzung, wie intern und extern in diesem Kontext zu verstehen sind. Interne Datenquellen beinhalten die Daten, welche in der eigenen Abteilung vorgehalten werden und auf welche ohne Einschränkungen zugegriffen werden kann. Für die Durchführung einer Datenanalyse werden interne Datenquellen benötigt; diese stellen aber nur einen geringen Anteil an den gesamt zu nutzenden Datenquellen dar. Bei externen Datenquellen ist der Zugriff bspw. Durch Zugriffsrechte eingeschränkt. Dies bezieht sich auf Daten, welche sowohl in der eigenen Organisation, als auch extern vorliegen. In diese Kategorie fällt auch externes Wissen, welches noch nicht in Datenform vorliegt. Externes Wissen kann z. B. durch Umfragen erhoben werden, um dieses in interne Daten umzuwandeln. Externe Daten können auch in dem eigenen Unternehmen vorliegen, aber nicht ohne Einschränkungen zugänglich sein. Dies kann an der Datenhoheit, wie zum Beispiel den Besitz der Datenquelle durch eine andere Abteilung, ungenügenden Zugriffsrechten oder Einschränkungen durch den Datenschutz sowie Corporate Compliance liegen.

Für typische Marketing-Intelligence-Aufgaben, wie etwa Customer-Opinion-Mining, besitzen Unternehmen heute viele unterschiedliche Ansätze, um Daten aus diversen Informationsquellen zusammen zufügen (Fan et al. 2015). Die Daten werden auf verschiedenen Wegen generiert.

Zum einen werden Daten erhoben, bei denen der Webnutzer den Inhalt aktiv generiert hat, wie etwa Tweets oder Posts. Durch diese Daten können aktuelle Ereignisse identifiziert oder die Stimmung zu bestimmten Themen ermittelt werden (Kimball and Merz 2000; Shugan 2004). Zum anderen werden Daten durch Beobachtungen, wie bspw. Beim Tracking, erhoben. Diese werden durch Server-Weblogs generiert, welche die Interaktionen – also das Verhalten – von Webnutzern aufnehmen. Durch diese Weblogs kann ermittelt werden, welche Aktionen in welcher Reihenfolge von einem Nutzer durchgeführt wurden. Diese Clickstreams können Einsichten in die Art der Websitenutzung gewähren (Kimball and Merz 2000; Shugan 2004; Tendick et al. 2016; Jacobs 2009). Beide Methoden können kombiniert werden, wenn eine Beziehung zwischen Benutzerverhalten und -absichten erforscht werden soll (Fan et al. 2015).

Praxis Use-Case

Um das Nutzerverhalten zu erheben und später analysierbar zu machen, wird auf Immonet.de mit Hilfe von Google Analytics (*Google Analytics ist ein User-Tracking-Tool, welches in einer Freien- und einer Premium-Lizenz zur Verfügung steht. (www.google.com/analytics)*) Premium das Nutzerverhalten clientbasiert verfolgt. Auch serverseitig wird das Nutzerverhalten für das Controlling erhoben und in Hadoop-Clustern (*Hadoop ist eine Open-Source-Softwarelösung von Apache. Als Framework ermöglicht Hadoop das Speichern und Analysieren (ggf. mit Zusatzsoftware) großer Datenmengen in einem speziell entwickelten Dateisystem (HDFS)*) aufgezeichnet. Für die geplante Analyse werden die clientseitig erhobenen Daten genutzt, da diese umfangreicher sind und einen höheren Grad der Granularität aufweisen und somit für eine detaillierte Analyse besser geeignet sind.

Ergänzend zu den mittels Google Analytics direkt über das Interface zur Verfügung gestellten Daten und Berichte, werden eventbasierte Rohdaten benötigt, die in Google Analytics-Interface nicht zur Verfügung stehen.

Google BigQuery (*Google BigQuery (cloud.google.com/BigQuery). Als Speicherplatz für große Datenmengen wird zusätzlich Google Cloud Storage benötigt.)* ist eine Datenbank-Softwarelösung, die es ermöglicht, große Datenmengen in der Cloud (*Als Alternative zu Google BigQuery (Cloud-Datenbank-System) sind hier Amazon DynamoDB, Azure DocumentDB zu erwähnen.)* zu verwalten und zu analysieren. Es zeichnet sich durch die bereitgestellte Rechenleistung und die daraus folgende Geschwindigkeit, in der eine Abfrage verarbeitet werden kann, aus. Das Analysieren von Terabyte großen Datenmengen ist daher mit Google BigQuery in akzeptabler Zeit möglich. Zusammenstellen und Aufbereiten der Daten für diesen Use-Case umfassen circa 850 GB. Die kumulierte Rechenzeit für alle datenvorbereitenden Maßnahmen beträgt circa 15 bis 20 min.

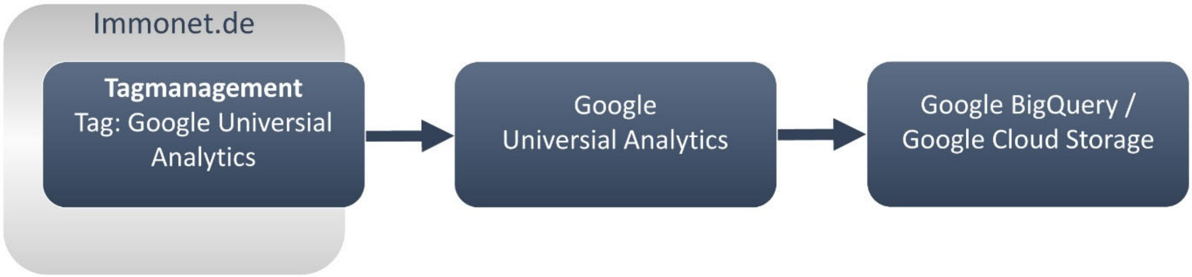

Immonet.de
Tagmanagement
Tag: Google Universial Analytics
Google Universial Analytics
Google BigQuery / Google Cloud Storage

Figure 6: Technische Implementierung des Trackings auf immonet.de, inkl. Datenflusses. *(Quelle: eigene Darstellung)*

Das Analysieren von Google Analytics Premium Daten ist mit Google BigQuery sehr komfortabel (*Die Überführung von Google Analytics-Daten in Google BigQuery ist nur mit einem Google Analytics Premium Account möglich. Mit der frei verfügbaren Version von Google Analytics ist das Überführen nicht möglich*.), da die Rohdaten (*Hit- bzw. Eventbasierte Daten, die nicht aggregiert sind.*) ohne großen Aufwand von Google Analytics Premium direkt an Google BigQuery/Google Cloud Storage übergeben werden können.

Es ist ebenfalls möglich, Daten anderer Tracking-Lösungen mit Google BigQuery zu verarbeiten, solange diese einen Export der Rohdaten ermöglichen. Der Import (*Siehe auch https://cloud.google.com/bigquery/preparing-data-for-bigquery.*) von Daten in Google BigQuery ist aktuell in den Formaten CSV, JSON oder direkt über die Google BigQuery-API möglich. Daten, die nicht von Google Produkten erhoben werden, können somit ebenfalls mit Hilfe von Google BigQuery performant analysiert werden.

User-Tracking auf immonet.de

Zielsetzung ist es, die CLT für verschiedene Produkte zu berechnen, vorherzusagen und die daraus gewonnenen Informationen für gezielte Marketing-Maßnahmen zu nutzen. Wie bereits beschrieben, werden die erhobenen Daten aus Google Analytics automatisch nach Google BigQuery exportiert.

Technische Implementierung des Trackings

Auf Immonet.de werden mithilfe eines Tagmanagementsystems Online-Marketing- Tags ausgespielt. Auch das Google (Universal) Analytics-Tag wird so auf allen Seiten eingebunden. Für eine bessere Erfassung des Nutzerverhaltens wird die Google Analytics Option *Enhanced Ecommerce* (*Siehe https://developers.google.com/analytics/devguides/collection/analyticsjs/ enhanced-ecommerce.*) genutzt.

Die Rohdaten jeder einzelnen User-Aktion werden anschließend an Google Big-Query/Google Cloud Storage weitergegeben. Figure 6 skizziert die technische Implementierung und den Datenfluss.

Technische Implementierung von Leads

Aufgrund des eingesetzten Trackings können User mithilfe einer ID eindeutig über Ihren kompletten Besuch hinweg (auch Session übergreifend) identifiziert werden. Jede von einem User durchgeführte Aktion (page view, event, ecommerce action etc.) kann somit direkt diesem zugeordnet werden. Leads, die für ein After-Sales-Produkt generiert werden, werden mithilfe eines Enhanced-Ecommerce-Trackings aufgezeichnet. Ein Telefonkontakt wird etwa mit folgender Struktur realisiert, die im folgenden Programmcode in JSON dargestellt ist.

JSON-Objekt eines Telefonkontakt Enhanced Ecommerce-Trackings

```
{
  “event”: “trackPhone”,
  “transactionId”: “1234567890”,
  “transactionTotal”: 0.00,
  “transactionProducts”: [ {
    “sku”: “Phonecontact/12345/MeineStadt”,
    “name”: “Phonecontact”,
    “category”: “Phonecontact/MeineStadt/Miete/Wohnen”,
    “price”: 0.00,
    “quantity”: 1.00
  } ]
}
```

Tracking-Rohdaten (Interne Daten)

Aus den Tracking-Rohdaten werden die in Table 6 aufgelisteten Datenfelder für die CLT ausgewählt.

Table 6: Selektierte Datenfelder aus Google BigQuery

Datenfeld	Datentyp	Beschreibung
`fullVisitorId`	**STRING**	Eindeutige Kunden-ID
`visitStartTime`	**INTEGER**	Timestamp
`date`	**STRING**	Datum der Sitzung (JJJJMMTT)
`hits.page.pagePath`	**STRING**	Seiten-URL
`totals.hits`	**INTEGER**	Anzahl der Hits innerhalb einer Session
`hits.item.productName`	**STRING**	Produktname
`hits.item.transactionId`	**STRING**	ID der Ecommerce-Transaktion

Legende: **STRING**: Hierbei handelt es sich um eine Zeichenkette, bestehend aus Buchstaben und/oder Zahlen und/oder Sonderzeichen

INTEGER: Ein **INTEGER** ist eine Ganzzahl

Table 7: Selektierte Datenfelder aus der Objekt-Datenbank

Datenfeld	Datentyp	Beschreibung
`fullVisitorId`	**STRING**	Eindeutige Kunden-ID
`visitStartTime`	**INTEGER**	Timestamp
`date`	**STRING**	Datum der Sitzung (JJJJMMTT)

Legende: **STRING**: Hierbei handelt es sich um eine Zeichenkette, bestehend aus Buchstaben und/oder Zahlen und/oder Sonderzeichen

INTEGER: Ein **INTEGER** ist eine Ganzzahl

Das vollständige Schema (alle Datenfelder), inklusive Beschreibung, stellt Google im eigenen Analytics-Support-Bereich als Übersicht zur Verfügung. (URL: https://support.google.com /analytics/answer/3437719?hl=de.)

Objekt-Daten (Interne Daten)

Als erstes Ziel wurde festgelegt, dass eine Vorhersage für die vier Kernprodukte (Haus/kaufen, Haus/mieten, Wohnung/kaufen und Wohnung/mieten) erfolgen soll. Eine Produktunterscheidung ist mit den Daten aus Google BigQuery nicht möglich, da keine Klassifizierung über die zwei Dimensionen kaufen/mieten und Haus/Wohnung nicht möglich ist. In der Objekt-Datenbank, welche nicht in Google BigQuery stattfinden kann, stehen diese Informationen zur Verfügung, sodass jedes Objekt eindeutig einem der vier Produkte zugeordnet werden kann. Table 7 enthält die Felder der Objektdatenbank, die für die Analyse relevant sind.

Phase III: Datenaufbereitung

In der Phase der Datenaufbereitung werden die Daten so transformiert, dass diese mithilfe einer Modellierungs-/Analysesoftware (vgl. Phase V – Praxis Use-Case) weiter verarbeitbar sind. Hierzu werden drei Schritte – Datensatzselektion (I), Datenverknüpfung (II) und Datenbereinigung (III) – durchgeführt. Wie bereits in Phase I beschrieben, nimmt die Datenaufbereitung mit einem Anteil zwischen 50 und 80 % den Großteil des zeitlichen Aufwandes eines Analyseprojektes in Anspruch (Granville 2015; Dasu and Johnson 2003). Ziele der Datenaufbereitung sowie der vorangegangenen Datenquellenauswahl sind es, die Gütekriterien Objektivität, Reliabilität und Validität der Daten zu sicherzustellen.

Die drei Schritte dieser Phase können in beliebiger Reihenfolge durchgeführt werden. In den meisten Fällen ist es sinnvoll, zunächst die Datensatzselektion – also das Filtern der Daten anhand von zielführenden Kriterien – durchzuführen und anschließend die Verknüpfung der Daten vorzunehmen. Die Verknüpfung – oder auch Integration – von Big Data aus unterschiedlichen Quellen zur Generierung von Marketing-Intelligence ist keine triviale Aufgabe (Fan et al. 2015) aufgrund der Big-Data-Eigenschaften – Umfang, Vielseitigkeit, Geschwindigkeit und Nutzen.

Durch die beiden Schritte *Datensatzselektion* und *-verknüpfung* können die Datenmenge erheblich dezimiert und die Durchführung der folgenden Schritte beschleunigt werden. Anschließend ist die Datenmenge überschaubarer und die Datenbereinigung bzw. das Data Cleaning kann durchgeführt werden. Im Mittelpunkt der Datenbereinigung steht die Datenqualität. Laut Felden besteht ein Zusammenhang zwischen der Datenqualität und der Entscheidungsqualität (Felden 2012). In der vorangegangenen Phase der Datenquellenauswahl wurde bei der Sichtung der Daten festgelegt, ob und welche Mängel hinsichtlich der Datenqualität in den ausgewählten Daten bestehen. Diese Abweichungen von der definierten Datenqualität werden in diesem Schritt der Datenbereinigung angepasst. Es werden Dubletten entfernt, Datenwerte standardisiert bzw. transformiert, fehlerhafte Datensätze entfernt, fehlende Daten mit Standardwerten aufgefüllt und ganze Datenspalten aus den bestehenden Daten abgeleitet. Im Rahmen

von Big Data, genauer automatisch generierten Daten, sind denkbare Gründe für eine Datenbereinigung etwa

- die Veränderung des Erfassungsprozesses und somit unterschiedliche Strukturen von Daten,
- spezielle Anforderungen des Weiteren Modellierungs- oder Verarbeitungsprozesses,
- Vereinigen von heterogenen Datenpools (z. B. aus unterschiedlichen Systemen) oder
- fehlerhafte, in Dubletten resultierende Verknüpfung von Datenquellen.

Entscheidend für die Datenqualität sind das Datenqualitätsmanagement bzw. der Erfassungsprozess. Durch die Dokumentation der Schritte kann das Vorgehen nachvollzogen oder gegenüber Dritten gerechtfertigt werden. Besonders wichtig ist das Dokumentieren des Vorgehens, wenn der Prozess anschließend automatisiert werden soll. Die Dokumentation kann in diesem Fall als eine „Richtschnur“ genutzt werden.

Praxis Use-Case

In der vorherigen Phase *Datenquellenauswahl* wurden die notwendigen Datenquellen und Datenfelder festgelegt. In dieser Phase der *Datenaufbereitung* werden die folgenden drei Schritte auf Grundlage der selektierten Datenfelder durchgeführt:

1. **Datensatzselektion**: Relevante Datensätze werden anhand von Kriterien gefiltert und ausgewählt.
2. **Datenbereinigung**: Eine hohe Qualität der Daten wird erzielt, indem notwendige Transformationen, wie beispielsweise Normalisierungen, vorgenommen werden.
3. **Datenverknüpfung**: Datenquellen werden miteinander verknüpft.

Datensatzselektion

Alle benötigten Daten liegen nun in einer Form vor, sodass diese später für das Modell genutzt werden können. Damit die CLT berechnet werden kann, muss vorher definiert werden, welche Datensätze für die Vorhersage genutzt werden sollen. Es wird festgelegt, dass als Datengrundlage ausschließlich Nutzer berücksichtigt werden, die innerhalb des Monats März des letzten Jahres den ersten Kontakt mit immonet.de hatten. User, die mindestens zwei Jahre nicht immonet.de besucht haben, werden als neue User identifiziert, da davon ausgegangen werden kann, dass diese eine neue Customer Journey beginnen. Alle weiteren Aktivitäten der ausgewählten User werden für den Zeitraum des nächsten Jahres aus den Rohdaten extrahiert.

Schritt 1: Selektion aller `fullVisitorId` -Werte der Neukunden auf immonet.de: Eine Selektion ist mit einer SQL-Abfrage (Table 8) in Google BigQuery möglich. Durch die Einschränkung `totals.newVisits = 1` wird sichergestellt, dass nur IDs von Erstbesuchern (Erstbesucher oder inaktiv seit mindestens zwei Jahren) selektiert werden. Durch die Einschränkung `totals.hits > 1` werden alle Nutzer, die nur eine Seite von immonet.de aufgerufen haben (Bouncer), entfernt. `###ALLE_DATENQUELLEN_AUS _MAERZ###` wird durch die entsprechenden Datenquellen ersetzt.

Schritt 2: Selektion aller Ecommerce-Trackingdaten aus zwölf Monaten:

In diesem Schritt werden alle Datensätze eines Jahres aus Google Big Query selektiert, bei denen die `fullVisitorId` aus der Ergebnismenge der Erstbesucher stammt (siehe folgende SQL-Abfrage), d. h., es werden alle Ecommerce-Aktivitäten, die Erstbesucher aus März innerhalb von zwölf Monaten (März bis Februar) wahrgenommen haben, ausgewählt.

Table 8: Selektion aller Neukunden des Monats März

SQL-Abfrage	Ergebnis

SELECT
  fullVisitorId
FROM
  ###ALLE_DATENQUELLEN_AUS_MAERZ###
WHERE
  totals.newVisits = 1
  AND totals.hits > 1
GROUP BY
  fullVisitorId;
	1000009876151677559
	1000003566143674436
	1000009374848362837
	1000004736573909009
	…

SQL-Abfrage: Selektion der Ecommerce-Trackingdaten der Neukunden aus Màrz hinweg ùber ein Jahr

```
SELECT
  marchIds.fullVisitorId,
  visitStartTime,
  visitNumber,
  date,
  hits.page.pagePath,
  totals.hits,
  hits.item.productName,
  hits.item.transactionId,
  hits.hitNumber
FROM (
  SELECT
    fullVisitorId
  FROM
    IDS_AUS_MAERZ)AS marchIds
JOIN (
  SELECT
    fullVisitorId,
    visitStartTime,
    visitNumber,
    date,
    hits.page.pagePath,
    totals.hits,
    hits.item.productName,
    hits.item.transactionId,
    hits.hitNumber
  FROM
    TRACKING_DATEN
 ) AS totalData
ON
  totalData.fullVisitorId = marchIds.fullVisitorId
WHERE
  totalData.hits.item.productName IS NOT NULL
ORDER BY
  marchIds.fullVisitorId,
  date,
  visitStartTime,
  visitNumber;
```

Datenbereinigung

Da die durch das Tracking erhobenen Daten die Gütekriterien der Objektivität, Reliabilität und Validität erfüllen, werden die Daten aus Google Analytics in sich als konsistent und qualitativ angemessen angesehen. Die Objekt-Daten sind ebenfalls vollständig und konsistent. Sollten Datensätze mit `NULL`-Werten vorkommen, werden diese ausgeschlossen und entfernt. Da die vorliegenden Daten eventbasiert sind und das Tracking korrekt implementiert ist, sind für diese Analyse die entsprechenden Felder i. d. R. aber immer gesetzt.

„Ausreißer“ werden entfernt, damit die Analyseergebnisse nicht verfälscht werden. User, die ein Produkt zeitlich stark abweichend vom *Durchschnittsuser* innerhalb ihrer Customer Journey kaufen, werden als Ausreißer definiert und ausgeschlossen. Als Ausreißer werden also diejenigen User bezeichnet, deren Kaufentscheidung nicht innerhalb des 5 – 95 % Konfidenzintervalls liegt.

Schritt 3: Anreicherung der Daten um Objekt-Attribute:

Im letzten Schritt sollen die Tracking- und Objektdaten miteinander verbunden werden. Eine Verknüpfung ist nicht direkt möglich, da in den Daten aus Google Big-Query keine ObjektId enthalten ist. Das Feld **`hits.item.transactionId`** setzt sich aus den zwei Informationen ObjektId/Timestamp zusammen. Dieser **`STRING`**-Wert muss dahin gehend angepasst werden, dass der Slash und der darauffolgende Timestamp entfernt werden. Diese Transformation kann mithilfe eines Substring-SQL-Befehls, wie in dem folgenden SQL-Abfragenausschnitt, skizziert realisiert werden.

```
Transformation-Extraktion der ObjektId aus dem Feld hits.item.
transactionId:
[…]
substring(hits.item.transactionId,1,
instr(hits.item.transactionId,“/”) -1) ObjectId
[…]
```

Dieser Zeilen Code speichert die Zeichen aus **`hits.item.transactionId`**, beginnend

bei Position 1 bis hin zur Position des Slashs -1 in der neuen Datenspalte **`Objectid`**. Durch diese Transformation steht nun die ObjektId für eine Verknüpfung zur Verfügung.

Datenverknüpfung

Für jede Ecommerce-Conversion12 des Produkts Exposé (aus Google BigQuery) müssen die entsprechenden Objektdaten (aus der Objekt-Datenbank) an den Datensatz angefügt werden. Eine Verknüpfung der beiden Datenquellen ist über die, in beiden Quellen enthaltene ObjektId möglich. Durch diese Verknüpfung wird die Menge der Datenfelder, wie in Tab. 5, erweitert.

Table 9: Datenfelder der verknüpften Tracking- und Objektdaten

Datenfeld	Datentyp	Quelle	Beschreibung
`ObjectId`	**`INTEGER`**	BigQuery/ObjectDB	Eindeutige Objekt-ID
`fullVisitorId`	**`STRING`**	BigQuery	Eindeutige Kunden-ID
`visitStartTime`	**`INTEGER`**	BigQuery	Timestamp
`date`	**`STRING`**	BigQuery	Datum der Sitzung (JJJJMMTT)
`hits.page. pagePath`	**`STRING`**	BigQuery	Seiten-URL
`totals.hits`	**`INTEGER`**	BigQuery	Anzahl der Hits innerhalb einer Session
`hits.item. productName`	**`STRING`**	BigQuery	Produktname
`hits.item. transactionId`	**`STRING`**	BigQuery	ID der Ecommerce-Transaktion
`marketingType`	**`INTEGER`**	ObjectDB	Miete/Kauf
`parentCat`	**`INTEGER`**	ObjectDB	Haus/Wohnung

Legende: **`STRING`**: Hierbei handelt es sich um eine Zeichenkette, bestehend aus Buchstaben und/oder Zahlen und/oder Sonderzeichen

`INTEGER`: Ein **`INTEGER`** ist eine Ganzzahl

Phase IV: Modellierung

Die Phase des Modellierens wird im MAP in vier Schritte unterteilt. Die Auswahl der Modellierungstechnik (I) erfolgt anhand der Zielstellung bzw. des Geschäftsproblems und der Dateneigenschaften (Liao et al. 2012). Anschließend werden Testvorgänge erstellt (II). In Schritt drei wird mindestens ein Modell erstellt (III). Laut Ngai sind die verbreitetsten Analysemethoden die Assoziations-, Klassifikations-, Cluster- sowie Regressionsanalyse (Ngai et al. 2009), aber auch deskriptive Analysen sind laut Haumer relevant (Haumer 2015). Abschließend werden in diesem Schritt die Modelle bewertet (IV). Hier ist eine intensive Zusammenarbeit des Data-Scientists mit der Fachabteilung sehr empfehlenswert (Shearer 2000).

Mithilfe von verschiedenen Modellen können im Rahmen von Big Data Kundensegmentierungen und -profilerstellungen, ortsgebundene Werbung, Analyse von Gruppendynamik, Erforschung der unternehmenseigenen Preispolitik, Wettbewerberanalyse, Marktübersicht, Produktreputationsmanagement, Analyse der Marketingmaßnahmen oder auch Empfehlungssystemen realisiert werden (Fan et al. 2015).

Laut Tendick et al. sind statistische Analysen weitgehend auf menschliche Aktivitäten – sowohl bzgl. des Verständnisses von Daten, als auch bzgl. des Prozesses und der Verarbeitung – angewiesen. Dieses Vorgehen ist für Situationen, in denen schnelle Maßnahmen oder gar Echtzeitanalysen notwendig sind, nicht geeignet (Tendick et al. 2016). Daher werden in dem MAP zunächst Modelle und Dateneinsichten durch den Fachbereich bzw. dem Fachbereich nahestehende Analysten gewonnen, um diese später zu automatisieren. Sharma merkt an, dass traditionelle Verfahren und Tools häufig nicht mächtig genug sind, um Big Data in seiner semi- und unstrukturierten komplexen Natur zu beherrschen. Hierfür existieren spezielle Software-Frameworks, welche Algorithmen und Techniken für Data Mining, Predictive Analytics sowie statistische Analysen bereitstellen. Des Weiteren besitzen etablierte Big-Data-Tools die Möglichkeit der Echtzeit-Datenvisualisierung (Sharma 2016).

Aktuelle Softwaretools bieten eine vergleichbare Qualität in Bezug auf Grundfunktionen; unterscheiden sich in Spezialanforderungen allerdings erheblich. An dieser Stelle wird darauf hingewiesen, dass für die richtige Wahl eines Softwaretools eine fundierte Kenntnis der Daten und Datenstrukturen unabdingbar ist (Judah et al. 2017).

Data-Scientists neigen dazu, die Güte der Modelle ständig zu verfeinern und die Fehlerwahrscheinlichkeit weiter zu minimieren. Im Fokus des MAP stehen jedoch die Machbarkeit und Wirtschaftlichkeit, nicht das *perfekte Modell*. Die Durchführung des MAP kann durchaus als Ziel haben, ein bestehendes Modell zu verbessern. Im Rahmen von Big Data sind aber vor allem neue Ansätze und Einsichten gefragt, welche monetarisiert werden können.

Praxis Use-Case

Alle relevanten Daten stehen nun für die Analyse bereinigt und verknüpft zur Verfügung.

Schritt 1: User-Segmentierung

Im ersten Schritt werden die User(-Sessions) den vier Produkten (Segmenten) *Haus/kaufen*, *Haus/mieten*, *Wohnung/kaufen* und *Wohnung/mieten* zugeordnet. Die Zuordnung erfolgt anhand der Häufigkeit der Vorkommen der Wertepaare **`marketingtype`** (kaufen/mieten) und **`parentcat`** (Haus/Wohnung).

Beispiel:

Ein User hat beispielsweise nachfolgenden Objekttypen gesucht:

- 10x Haus/kaufen
- 4x Haus/mieten
- 1x Wohnung/kaufen
- 2x Wohnung/mieten

Daraus ergeben sich folgende Werte

Parentcat		**Marketingtype**	
Haus	Wohnung	**Kaufen**	Mieten
14	3	**11**	6

Anhand dieser Werte wird der User etwa dem Segment Haus/kaufen zugeordnet. User, die nicht eindeutig zugeordnet werden können, werden einer weiteren Gruppe Rest zugewiesen, welche im Rahmen der folgenden Betrachtung nicht berücksichtigt wird. Dadurch reduziert sich die Fallzahl um weniger als 1 %.

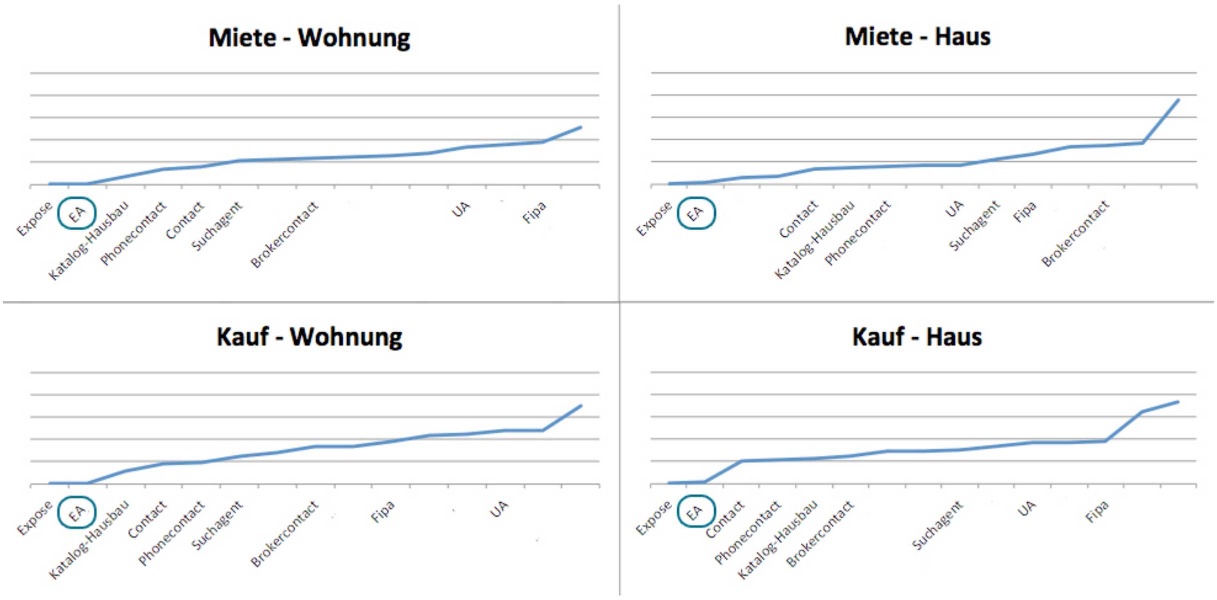


Figure 7: Ergebnis der vier Segmente. *(Quelle: eigene Darstellung)*

Schritt 2: Analyse der einzelnen Segmente

Im zweiten Schritt werden die Customer-Journeys pro Segment detaillierter analysiert, d. h. es wird überprüft, zu welchem Zeitpunkt ein User welches After-Sales-Produkt gekauft hat. Für jedes Produkt wird als erster Schritt der Mittelwert der Kaufzeitpunkte gebildet. Die After-Sales-Produkte können pro Segment dann in eine erste Reihenfolge (entspricht der CLT) gebracht werden. Abbildung Figure 7 zeigt die Mittelwerte aller Produkte für das jeweilige Segment. Auf der Y-Achse sind die Tage dargestellt, ab wann ein Produkt (X-Achse) für einen User des Segments relevant ist.

Phase V: Modellevaluierung

In der Phase der Modellevaluation werden die – aus der vorigen Phase resultierenden – Modelle überprüft (Shearer 2000). Die – aus der Sicht der Analysten – technisch einwandfreien Modelle sind nun hinsichtlich der fachlichen Zielsetzung zu inspizieren. Des Weiteren ist das Vorgehen zu validieren, welches zu der Erstellung der Modelle geführt hat. Es werden demnach auch noch einmal die Datenquellen und die Datenaufbereitung überprüft. Es wird also validiert, ob bisher alles Relevante bedacht worden ist, um mit den Modellen das definierte Ziel zu erreichen.

Praxis Use-Case

Auffällig ist, dass bei allen vier Segmenten das Produkt Einzelanzeige (EA) an zweiter Stelle erscheint. EA ist ein für Privatpersonen erstelltes Produkt, die ein Objekt anbieten möchten. An dieser Stelle wird die Datengrundlage nochmals angepasst. Die Auswertung lt. Zielvorgabe berücksichtigt den suchenden, nicht den anbietenden Nutzer mit Kaufinteresse. Alle Sessions, die das Produkt EA enthalten, werden vor der Analyse aussortiert. Der MAP sieht für diesen Fall einen Rücksprung zur Phase III (Datenaufbereitung) vor.

Nachdem die Daten bereinigt und alle Sessions, die das Produkt EA enthalten, aussortiert sind, wird die Analyse mit dem aktualisierten Datenbestand erneut durchgeführt. Figure 8 zeigt die Ergebnisse.

Mittels einer Varianzanalyse wird geprüft, ob sich die Mittelwerte der einzelnen After-Sales-Produkte zwischen den vier Segmenten signifikant voneinander unterscheiden. Hierfür wird mithilfe von R, einem open-source-Programm für statistische Auswertungen und grafische Darstellung, für jedes einzelne Produkt eine Varianzanalyse durchgeführt. Diese wird auch als ANOVA (analysis of variance) bezeichnet.

Alternativ zu R sind Python, RapidMiner und SQL, aber auch Microsoft Excel als Modellierungstools und -sprachen zu erwähnen.

Es zeigt sich, dass sich die Mittelwerte zwischen den Segmenten für den ersten Contact, Phonecontact, Suchagent und Suchanzeige-Privat statistisch signifikant voneinander unterscheiden. Die After-Sales-Produkte weisen jedoch keine signifikanten Differenzen auf. Dies deutet auf eine zu geringe Fallzahl oder auf eine ähnliche CLT der After-Sales-Produkte hin.

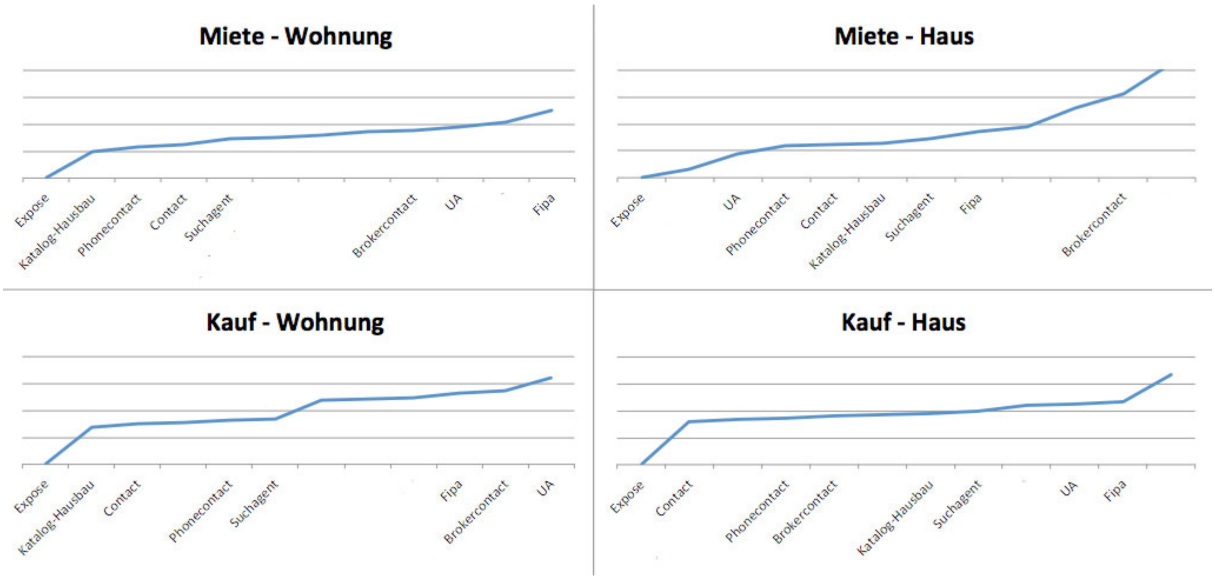


Figure 8: Ergebnis der vier Segmente mit bereinigter Datengrundlage. *(Quelle: eigene Darstellung)*

Phase VI: Handlungsempfehlungen

Die letzte Phase des MAP beinhaltet die Formulierung der Handlungsempfehlungen. Eine häufige Handlungsempfehlung stellt die Implementierung bzw. Automatisierung der Modelle dar. Eine Handlungsempfehlung ist ein Bericht, welcher den durchgeführten Prozess dokumentiert, Resultate – wie zum Beispiel Modelle, Erkenntnisse oder entstandene Artefakte – beschreibt und visualisiert und auf deren Basis eine Empfehlung für das weitere Vorgehen ausspricht. Die Handlungen werden dabei mit Fokus auf den Return on Investment (ROI) empfohlen. Alternativen werden beschrieben und bewertet und mögliche Auswirkungen – sowohl strategischer, als auch operativer Herkunft – auf das bestehende Geschäft erläutert.

Im Rahmen von Big Data Analytics ist vor allem die Visualisierung von Prozessen oder Resultaten im Hinblick auf die Beherrschung der Komplexität relevant. Durch Visualisierungstechniken

und -formen wird die Entscheidungsqualität erhöht und Wissen zugänglich, welches ohne Visualisierung nicht ersichtlich ist (Schoeneberg and Pein 2014).

Praxis Use-Case

Das beschriebene Modell basiert auf der Bildung von Mittelwerten. Es ermöglicht einen guten Überblick der CLT und ist ein Indikator dafür, dass eine verfeinerte Analyse an dieser Stelle mit großer Wahrscheinlichkeit gewinnbringend sein wird. Dies ist darin begründet, dass pro Segment ein unterschiedliches Nutzungsverhalten festgestellt wurde, welches aber nicht durch eine statistische Signifikanz belegt ist.

Mittelwerte sind für eine statistische Wahrscheinlichkeit, zu welchem Zeitpunkt ein Produkt für einen Kunden relevant ist, nicht exakt. In einer weiteren Iteration des MAP wird daher eine sog. Event-Analyse durchgeführt. Das dabei entstehende Modell berechnet eine statistische Wahrscheinlichkeit pro Tag für jedes After-Sales-Produkt. Die Ergebnisse dieses Modells sind für den Einsatz von Online-Marketing-Maßnahmen, wie beispielsweise E-Mail-Kampagnen, besser geeignet, da die Ergebnisse nicht durch die Mittelwertbildung, bei der Informationen verloren gehen, verfälscht werden.

MAP sieht in dieser Phase die Formulierung konkreter Handlungsempfehlungen vor. Folgende Ziele werden vor Beginn der Analyse in der ersten Phase des MAP definiert:

1. Erstellung einer CLT für die vier Kernprodukte: OK
2. Erstellung eines Forcast-Modells: OFFEN
3. Integration des Forcast-Modells: OFFEN

Für die in diesem Use-Case initial durchgeführte Iteration lautet die Handlungsempfehlung eine erneute Iteration des MAP, da die Ziele 2 und 3 noch nicht erreicht werden. In der zusätzlichen Iteration wird eine Event-Analyse mit einer größeren Stichprobe und weiteren Datenfeldern durchgeführt, die die Erstellung eines Forecast-Modells und dessen Implementierung ermöglicht. Die Durchführung dieser Phase ist jedoch noch nicht das Ende des gesamten Projekts. Eine Implementierung des MAP und die damit einhergehende Verfeinerung der Modellgüte durch weitere Iterationen dienen dazu, das eigene Produkt von denen der Mitbewerber abzusetzen.

Implementierung

Die Implementierung zur automatischen Ausführung der zuvor entwickelten Modelle stellt im Rahmen von Big Data in der Marketing-Intelligence den logischen Schluss dar. Durch das automatisierte Ausführen kann eine permanente Neuberechnung erfolgen. Eine ständige Optimierung der Modellparameter vermag so zu einer kontinuierlichen Steigerung des ROI zu führen.

Laut Sharma existieren zwei unterschiedliche Arten der automatisierten Datenverarbeitung: Streaming und Batch-Verarbeitung. Während beim Streaming die Daten in Echtzeit verarbeitet werden, werden bei der Batch-Verarbeitung Daten in langlaufenden Batch-Aufträgen – oft skaliert über große Servercluster – abgearbeitet (Sharma 2016).

Streaming erfolgt durch direkte Verarbeitung von Daten- oder Ereignisströmen. Diese Datenverarbeitung ist nur sinnvoll, wenn die Daten schnell verarbeitet werden sollen und ein Resultat umgehend oder durchgehend notwendig ist. Ein Beispiel hierfür ist die Darstellung individueller Web-Displays.

Gemäß Chen et al. implementieren aktuelle Systeme für die Ausführung einer Batch-Verarbeitung das MapReduce-Framework (Chen et al. 2012). MapReduce ist ein höchst skalierbares Framework zur parallelen Datenverarbeitung. Es ist beispielsweise eine der Kernkomponenten des Hadoop-Systems (Sharma 2016). Laut Bello-Orgaz et al. stellt MapReduce eine exzellente Technik dar, um große Mengen von Daten zu verarbeiten. Voraussetzung für das schnellere Verarbeiten von großen Datenmengen durch MapReduce ist, dass die Algorithmen auf kleinen Mengen der Daten parallel angewendet werden können (Bello-Orgaz et al. 2016). Batch-Verarbeitung ist zu wählen, wenn die Datenanalyse nicht in Real- oder Neartime, sondern zu definierten Zeitpunkten, z. B. täglich oder monatlich, durchgeführt werden soll. Die Datenerhebung ist hiervon nicht betroffen. Es entstehen keine Lücken in der Aufzeichnung. Ein Beispiel für die Batch-Verarbeitung ist etwa der Bericht der täglichen/monatlichen Verkaufszahlen.

Fazit

Das Ziel von Big-Data-Analytics ist es, große Datenmengen möglichst in Echtzeit zu analysieren und zu interpretieren, um dem Business Informationen zur Generierung eines Wettbewerbsvorteils zu liefern. Dabei zeigen aktuelle Studien, dass es bei 85 % der mittelständischen Unternehmen derzeit an ausreichend qualifiziertem Personal fehlt, Projekte dieser Art durchzuführen (Haumer 2015).

Das Marketing-Controlling, dessen Aufgabe die marktorientierte Unternehmensführung auf Basis von Daten ist, muss sich neuen Herausforderungen stellen. Durch die Digitalisierung steigen die diesbezüglich fachlichen Anforderungen in qualitativer wie quantitativer Sicht. Darüber hinaus sind unternehmensweit Prozess- und Organisationsstrukturen zu implementieren, die die Digitalisierung unterstützen und begünstigen. Das Marketing-Controlling selbst steht dabei vor der Herausforderung, unter Zuhilfenahme der Digitalisierung mittels Unternehmens- und Wettbewerbsdaten strategische wie operative Wettbewerbsvorteile zu generieren.

Der hierfür entwickelte MAP stellt ein Framework dar, um Big Data Analytics Projekte für das Marketing-Controlling erfolgreich umzusetzen. Das entwickelte Vorgehensmodell basiert auf prozessualen wie agilen Komponenten und folgt einer strukturierten Vorgehensweise in sechs Phasen. Durch den klaren Aufbau kann es damit leicht analysiert und auf das eigene Business adaptiert werden.

Der dargestellte durchgängige Best Practice Use-Case je Phase bietet Analysten wie Marketing-Controllern zusätzlich wichtige Hinweise zur Umsetzung des MAP. Die dargestellten Aspekte zur fachlichen wie technischen Implementierung stellen eine wertvolle Basis zum disziplinübergreifenden Austausch dar. Sie bieten darüber hinaus Ansätze zum Transfer der vorgestellten Inhalte auf das eigene Business dar und können als Grundlage zur Implementierung in die eigene Organisation dienen.

Der aus wissenschaftlicher, praktischer und technischer Sicht dargestellte MAP löst damit die Grenzen zwischen dem Fachbereich Marketing und Informationstechnologie auf. Ein in der Zukunft erfolgreiches Marketing Analytics ist mehr denn je von beiden Kompetenzen abhängig. Durch die voranschreitende Digitalisierung hat eine Verschmelzung dieser Disziplinen längst begonnen.

Literatur

Bello-Orgaz, Gema; Jung, Jason J.; Camacho, David (2016): Social Big Data: Recent Achievements and New Challenges. In Inf. Fusion 28 (1), pp. 45–59. DOI: 10.1016/j.inffus.2015.08.005.

Borden, Neil H. (1964): The Concept of the Marketing Mix. In Marketing management and administrative action, pp. 31–40.

Chen, Yanpei; Alspaugh, Sara; Katz, Randy (2012): Interactive Analytical Processing in Big Data Systems. A Cross-Industry Study of MapReduce Workloads. In Proc. VLDB Endow. 5 (12), pp. 1802–1813. DOI: 10.14778/2367502.2367519.

Dasu, Tamraparni; Johnson, Theodore (2003): Exploratory Data Mining and Data Cleaning. Hoboken, NJ: Wiley-Interscience (Wiley series in probability and statistics). Available online at http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10299281.

Fan, Shaokun; Lau, Raymond Y. K.; Zhao, Leon J. (2015): Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix. In Big Data Research 2 (1), pp. 28–32. DOI: 10.1016/j.bdr.2015.02.006.

Felden, Carsten (2012): Datenqualitätsmanagement. Enzyklopädie der Wirtschaftsinformatik. Edited by Norbert Gronau, Jörg Becker, Karl Kurbel, Elmar Sinz, Leena Suhl. Available online at http://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/daten-wissen/Datenmanagement/Datenmanagement--Konzepte-des/Datenqualitatsmanagement, updated on 10/31/2012, checked on April 5th, 2016.

Goi, Chai L. (2009): A Review of Marketing Mix. 4Ps or More? In IJMS 1 (1). DOI: 10.5539/ijms.v1n1p2.

Granville, Vincent (2015): Reducing Data Cleansing Time to Get Actionable Insights Faster. Available online at https://www.datasciencecentral.com/profiles/blogs/reducing-data-cleansing-time-to-get-actionable-insights-faster, updated on June 15th, 2015, checked on June 2nd, 2016.

Haumer, Florian (2015): Was Marketingcontrolling und Big Data Analytics gemeinsam haben. Available online at http://www.sputnika.de/dresden/magazin/details/article/was-marketingcontrolling-und-big-data-analytics-gemeinsam-haben/, updated on Mai 5th, 2015, checked on June 6th, 2016.

Jacobs, Adam (2009): The Pathologies of Big Data. In Commun. ACM 52 (8), p. 36. DOI: 10.1145/1536616.1536632.

Judah, Saul; Selvage, Mei; Jain, Ankush (2017): Critical Capabilities for Data Quality Tools. In Gartner Report. Available online at https://www.gartner.com/doc/3835263/critical-capabilities-data-quality-tools, checked on August 11th, 2018.

Judd, Vaughan C. (1987): Differentiate with the 5th P: People. In Industrial Marketing Management 16 (4), pp. 241–247. DOI: 10.1016/0019-8501(87)90032-0.

Kimball, Ralph; Merz, Richard (2000): The Data Webhouse Toolkit. Building the Web-Enabled Data Warehouse. New York: Wiley. Available online at http://www.loc.gov/catdir/bios/wiley042/99055652.html.

Liao, Shu-Hsien; Chu, Pei-Hui; Hsiao, Pei-Yuan (2012): Data Mining Techniques and Applications – A Decade Review from 2000 to 2011. In Expert Systems with Applications 39 (12), pp. 11303–11311. DOI: 10.1016/j.eswa.2012.02.063.

Malgara, Andrea (2014): Big Data und Attribution Modelling: Total Marketing Controlling in Echtzeit (New Business, 44/2014). Available online at http://www.mediaplus.com/de/presse-detail/big-data-und-attribution-modelling-total-marketing-controlling-in-echtzeit.html, checked on March 24th, 2016.

McCarthy, Jerome E. (1964): Basic Marketing. Homewood, IL, USA: Irwin.

Ngai, Eric W.T.; Xiu, Li; Chau, Dorothy C. K. (2009): Application of Data Mining Techniques in Customer Relationship Management. A Literature Review and Classification. In Expert Systems with Applications 36 (2), pp. 2592–2602. DOI: 10.1016/j.eswa.2008.02.021.

Piatetsky, Gregory (2014): CRISP-DM, Still the Top Methodology for Analytics, Data Mining, or Data Science Projects. KDnuggets. Available online at https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html, checked on August 3rd, 2018.

Schoeneberg, Klaus-Peter; Pein, Jennifer (2014): Entscheidungsfindung mit Big Data. Einsatz fortschrittlicher Visualisierungsmöglichkeiten zur Komplexitätsbeherrschung betriebswirtschaftlicher Sachverhalte im Unternehmen. In Klaus-Peter Schoeneberg (Ed.): Komplexitätsmanagement in Unternehmen. Herausforderungen im Umgang mit Dynamik, Unsicherheit und Komplexität meistern. Wiesbaden: Springer Gabler, pp. 309–354.

Schoeneberg, Klaus-Peter; Zerres, Christopher; Frass, Alexander; Igelbrink, Jörg (2016): Textmining. Markenführung mittels Social Media Analytics. In Michael Lang (Ed.): Business Intelligence erfolgreich umsetzen. Von der Technologie zum Geschäftserfolg. 1. Auflage 2016, neue Ausgabe. Düsseldorf: Symposion Publishing, pp. 75–99.

Sharma, Sugam (2016): Expanded Cloud Plumes Hiding Big Data Ecosystem. In Future Gener Comput Syst 59, pp. 63–92. DOI: 10.1016/j.future.2016.01.003.

Shearer, Colin (2000): The CRISP-DM Model: The New Blueprint for Data Mining. In Journal of data warehousing 5 (4), pp. 13–22.

Tendick, Patrick H.; Denby, Lorraine; Ju, Wen-Hua (2016): Statistical Methods for Complex Event Processing and Real Time Decision Making. In Wiley Interdiscip. Rev. Comput. Stat. 8 (1), pp. 5–26. DOI: 10.1002/wics.1372.

5 Publication 2 Attribution Modelling in an Omni-Channel Environment - New Requirements and Specifications from a Practical Perspective

Attribution Modelling in an Omni-Channel Environment **New Requirements and Specifications from a Practical Perspective**	
DOI	Not assigned yet
Format	Journal article
Journal	International Journal of Electronic Marketing and Retailing
Language	English
Status	Accepted (May 3rd, 2018)
Abstract	How much am I, the customer, currently worth to a company? The answer to this question is very important for the marketing team but difficult to obtain. In an omni-channel environment, the degree of complexity for answering this question has reached a new level. Based on a structured literature research process, existing dynamic budget allocation approaches are identified and evaluated regarding their applicability in an omni-channel environment. For the evaluation process of these identified models, assessment criteria are needed. Structured interviews are conducted with experts in the field of attribution to formulate evaluation criteria, which are being used to evaluate the applicability of the defined models. This article describes why existing dynamic attribution models are not suitable for an omni-channel environment and what features need to be part of a new future-ensured omni-channel attribution model. The authors conclude by presenting questions for future research in the field of dynamic attribution.
Keywords	*omni-channel attribution, practical requirements for omni-channel attribution, online advertising, dynamic attribution, dynamic attribution model, omni-channel attribution modelling, multi-touch attribution (MTA), budget allocation, data-driven attribution modelling, real-world attribution*

Introduction

The objective of this article is to identify requirements and specifications towards attribution modelling in an omni-channel environment from a practical perspective and to analyse which abilities existing models already fulfil.

This article aims at offering the following contributions to the academic science community. (1) Identify existing dynamic attribution models in the science community based on a structured literature research process. (2) Present criteria for dynamic attribution models in an omni-channel environment based on expert interviews. (3) Evaluate identified models based on the resulting criteria from the expert interviews. (4) Formulate and define exigencies, research fields and research questions for further research.

This article has been organized the following way. The introduction draws the objectives and the theoretical background. Additionally, the state of art is described in the research areas of *attribution*, *marketing performance*, *budget allocation* and *dynamic attribution modelling*. The latter one is realized by a structured literature research process which identifies existing attribution approaches. The literature research process is also placed in the introduction since it is not part of the primary research. Its results enable the current study to fill the research gap whether existing attribution models are applicable in an omni-channel environment, or not. In the chapter research methodology, the approach of how to identify requirements and specifications from a practical perspective is explained in detail. Both, the results from the literature research process and the results from the experts' interviews are presented in the results section followed by a discussion including an outlook for areas of further research. This article finalizes by a conclusion.

Theoretical Background

On his way to purchase the customer leaves, a very detailed and granular footprint, which is traced by new technologies to support companies. Important trace data is not generated anymore by the customer on a firm's website alone (online). Offline touchpoints such as a purchase in a local store or a call at the customer's support desk are also relevant information about the ways a customer gets in touch with a company. Interaction with customers indicates that they leave very detailed usage data in various places, using company-specific apps and other marketing channels, e.g., social networks, display advertisement, paid search, and mailings. Sources of user-generated data are decentralized by default if different systems are involved to track user behaviour, e.g., CRM, website tracking tool and third-party services. Offering different, independent channels to communicate and interact with one's customers is a widely applied strategic approach (Econsultancy 2015). This strategy is called a multi-channel approach where responsible channel marketers often act as independent departments mainly using their channel generated data (Neslin, Shankar 2009). The following example outlines weaknesses of a multi-channel approach.

Responding to a company-initiated survey, a customer may state that he/she is not interested in a particular product category (e.g., sports shoes). A few days later this customer receives a newsletter promoting this specific product category in which he/she is not interested.

This situation shows that this particular information is kept in the survey results and not transmitted to the email/newsletter channel since each channel department still uses its own generated data and logic to perform channel actions.

The next more advanced strategic approach to communicating with customers is called omni-channel strategy (Camiade 2013). From a data-driven perspective, the most relevant difference between a multi-channel and omni-channel setup is a centralized data source, containing or connecting the data from various channels and sources. Figure 9 illustrates the different data streams in a multi-channel and omni-channel configuration. From a customer perspective, a seamless communication using different channels is feasible in an omni-channel environment. This approach is realized by adding one logic layer connecting all channels. All customer actions are filtered through this layer.

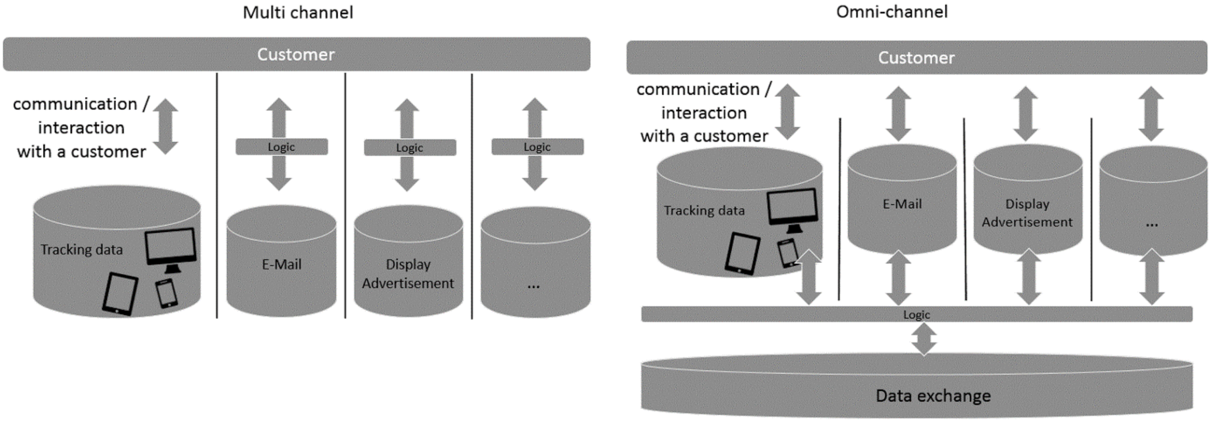


Figure 9: Data flow in a multi-channel and an omni-channel environment. Based on Paccard (2017)

Deriving Towards Omni-Channel Marketing

In the following paragraphs, the development towards an omni-channel marketing approach is outlined and explained. Single-channel marketing implies that a company or a brand only offers one single channel for interacting with a customer and vice versa. A further developed presence of a company or brand in a market is a so-called cross-channel or multi-channel setup. The phrase "cross"-channel marketing is derived from the lat. term Crux, meaning to go across. Multi-channel is derived from the word *multus*, which means multiple or many. The multi-channel approach comprises an interaction between customers and a company through different independent channels such as social media, paid search, banner advertisement, mailings, etc. This strategic approach is still widely spread.

Verhoef et al. (2015) describe the necessary shift from multi-channel to omni-channel in a retailing context. The term “Omni”-channel (lat. omnis), translated “all,” signifies an even more complicated approach of how a firm or a brand needs to interact with their customers. The main difference between multi- and omni-channel is the elimination of borders between channels toward a seamless experience through integrated channels. Compulsorily, all channels need to connect or join their generated user data and channel data in one destination (e.g., DMP data management platform) to achieve such a seamless experience. All customer interaction data is stored in or connected to this destination and can be used to predict how to get or stay in contact with a customer and how to motivate a user to buy a product or service.

It is a strategic management decision to move established independent channel departments within a company to a holistic omni-channel structure. To be able to offer a seamless user experience across all channels, information about a customer such as their needs, their attributes, or the state within the buying process needs to be accessible and processed by every channel offered.

Already in 2006 Neslin et al. describe a customer data integration approach as the ideal data setup in a multi-channel context. Today, in a market application data integration is becoming achievable and comprises the basis for an omni-channel approach.

A Brief Background on Budget Allocation

To understand the need for dynamic budget allocation approaches, in the following paragraphs, different budget allocation strategies are presented and evaluated with respect to an application in an omni-channel environment.

A rule, a set of rules, or an algorithmic approach are the foundation to determining how much credit or budget is assigned to a certain source, e.g. channel, for conversions, leads, or sales. These are basic specifications in an attribution model. In marketing publications, attribution models are often distinguished into two categories: *static* attribution models and *dynamic* attribution models (Anderl et al. 2016a; Li and Kannan 2014; Shao and Li 2011).

Jayawardane et al. (2015) distinguish between three categories, instead of two: *simplistic* models, *rule-based* models, and *algorithmic* models. The category of static attribution models introduced in the preceding paragraph is split up into simple models and rule-based models. The two groups of dynamic and algorithmic approaches are congruent (see Figure 10).

Models belonging to the simplistic category assign the complete conversion credit to a single touch point (Google Inc. 2017; Jayawardane et al. 2015) such as

- *last click/last interaction* which assigns 100% credit to the last touchpoint,
- *last non-direct click* which assigns 100% credit to the channel that the customer came from before converting (direct traffic is ignored) or
- *the first interaction* which assigns 100% credit to the first touch point.

Models contained in the category *rule-based* are known as heuristic models. By defining a static rule to spread the credit to all touch points which lead to a conversion, these models address the essential limitation of simplistic models (Jayawardane et al. 2015; Uniquedigital 2012; Lee 2010). Examples are

- *linear* which assigns the same amount of credit to all channels, or
- *position,* which assigns 40% to the first touch point, 40% to the last touch point and 20% equally to all touch points in between.

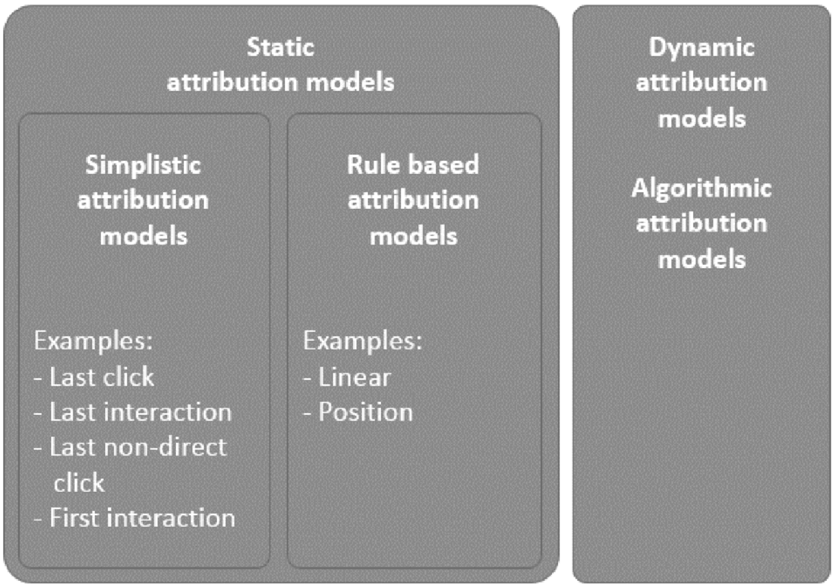


Figure 10: Delimitation: Static, Simplistic, Rule-based, Dynamic and Algorithmic attribution models

All the models mentioned above assign credit by static rules and neglect individual user behaviour. Whole user sessions which do not lead to a conversion are disregarded as well (Petersen et al. 2009). Although these *static* and *rule-based* models are inaccurate and their results are uncertain and questionable regarding representing the reality correctly, the *last click/last interaction* model, for example, is still widely used (eMarketer 2016). Reasons, why these attribution approaches are still so widely in use, are evident: Joining data from different channels and channel vendors is a challenging task.

Research on cross-channel customer satisfaction, the spillover-, and the carryover-effect (Nottorf and Funk 2013; Rutz and Bucklin 2011; Hammerschmidt et al. 2015) proves that there is an influence between channels which is neglected by *static* (*simplistic* and *rule-based*) attribution modelling approaches.

Because of their characteristics (ignoring the influence between channels and non-converting user sessions) *simpl*e and *rule-based* attribution approaches are not applicable to an omni-channel environment if the results need to represent the reality accurately.

In the past decade, different authors tried to fill this research gap by presenting various dynamic attribution models using different statistical approaches. A list of relevant models is discussed in the current study.

State of the Art

Online marketing performance and the always implied question of efficient budget allocation are widely studied areas. The search request for "online marketing performance" and a restriction on the timespan to “2010-2017” utilizing the web of science offered by Thomson Reuters was conducted on March 24th, 2017. The result set contained almost 1900 contributions since 2010. The following three primary research areas *channel performance, challenges of marketing structures* and *customer satisfaction* are identified and described for budget allocation in an omni-channel environment.

Channel Performance

There are various articles focused towards the field of channel or cross-channel performance, including the aspect of spending marketing budget in a more effective way (Archak et al. 2012; Dinner et al. 2011; Gallino and Moreno 2014; Haan et al. 2016; Joo et al. 2014; Olbrich and Schultz 2014; Voorveld 2011; Wiesel et al. 2011). Dinner et al. (2011) analyse cross-channel effects of digital vs. traditional advertisements while Gallino and Moreno (2014) analyse the impact of shared availability of inventory information and Haan et al. (2016) compare different forms of advertising in their long-term effectiveness. Joo et al. (2014) focus on television and search advertisement and identify a need for considering cross-media effects during planning, executing and evaluating campaigns. Within the context of social networks, Alon et al. (2012) propose different models to capture influences. Aspects of channel migration (Ackermann and Wangenheim 2014; Polo and Sese 2016; Woo et al. 2015) or mobile advertising (Grewal et al. 2016) are analysed as well.

Effects between channels such as the spill-over and carry-over effects (Leone 1995; Nottorf and Funk 2013; Rutz and Bucklin 2011) are analysed to improve performance and reduce costs. Archak et al. (2012) focus on positive spill-over effects. Furthermore, the performance of different advertisers is analysed as well (Berman 2015). The interactive effects of online and offline activities and their interaction (Wiesel et al. 2011; Naik and Peters 2009) are well observed and studied.

Challenges of marketing structures

Within marketing structures (e.g. offered channels by a company) the complexity itself and how to reduce it is another area of research (Anderl et al. 2016b; Feit et al. 2013; Lewis and Rao 2013). Saldanha et al. (2013) have registered a US patent for attributing conversion credit for transactions by users.

Research in retail is also an important research area. Verhoef et al. (2015) show how to shift from multi-channel retailing to omni-channel retailing in general.

Customer Satisfaction

Customer satisfaction (Hammerschmidt et al. 2015) and channel conflicts (Rusko 2015) in an omni-channel environment are also aspects which have an influence on the online marketing performance and are an aspect of research.

Dynamic Attribution Approaches

Marketing performance and the implicit question of how to allocate the marketing budget in an efficient and effective way is already a complex and important question in a multi-channel environment (Dalessandro et al. 2012; Xu et al. 2014). Allocating an appropriate credit for a certain customer action to each marketing touch point across all online and offline channels is the definition of attribution modelling from a practitioners' perspective (Moffett et al. 2014). This challenging question gains complexity in an omni-channel environment due to more channels, it's linking, and data sources, containing detailed user interaction data. The Marketing Science Institute announced attribution modelling to be the number one priority research area in the years of 2016 to 2018 (MSI 2016). Marketers need to decide how to allocate the marketing budget across the offered channels. Therefore, attribution models which divide up the given budget and assign it to a source e.g. marketing channel in ratio to its performance, are used to support marketers.

To the best of the author's knowledge, there is no other publication dealing either with the topic of comparing dynamic attribution models or evaluating them with respect to omni-channel requirements. There is only one related paper in which the authors classify dynamic attribution models from a statistical perspective (Jayawardane et al. 2015). A structured and comprehensive analysis is not within the scope of their article. Based on seven identifiable model features, a classification of the statistical approach within a model has been analysed. This paper concentrates on the mathematical and statistical approach.

In their introduction to the special section, a brief overview of dynamic attribution approaches is given by Kannan et al. (2016). Future research areas from a scientific perspective are introduced as well. These research areas are formulated only from a scientific perspective. At this point, a structured research process is necessary to build upon a scientific reliable foundation. This article offers a holistic approach. The presented dynamic attribution models by Kannan et al. (2016) and Jayawardane et al. (2015) were not identified through a structured research approach and cannot be applied for this research because the authors did not analyse that their results present a holistic overview of dynamic attribution models.

In order to proceed with a content analysis and tackle the mean issue of our research, we situate the subject and its boundaries in a specific area: the omni-channel environment, to concentrate later on the research topic, the analysis of requirements and specifications for attribution modelling in an omni-channel context. As already described it is one goal to identify existing literature containing dynamic attribution approaches to state the status quo. To identify all relevant existing dynamic attribution approaches, a protocol-driven systematic literature research methodology for the research process in accordance with Greenhalgh and Peacock (2005) and Webster and Watson (2002) is chosen to ensure a holistic output. Webster and Watson (2002) define a literature review to be concept-centric and not author-centric because the latter fails to synthesize the literature.

To ensure a comprehensible and understandable research process, all papers which need to be evaluated were listed in a spreadsheet table. Cronin et al. (2008) recommend a table holding process relevant information about topic-specific data and the article. The table holds, next to those process relevant information, columns such as title, abstract, no. of citations, DOI, topic, *channel count*, *is dynamic approach* and *uses cross-device data*. *Channel count* and *is dynamic approach* represent the two concepts which are described below.

The initial search strategy was defined at the beginning of the study. Its result was combined with a *forward snowballing* and *backward snowballing* approach (Webster and Watson 2002) letting the search strategy partly emerge as the investigation unfolds. This structured literature approach ensures a for this study relevant holistic result where the research process ends if no new relevant literature is found (Salipante et al. 1982).

Boundaries of the Literature Research

According to Bacharach (1989) and Whetten (1989), the following boundaries were set for the search. During the structured literature process, only those papers were selected which are published in or after 2010 and meet at least one of the following two concepts.

(1) A dynamic attribution model is defined (*is dynamic approach*). This is the case if the primary focus of the publication is to develop a dynamic attribution model. Papers which contain an analysis of an individual online marketing problem (e.g., mutual influence from two channels) are not considered.

(2) Alternatively, the publication focus comprises an analysis of the performance of at least two marketing channels (*channel count*).

Models established before 2010 are not relevant for omni-channel attribution because in 2010 the IDC Retail Insights Report first predicted a strong reliance on omni-channel for successful marketers in the following years (IDC 2010). Therefore, companies were not acting in an omni-channel context before 2010.

Publications from ranked journals, as well as conference papers and publications in books, are considered to obtain a comprehensive overview.

The search is not confined to a particular set of journals, research methodology or geographic region – it is the goal to get a broad overview of existing dynamic attribution models in science.

Executing the Literature Review Analysis

To identify all relevant dynamic attribution approaches the Social Sciences Citation Index (SSCI) was selected as a data foundation for the initial search. In September 2016 the SSCI was inquired with the initial search query. This query contained the following topic terms: *omni-channel attribution, dynamic attribution, multi-channel attribution, attributing conversions* and *online conversion*.

The resulting initial data set has been refined by only considering publications listed in the Business or Management category. The first outcome contains 123 articles. These were evaluated concerning the two pre-defined concepts (1) and (2). For the evaluation, the following order of information was consulted. First, the year, the title of the publication, the abstract and the number of citations were considered. If these four attributes were placed within the boundaries and the abstract indicates at least one of the pre-defined concepts to be fulfilled, the paper is identified for further proof (see Table 10). Within each iteration all articles identified for further proof were analysed closely. The article was kept if the content still meets at least one of the two concepts. The snowballing approach was applied for retained publications (Webster and Watson 2002). All used references of a kept publication and all publications which in turn use a publication as a source are evaluated in the next iteration. The Web of Science provided by Thomson Reuters was used to identify those forward-citations.

During the initial iteration, six articles were identified for further proof. These six articles were analysed to see whether they fulfil at least one of the pre-defined concepts. If they do the publication is kept. This research design is concept-centric. Table 10 shows the number of publications identified for further proof during each iteration. The column *Publications kept* represents the number of papers in each iteration, which still meet the concept after a detailed analysis of the publication.

Table 10: Structured literature process: identified publications based on concept 1 and concept 2

	Concept 1			**Concept 2**		
Iteration	Publications identified for further proof	Publications kept	No. of publication (see Table 13)	Publications identified for further proof	Publications kept	No. of publication (see Table 13)
0 (initial)	2	2	[5,8]	4	0	-
1	3	1	[1]	7	0	-
2	8	6	[2,3,4,6,7,9]	5	0	-
3	0	0	-	2	0	-
SUM		**9**			**0**	

During the process, the second concept turned out not to support the main objective because publications identified for further proof focus mainly on analysing cross-channel influences and not on attribution. Therefore, no paper was kept because of concept 2. After a total of four iterations (0 to 3), the search process was finalized. During the last iteration, no new publication was either identified for further proof or kept. This approach ensures that all relevant publications are already examined (Salipante et al. 1982). The results – we identified nine articles which are placed within the pre-defined boundaries – of the structured literature research are presented in the chapter *results*.

Hypotheses

The current research is led by the hypotheses listed in Table 11. At this point, to best of the authors' knowledge, the applicability from a practical perspective of existing attribution approaches in an omni-channel environment has not been analysed yet.

Table 11: Formulated hypotheses for the current research

Hypotheses	References
H1: In an omni-channel environment, new requirements are requested for attribution modelling from a practical point of view.	▪ Some approaches are limited to a certain amount of marketing channels. Furthermore, the authors claim that richer data on user level is needed (Abhishek et al. 2012; Nottorf 2014; Zhang et al. 2014). ▪ None of the identified attribution models/ approaches is described to be applicable in an omni-channel environment (Anderl et al. 2016a; Abhishek et al. 2012; Dalessandro et al. 2012; Geyik et al. 2014; Li and Kannan 2014; Nottorf 2014; Shao and Li 2011; Xu et al. 2014; Zhang et al. 2014). ▪ Verhoef et al. (2015) explain the shift from multi-channel towards omni-channel in the retail context. The authors estimate a comparable change in the context of attribution.
H2: Existing attribution models are not effectively applicable in an omni-channel environment from a practical perspective.	

Research Methodology

Empirical research is defined as a systematic, intersubjective verifiable collection control and criticism of experiences by Früh (2015). According to this author, an idea or a research question needs to be formulated at the beginning of the research. Based on the following research question the investigation is implemented.

"From an omni-channel perspective: What attributes and abilities do future proofed dynamic attribution models need to fulfil?"

In social sciences, a distinction is made between three various methodologies for empirical research: the quantitative, the qualitative and the mixed-method approach (Gläser and Laudel 2010; Creswell 2014) whereby the mixed-method approach consists of a combination of the first two approaches. A quantitative method verifies existing theories or assumptions in contrast to a qualitative approach which is theory-generating and also termed as mechanism-orientated. This latter explanation strategy offers a direct access to the mechanism (the theory) and allows the use of expert interviews as a survey methodology (Gläser and Laudel 2010) which was applied for identifying evaluation criteria and requirements from a practical point of view.

Requirements and specification criteria for dynamic attribution approaches in an omni-channel environment

Criteria for attribution in an omni-channel environment were needed. Hopf, Schmidt (1993) differentiates two intents for interviewing experts in a qualitative research. First, the experts are interviewed because they are experts for a special configuration, or second the interviewees are asked to gather interpretations, views and attitudes of the interviewees. For this study, the expert interviews were conducted with the first intend. All experts were interviewed because of their special knowledge about attribution due to their professional position (Bogner 2005). As a classification of expert interviews, a semi-structured expert interview was applied.

The data collection, as well as the evaluation, were carried out based on the qualitative content analysis by Gläser and Laudel (2010). Their modified analysis approach based on Mayring (2010) is more flexible and allows predefined categories to be adjusted. Gläser and Laudel (2010) underline that there is no holistic representation of the procedure model in social research and condemn that existing literature focus primarily on the fundamental principles of qualitative research, which leads to a more intuitive rather than systematic research. Bogner et al. (2014) describe that there is not *the one* method for evaluating expert interviews, which allows a conjunction of different approaches.

This exploratory methodological setup (Creswell 2014; Kuckartz 2014) was chosen because the number of available experts in this field was limited. Furthermore, the willingness of sharing insights was low because the ability of efficient attribution is a competitive advantage in the market. Additionally, performing in an omni-channel environment is relatively new (Verhoef et al. 2015) and finding real experts was a challenging task.

Expert Sampling

Expert interviews are a proven means to collect data if the number of available experts is limited (Gläser and Laudel 2010). These interviews were conducted to identify specifications, features or *nice to have* attributes of future proofed attribution models from a practical perspective.

All expert interviews were conducted during October 2016 and February 2017. The group of experts was divided up into the two units, to obtain sophisticated and reliable results. A purposive sampling (Flick 2007; Diekmann 2007) was chosen and applied to select experts. Summarizing, experts were classified into one of the following two groups:

1. practitioners (user or operators), and
2. publisher

This study focused on the practical perspective. To each interviewee, an id was assigned. Table 12 displays an overview of the conducted interviews, the expert group and the field of work the interviewee belongs to (n=9).

Table 12: List of interviews consisting of an Identifier, the assigned expert group and the field of work of the interviewee.

ID	Expert Group	Field of Work
[01]	Practitioner	Head of Digital Marketing & Analytics
[02]	Practitioner	Head of M&A
[03]	Practitioner	Head of SEO
[04]	Practitioner	Digital Data Analyst Expert
[05]	Practitioner	Digital Marketing Expert
[06]	Practitioner	Data Mining Expert
[07]	Practitioner	Data Innovation Expert
[08]	Publisher	Data Performance Expert
[09]	Publisher	Head of Data Management

To be acknowledged as an expert in this analysis all interviewees need to match the following criteria:

1. Solid working experience in their field for at least four years
2. Integrated technical and strategic knowledge in marketing
3. Specialist in the field of attribution or a related area

The first two criteria ensure that technical and strategic expertise in marketing is available grounded on a substantial working experience. In combination with being a specialist in the field of marketing or a related area people meeting all three criteria were considered to be an expert in this investigation. Before each interview, the interviewee is informed about ethical principles and the goal of the research.

The first interviews were conducted with experts belonging to the group of practitioners. Nine experts were interviewed in total. A saturation within each group was reached when there was no new requirement and specification criteria identified. A saturation for the group of publishers was achieved when there was no further assessment criteria defined – this is based on the results from the prior interviewed group of practitioners and the identified requirement and specification in the current group of publishers.

Interview Guideline

This current qualitative research process is inspired by the Qualitative Content Analysis by Schreier (2012) and by (Gläser and Laudel 2010). The coding frame consists of the following main research question for the interviews:

“From an omni-channel perspective: What attributes and abilities do future proofed dynamic attribution models need to fulfil?"

As an initial setup, the main categories (1) *data flow*, (2) *personalization* and (3) *integration in a productive environment* were predefined and pretest-verified. The category (1) *data flow* consists of the following sub-categories

1. input data
2. data quality,
3. the mathematical/statistical approach,
4. calculation, and
5. output.

This main category was inspired by the IPO-model (input-process-output) developed by Grady (1995). The *input* consists of the *input data* itself and the *data quality*. The *process* section was separated into *mathematical/statistical approach* and *calculation*.

The main categories (2) *personalization* and (3) *integration in a productive environment* do not contain any sub-categories. All experts were interviewed using the guideline consisting of the main- and sub-categories summarized in Figure 11.

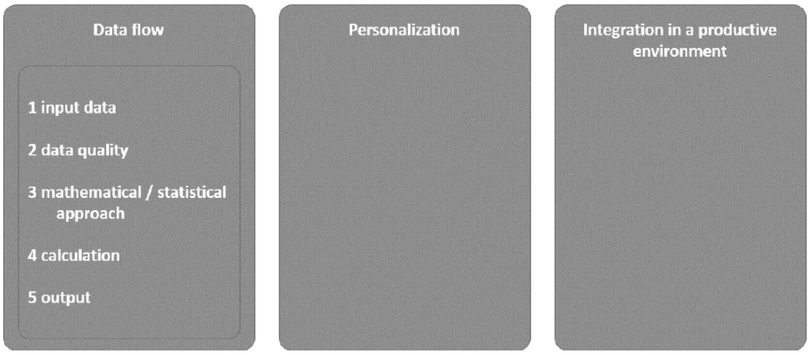

Data flow
1 input data
2 data quality
3 mathematical / statistical approach
4 calculation
5 output
Personalization
Integration in a productive environment

Figure 11: Guideline for the expert interviews: Main- and sub-categories

Conducting the Semi-Structured Expert Interviews

All nine interviews were conducted in German either on-site or by telephone. Three interviewees did not allow a voice recording. According to Gläser and Laudel (2010), an interview report including a memory record was constructed for those three interviews. Since only abilities and features of a future-proofed attribution model are relevant for the objective, these reports a limited to such information. Those three interviewees reviewed their summarized input by email afterwards and corrected statements if necessary. This process ensures correctness of the answers and that their meaning is accurate. The length of the other seven interviews varied from 11 to 21 minutes.

Before each interview, the two terms *dynamic attribution* and *omni-channel* (marketing) were explained by the interviewer to ensure the same understanding of those terms. All categories of the coding frame were shortly introduced during the interview, and possible questions of comprehension were answered. Once all interviews were conducted their audio records were literally transcribed (Mayring 2016) with corresponding timestamps to know who said what and when. These "literal transcription with literary script" present the foundation for the following evaluation.

Evaluation Methodology

For the evaluation, a content-structured content analysis was applied. As the primary guideline, the eight steps defined by Schreier (2012) were followed except for the trail coding. Gläser and Laudel (2010) extend the approach of Mayring (2010) by allowing categories to be modified and adjusted during coding. This method was applied to the procedure of Schreier (2012). The predefined categories turned out to be a well-chosen initial setup for categorization because only minor changes to the structure were necessary during analysis.

Identified requirements within a category were directly coded using the in-vivo coding methodology. As described by Saldaña (2009) this search for pattern enables and supports the analysis. To answer the research question only requirements and specifications were essential for this analysis. Statements by the interviewee that support a certain requirement were copied into the corresponding category for further investigation.

To ensure an appropriate degree of coding quality, a third independent person re-assigned extracts from the interviews to the existing categories. Based on Guest et al. (2012) an intercoder agreement helps to increase objectivity, reliability and validity.

During the last step of the evaluation, the requirements regarding sub-categories within the pre-defined categories were analysed and named. Often the names for the category were mentioned directly during the interview or were closely described.

Results

This chapter is separated into two parts. The first one consists of the results from the structured literature review (identification of existing dynamic attribution models) described in the introduction. In the second, the significant part the results from the expert interviews, the specifications and requirements towards attribution modelling in an omni-channel environment, are presented. This chapter closes by bringing these two result sets together regarding analysing the applicability of existing attribution models in an omni-channel environment.

Identify Dynamic Attribution Approaches

The outcome of the formal research process is presented in Table 13. In total, nine articles containing a dynamic attribution approach are kept in consideration of the literature search assumptions.

Table 13: Dynamic Attribution Approaches in Science: Results of the structured literature research process

No.	Author / Year of publication	Title
[1]	Abhishek et al. 2012	Media Exposure Through the Funnel: A Model of Multi-Stage Attribution. A Model of Multi-Channel Attribution
[2]	Anderl et al. 2016a	Mapping the customer journey. Lessons learned from graph-based online attribution modelling
[3]	Dalessandro et al. 2012	Causally motivated attribution for online advertising
[4]	Geyik et al. 2014	Multi-Touch Attribution Based Budget Allocation in Online Advertising
[5]	Li and Kannan 2014	Attributing Conversions in a Multichannel Online Marketing Environment
[6]	Nottorf 2014	Multi-channel Attribution Modeling on User Journeys
[7]	Shao and Li 2011	Data-driven multi-touch attribution models
[8]	Xu et al. 2014	The path to Purchase. A Mutually Exciting Point Process Model for Online Advertising and Conversions
[9]	Zhang et al. 2014	Multi-touch Attribution in Online Advertising with Survival Theory

In the following paragraphs, the identified dynamic attribution approaches are briefly described in an alphabetical order to understand the author's approach.

[1] Abhishek et al. (2012): The authors explain that at any given time, a customer's state within the conversion funnel can only be inferred through trackable actions such as clicks on an advertisement, a conversion or page views. Abhishek et al. therefore model user behaviour as a Hidden Markov Model (HMM). To perform attribution, they attribute actions to advertisemts which cause the user to change its latent state.

[2] Anderl et al. (2016a) model individual-level multichannel customer journeys as first- and higher-order Markov graphs. They use a property *removal effect* to determine the contribution of online channels and channel sequences. Their model outcome includes the conversion probability of a customer which can be used for third-party vendors such as real-time bidding. Anderl et al. apply their model to four data sets from three different industries.

[3] Dalessandro et al. (2012): The authors' approach is motivated by the need to standardize the data-driven multi-touch attribution field. They formulate multi-touch attribution as a causal estimation problem. To fit attribution into a game-theoretic framework, they make simplifying assumptions about the data. Their approach uses the concept of Shapley value (Shapley 1953). The authors claim their model to be more suitable from a practical perspective. However, the actual benefits come at the cost of accuracy.

[4] Geyik et al. (2014) focus on efficient advertisement attribute auctions in a campaign hierarchy. Their approach includes a MapReduce algorithm on Hadoop which makes it easy to parallelize the calculation. Apache Hadoop is an open source project. The Hadoop framework is used for distributed storage and bigdata processing (for MapReduce see White (2012)). This method is necessary because they are using "tens of terabytes of user profile data." In their opinion, the data foundation “represents perfectly the nature of real-world online advertising systems.”

[5] Li and Kannan (2014) are using a purchase decision hierarchy. They developed a conceptual framework to analyse the nature of carryover- and spillover-effects across online marketing channels through which customers visit a firm's website. Li and Kannan distinguish between customer-initiated and firm-initiated channels. The presented framework provides the basis for their three-level measurement model. The conversion decision of a customer at an online site differentiates between consideration, visit decision, and the purchase decision.

[6] Nottorf (2014): With the proposed model the author analyses the effect of advertising on the individual behaviour of consumers. The model is build up on a binary logit model with a Bayesian mixture approach to model consumer clickstreams across multiple types of online advertising. Using anonymized user-level data, this model helps to understand the effects of specific advertising channels on individual consumer behaviour and online purchasing processes.

[7] Shao and Li (2011) develop a bagged logistic regression model. The classification accuracy is comparable to logistic regression, but the estimation of individual advertising channel contributions is much more stable. The authors point out that their model has a reproducible result and claim their model to be "easy to interpret." Furthermore, according to the authors their model "is the industry's first data-driven multi-touch attribution model commercially available."

[8] Xu et al. (2014): "[…] develop a stochastic model for online purchasing and advertisement clicking that incorporates mutually exciting point processes with individual heterogeneity in a Bayesian hierarchical modelling framework. The mutually interesting point process is a multivariate stochastic process in which different types of advertisement clicks and purchases are modelled as various types of random points in continuous time." "[Xu et al.] develop a […] modelling approach that captures the exciting effects among advertisement clicks to contribute to the attribution models for properly evaluating the effectiveness of online ads using individual-level online clickstream data."

[9] Zhang et al. (2014) evolve an entirely data-driven model for the multi-channel attribution problem in online advertising. They use an additive hazard model based on survival theory. Next, to the time-decaying effect, the model considers the different levels of impact of various advertising channels.

So far, all consistent dynamic attribution approaches identified through the structured literature review process are presented and briefly described. All dynamic attribution approaches need to be evaluated concerning their applicability in an omni-channel environment to meet the objective. In the following evaluation criteria is defined and applied to the approaches identified in this chapter.

Evaluation Criteria for Attribution Approaches in an Omni-Channel Environment

During each interview, all main- and sub-categories were shortly introduced to the experts by the interviewer to outline the scope of each section. The interviewee is asked to give their opinion and appraisal based on their expert experience and expertise regarding requirements and specifications for future attribution modelling in each category. Table 14 contains the results of the interviews. Each identified evaluation criteria is briefly described to understand the interviewees' state. All criteria within a category are prioritized from very important to less critical, based on how often a criterion is mentioned and how crucial it is for the experts.

The main-category *personalization* turned out not to be supportive and is removed due to the lack of evaluation criteria and redundancies. Within the main category *data flow*, the initial sub-categories *mathematical/statistical approach* and *calculation* are combined and labelled as *calculation*, because the identified evaluation criteria turned out to be very similar. During the analysis, it becomes evident that some requirements or specifications mentioned by the interviewees are not a request which can be realized within an advanced attribution model. Some requests require special skills for marketing experts and others call for specialized input data. Therefore, after all definition of the requirements were completed, three classes were derived to distinguish between *model feature (requirement) / specification* (MF/S), *data requirement* (DR) and *other requirement* (OR). This assignment depends on whether the criterion is a requirement which needs to be handled within a model or the criterion requires special raw (data/) information.

A change of demanded skills and new requirements of what marketing experts need to know was identified during the analysis. Basic skills from the Business Intelligence (BI) [1, 3, 4, 8, 9] sector and a basic understanding of technical aspects [1, 4, 8, 9] were claimed. The primary focus is not on being able to develop sophisticated statistical models, but on understanding the data and the data sources [1, 3, 5, 7, 9]. The ability to identify promising measures and strategies will become a standard [1, 3, 6, 8, 9]. These requirements are assigned to the class *other requirement* (OR), are not part of the presented results in Table 14. In the following analysis all requirements within the class *other requirement* are not considered, because these findings are not supportive for answering the research question.

Table 14: Results: Evaluation criteria for attribution models in an omni-channel environment from a practical perspective. Each category is sorted in descending order of importance.

	Category	**Criteria**	**Category**		**Description**
			MF/S	**DR**	
DATA FLOW	**Input data**	Ability to handle input sources[1] containing hard facts [3, 4, 5, 7, 8, 9]	X	X	During the interviews, a differentiation between hard facts and soft facts evolved. Experts determine hard facts to be "real facts" such as • tracking data (user behaviour on a website or in an app), • company internal data sources (product data, CRM or DWH data), • external data sources (weather data, information about the user geo-location or other statistical data) • channel data (user behaviour data as far it is available for analysis) • offline data (such as purchases in a local store) etc. Soft facts, on the other hand, represent a meta-level of information which is derived from a users' behaviour or situation. The assumption of a users' feelings, attitude or position is determined to be a soft fact.
		Ability to handle input sources containing soft facts [3, 4, 5, 7, 8, 9]	X	X	
		Ability to add/remove data sources [2, 3, 4, 7, 8, 9]	X		Based on the expert practitioners' experience nearly all of them were confronted with the challenge to add or disregard a technology and a vendor within the last year. Future proofed attribution system needs to be flexible regarding easily adding and removing data sources.
	Data quality	Highest possible data granularity of input sources [1, 4, 6, 7, 8, 9]		X	For a useful calculation, all experts agree on the need to be able to access the raw information and not aggregated data.
		Stitch ability of a single user cross-devices [2, 4, 6, 7, 8, 9]	X	X	All experts mentioned unanimously that different data sources are a required foundation for calculations for a future-insured attribution model. As a fundamental requirement, a linked ability from data sources is indispensable. Furthermore, users using different devices

[1] Input sources are meant to be company internal generated or third-party data sources such as tracking or channel data or weather information.

		Linkable data sources [2, 4, 6, 7, 8, 9]		X	(personal computer, tablet, smartphone) need to be stitched towards one user profile.
	Calculation Combination of the two sub-categories **mathematical/ statistical approach** and **calculation**	Ability to calculate in real-time [2, 3, 4, 5, 6, 8, 9]	X	X	Adding a customer to a display advertisement campaign or segment in real-time becomes a major action in the online marketing context. Therefore, real-time calculation becomes a basic functionality. While a user is performing actions within the company offered channels (website, app, emails, etc.), it is necessary to be able to calculate the value of the user in real-time to perform user value appropriate actions.
		Incremental learning process [2, 3, 5, 6, 8, 9]	X		Learning assessed by the attribution model. The accuracy of the model should get better over time by taking newly-learned experiences into consideration.
		Ability to predict future actions [2, 3, 5, 6, 8, 9]	X		Budget allocation (attribution on a channel, audience or user basis) is the principal objective of an attribution model. Static approaches do neglect most of the user interaction data. A future-proof dynamic approach needs to be able to predict whether a user is going to perform a certain activity (e.g. purchase, sign up for a service) or not. A predictive functionality should, therefore, be a basic functionality.
		Value calculation on user level [1, 2, 3, 4, 5, 8, 9]	X	X	In practice, it is still prevalent to calculate the budget spending on a channel basis. By doing so, user information such as their intent and attitude are neglected and aggregated within a channel value. Actions can't be done on a user or audience basis if data is aggregated to a channel level. To be able to be as close to a user as possible a value calculation on a user level is indispensable. Knowing that a value estimation on a user level is very complex and depends on much data and good data quality, most of the practitioners pointed to a calculation on an audience basis as an intermediate step. If the necessary information is available, the calculation on a user basis is preferred.
		Value calculation on audience basis [1, 2, 4, 5, 8, 9]	X	X	
		Machine learning /	X		To interact with every single user in a productive way, at best the marketing team has information

		Artificial Intelligence approach [1, 2, 3, 6, 8, 9]			available implying knowledge about the mood, situation and other attributes of every single user. Because of the fact, that user behaviour is dynamic the model needs to be able to take this into consideration.
		Data-driven calculation – not rule-based [8, 9]	X		The model should be able to configure itself dynamically based on the data from the input sources. Static manual rules always lead to inaccuracy because they are not able to handle dynamic changes.
	Output	High-quality output [1, 2, 4, 6, 8, 9]	X	X	This request can be split up into the two following characteristics: A self-evident request is the correctness of the output. Furthermore, this requirement contains the ability for further process-ability in terms of interpretability. Due to this aspect, one of the largest search engines still uses variants of the Shapley value (Shapley 1953) for attribution.
		Ability to connect (third party) vendors directly (automated connection) [2, 4, 5, 7, 8, 9]	X	X	Directly connect third-party vendors. For example, adding/removing a user to a display advertisement campaign. These actions need to be automated to act in real-time without manual delays.
		Performance test of the model outcome/data validation [2, 3, 5, 7, 8, 9]	X		To ensure the estimations of the model, a validation process is requested.
		Intuitive interface [2, 5, 7]	X		Practitioners require an intuitive interface to be able to control the model, see the performance and manage actions, such as communication with third-party vendors.

INTEGRATION IN A PRODUCTIVE ENVIRONMENT	**Integration (technical acceptance)**	Interface driven design [2, 4, 6, 8, 9]		X	A future-insured model needs to be able to hook up to existing input sources. Criteria for exclusion – as well as result quality - is the ability to integrate already existing data source within a present system environment.
		Interface definition / standards [2, 4, 6, 8, 9]		X	There is the need to define standards and interfaces. Already in 2012 Dalessandro et al. (2012) tried to bring more standardization in the field of dynamic attribution. A big challenge for marketing experts is to try to get different products from various vendors to work together from a data perspective. The expectations about quickly adding and removing (third party) data sources and applications from third-party vendors are almost unanimous. The tag management approach (Tealium iQ (www.tealium.com), Google TagManager (www.google.com/tagmanager) or TagCommander (www.commandersact.com)) is a good base, but the vendors themselves have different interfaces which are not standardized.
		Plug and play [2, 4, 9]	X	X	

During the analysis of the interviews, it turned out that some requirements such as the ability to stitch a customers' journey across different devices are not only a requirement for the model. The input data needs to include this piece of information as well. Aspects such as using only data with the highest granularity is also a data requirement. Being able to calculate in real-time is both, a model requirement and a data requirement. One the one hand the model performance needs to be so efficient to do the calculation in real-time, on the other hand providing the data can be a limitation as well.

Evaluation of Existing Dynamic Attribution Models Towards their Applicability in an Omni-Channel Environment

The required features and specifications identified by experts can be understood as evaluation criteria for existing attribution models. Next, these evaluation criteria, are applied to evaluate if the existing dynamic attribution approaches (still) meet the experts' requirements and specifications.

Table 15 contains the evaluation of the identified dynamic attribution models towards the applicability in an omni-channel environment by applying the requirements and specifications formulated by experts. Requirements and specifications only included in the class *data requirement* are greyed out because such elements are not part of the attribution model. The following evaluations for those requirements are made based on the described data foundation used for the constructed attribution model.

Table 15: Evaluation of the identified attribution models towards identified requirements and specifications from a practical perspective

Cat.	Criteria	1	2	3	4	5	6	7	8	9
		Abhishek et al. (2012)	Anderl et al. (2016a)	Dalessandro et al. (2012)	Geyik et al. (2014)	Li and Kannan (2014)	Nottorf (2014)	Shao and Li (2011)	Xu et al. (2014)	Zhang et al. (2014)
Input	Using hard facts for calculation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Using soft facts for calculation	Yes							Yes	
	Ability to add/remove data sources	Yes	Yes			Yes		Yes	Yes	
Data quality	Highest possible granularity of data sources used	Yes	Yes		Yes	Yes	Yes	Yes		
	Stitch ability of a single user cross-devices									
	Linkable data sources	Yes				Yes	Yes	Yes	Yes	Yes
Calculation	Calculation in real-time									
	Incremental learning process	Yes								
	Predictive approach	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes
	Value calculation on user or audience basis	Impact of every ad impression at an individual level			Calculation on ad basis					
	Machine learning / Artificial Intelligence approach									
	Data-driven calculation – not rule-based									Yes

Output	High-quality output									
	Ability to connect (third party) vendors directly (automated connection)		Yes				Yes			
	Performance test of the model outcome/data validation	Yes		Yes	Yes	Yes		Yes	Yes	Yes
	Intuitive interface									
Integration	Interface driven design									
	Interface definition / standards									
	Plug and play									

With existing attribution models identified and described, the primary objective of the current research can be studied and the both hypotheses can be verified.

H1: *In an omni-channel environment new requirements are requested for attribution modelling from a practical point of view*. The hypothesis H1 has been verified. The authors identified several new requirements and specification (see Table 14) which were not realisable in a multi-channel or cross-channel environment.

H2: *Existing attribution models are not effectively applicable in an omni-channel environment from a practical perspective.* The hypothesis H2 has been verified as well. Table 15 provides the results of the analysis of the applicability of current attribution models in an omni-channel environment. Abhishek et al. (2012) provide the best fitting model in an omni-channel environment.

Discussion

We discovered a major change regarding requirements towards attribution modelling in an omni-channel environment. Based on the identified requirements from a practical point of view the most critical difference consists of how attribution is performed. This change is not meant regarding calculation or performance; it is rather the granularity of the attribution model output. An attribution per channel is not requested primarily anymore. The fact that the majority of interviewed experts state that future attribution needs to be calculated on an audience or a user level implies a new way of attributing in an omni-channel environment. Attribution is no longer only a supportive action for marketing experts how to split their marketing budget. The previous focus on the channel is moved towards the user, who takes centre stage and becomes the primary target. This opens up a variety of new research fields as described in the next paragraphs.

Data Granularity

Next, to hard facts such as real usage data, there is a request to take soft facts into consideration. Hard facts are the basis for all identified existing attribution models. From a practical point of view, an attribution based only on behavioural data is not sufficient anymore. This focus is because essential insights e.g. the users current feeling, attitude, intent and situation are characterized to be important as well. Some authors mentioned this also as a limitation for their attribution approach (Nottorf 2014). Taking such meta-information about the user into consideration opens up a new area for further research. What kind of meta-attributes do exist and what impact does each meta-information have? The need for more information about the user is also supported by the third request in the category of *input data* – the ability to add and remove different data sources. Next, to flexibility, this request implies a strived exploratory discovery to identify new valuable insight of a customer which have a significant impact on attribution.

All identified requirements in the category of *calculation* imply that the research area *attribution* in the field of marketing and business intelligence have to grow further together. This circumstance is also mentioned directly by the majority of experts and requires new skills for marketing practitioners.

A *stitch ability* of users across different devices underlines the need for more research in the fields of user detection. There are two different approaches on how stitching can be realized: A deterministic approach – the user has to enter login credentials – or a probabilistic approach where a cross-device tracking is implemented based on non-personal information such as IP-address, location data, etc. (Whitener 2015). Diaz-Morales (2015) try to detect a user on different devices by using semi-supervised machine learning methods. Such probabilistic approaches are never 100% precise, and their applicability has not been analysed in a marketing attribution context.

Linkable data sources require a setup which ensures and enables linking information in all data sources. This requirement should be implemented in a way while the data is being collected. As a rather technical issue than a strategic one, this requires knowledge from data science experts. Current discussions about data privacy will have an impact on what can be achieved in this field.

Real-Time Prediction

As already mentioned in the introduction of the discussion, attribution is no longer a tool only for marketing experts to split up their marketing budget in the most efficient way. Attribution experts require attribution data in real-time to enable purposeful actions when the user interacts with the company (e.g., through an app, the website or any other offered interface). This need requires the attribution model to be part of the whole environment a company provides to interact with one's customers. How such integration should be realized becomes another area of research in the IT/ e-commerce field.

A prediction in real-time necessitates a quick responding model and a predefined interface of what a request towards the model consists of. Furthermore, the output of such an attribution model needs to be defined. Useful information could be the current value of the user, the probability of a conversion or the preferred marketing channel(s). From a statistical point of view, a reasonable amount which could/should be spent on marketing activities for a single user could also be such an output. This output information can be used to for example add/ or remove the user automatically to an advertisement campaign. Defining such output information by investigating the impact of each value opens up different knowledge gaps for further research.

Prediction in real-time combined with the request of an approach consisting of machine learning / artificial intelligence shift the attribution problem from the marketing field towards the field of research in machine learning. As it was realized in a cross-/multi-channel environment - different statistical approaches were applied to the question of attribution - the authors expect the development of new models with a machine learning approach in the future. The provided model from Abhishek et al. (2012) also does not meet all requirements and is therefore only suitable for the omni-channel environment to a limited extend.

The primary goal for an advertiser and website provider is to make the highest possible amount of money (through e.g. conversion, purchase, or sign up for a newsletter) out of the performed action (Geyik et al. 2014). This will continue to be the primary goal. The authors estimate a significant shift from a widely spread current perspective which focuses on revenue, not having the user in focus (e.g. general newsletter mailings to all know email addresses, not considering whether a user has an interest in the content or not) towards a user-centred perspective. “All that customers are concerned about is finding an answer to their current needs or desires in a way that is convenient, enjoyable and offers them real value, both regarding money and use of their time.” (Cook 2014) There are significant players such as Amazon, eBay, Google or Facebook which develop effectively in their way of interacting with their customers by focusing on the customer’s intent. Companies which are unable to correspond with their customers in an individualized accepted and supportive way or companies staying with the revenue-focused strategy probably stay satisfied with their marketing actions initially, but not in the long term. Spoiled internet users, used to appropriate addressed touch points with a company, will get a bad impression of enterprises offering inappropriate touchpoints. This will produce a negative attitude towards companies behaving money-centralized. The need to be able to predict the customer's necessities, taking into account the budget, will determine whether a marketing strategy is successful or not. A precise attribution of a budget per customer or audience is indispensable. Only Abhishek et al. (2012) and Geyik et al. (2014) perform a calculation based on the impact of an advertisement impression or the impression itself. This part of budget allocation needs to be further developed in future omni-channel attribution approaches.

Critical Examination of the Results

Most identified requirements presuppose a reliable, correct and holistic data foundation. From a practical perspective, this is still a challenge and a critical area for companies and brands for further development. This challenge probably will be the most challenging task to accomplish because such a data foundation can’t be achieved without management guidelines. A marketing department itself is not able to realize it independently without the IT and data science experts.

Before an attribution model can be applied to different data sources, those sources need to be pre-processed in some ETL (extract, transform, load) process. A required plug and play environment using standardized interfaces to connect different data sources is not given. The *plug and play* requirement is hard to realize in practice because this presupposes a standardized interface. Data source providers, such as Google, Facebook, advertisement vendors, etc., do not necessarily have the interest to have their data used externally. Already in 2012 Dalessandro et al. were motivated to present a dynamic attribution model “by a need to bring […] more standardization and data-driven intelligence [in this field]”. This motivation might be reconsidered by scientists in the future. Useful attribution is a competitive advantage in the market. Due to this circumstance developed approaches in businesses are not published to keep the advantage internally. In the author's opinion, it is difficult for the science community to be ahead of the real market. By providing structured approaches and evaluating influencing attributes, the science community should lead the practice to a more formal environment.

Future Research Fields and Research Questions from a Practical Perspective

Based on the results of the expert interviews and the applicability of existing attribution models in an omni-channel environment, research areas are identified, and research issues and research topics are formulated. These issues and topics can be used as an input for future research which also meets practitioners' demands. Results, presented in Table 16, are separated in the following research areas *input data*, *data quality*, *mathematical/statistical approach and calculation*, *output*, and *implementation and management perspective*. All questions and topics are placed in the field of dynamic attribution and the corresponding environment. Research questions and research topics are prioritized from important to less critical. Within a research area, research issues and research topics are prioritized based on how relevant they are to the interviewee from their practical point of view and how often they were mentioned during the interview.

Table 16: Proposed research agenda for further research in the field of dynamic attribution modelling from a practical perspective

Research areas	**Research question / research topics**
Input data	→ Which data (granular information) has what degree of impact on the quality of the attribution. Is *as much data as possible* a reasonable approach? → Overview of reasonably available data sources. → Identify necessary steps within an organization to have a solid data foundation for attribution in an omni-channel environment. → Analyse the optimal/minimum amount of data needed for successful attribution modelling. Analyse derived features. → In terms of the quality of the outcome analyse the impact of data sources such as Facebook data, Google AdWords data, etc. → User tracking in a cross-device environment. - Evolve alternatives to cookie based-, statistical- and logged in based approaches. - How to handle users which are not logged in. - From a technical point of view: What information needs to be presented from an operating system such as Windows, Mac OS, Android, etc. to be able to identify a user and being able to stitch users across different devices. → Calculate soft facts. What are the most important soft attributes having what influence and how to calculate them → Define a standard for input sources or a standard for attribution models based on the given data to facilitate a plug and play functionality.

Data quality	→ Analyse different data sources and score them regarding quality and quality gain for an attribution model → Bring the two fields business intelligence (BI) and online marketing together and apply BI methods on marketing questions taking real-time calculation, value calculation on audience/user basis, or incremental learning process into account. → User profiling: Identify industry-specific user attributes and general user attributes and their value to the outcome of an attribution model.
Mathematical/statistical approach and calculation	→ Build an attribution model which implements an incremental learning mechanism. → Build an attribution model using an artificial intelligence (AI) and/or machine learning approach which considers the change in customer behaviour. → Geyik et al. (2014) present an approach using MapReduce on Hadoop. What attributes does a sophisticated server structure need to fulfil? → The real-time calculation of a decentralized data source environment. Limitations and challenges.
Output	→ Identify requirements due to new challenges in an omni-channel environment. → Identify mutual channel influences based on a valid cross-device data foundation in an omni-channel environment. → Estimate if culture-specific approaches are needed within one industry. (Vuylsteke et al. (2010) discovered Chinese to have a different online search process compared to Western Europeans). → Research on how much marketing budget is (still) wasted with omni-channel approaches. → Develop a dynamic attribution model which meets all the identified requirements of this current study. → Model a standardized output (A standardized output promote the attributes fairness and interpretably claimed by Dalessandro et al. (2012))
Implementation and management perspective	→ Data warehouse (DWH), Data management platforms (DMP) and the approach of a dynamic attribution model need to grow together. What are the barriers and people within a company which you need to be aware of? → Getting ready for omni-channel. What are pitfalls and what needs to be considered for a prosperous implementation (from a marketing and/or technical perspective)? → Change Management: How to communicate and implement a change process from cross-channel marketing (various data silos) towards an omni-channel approach (combine data silos) → Identify and formulate new required skills of practitioners in the online marketing business → Analyse the costs of a change process from cross-channel towards omni-channel marketing.

	→ Develop a structured incremental learning framework or model. Define a structure or framework which offers the ability to add, test and remove *knowledge bricks (*knowledge gained from specific research in the marketing performance field. i.e. papers which deal with channel performance or in between channels*)*. → Analyse savings potentials → Define a complexity per model (for practical application) a) Regarding implementation b) Regarding required marketing knowledge for a successful use

Conclusion

The primary objective of this paper consists of identifying requirements and specifications towards attribution modelling in an omni-channel environment from a practical perspective. Identifying existing attribution models and the abilities these current models already fulfil was also part of the objective. The answer to whether existing models are applicable in an efficient way in an omni-channel environment is: *no*. No identified model offers all or a majority amount of the requirements and specifications. For example, no model consists of the ability to perform the calculation in real-time, and only one model (Abhishek et al. 2012) consist of an *incremental learning process* which are rated to be the most essential characteristics in the category *calculation*.

The identified new requirements and specifications from a practical perspective can be understood as the basis for a call for new research in this area. Furthermore, those requirements underline a practical necessity. Both hypotheses (H1 and H2) were verified. From a practical point of view, there are new requirements for attribution modelling in an omni-channel environment. Furthermore, the second hypotheses H2 stating that existing attribution models are not effectively applicable in an omni-channel environment was verified as well.

The input from practitioners should ensure that scientists understand demands, challenges and problems from a practical perspective. According to Ulrich (Ulrich 1995), the applied research design begins with practical problems which are unresolved. These problems are analysed using available literature and theories. His design justifies scientific-based research based on the needs of practitioners (Ulrich et al. 1976). Because the Marketing Science Institute addresses *Attribution* to be the number one priority research area in the year 2016 to 2018 (MSI 2016), this is an important research area for both, practitioners and scientists. New requirements for a future-proofed attribution system with the ability to perform the calculation in real-time and calculate the value on an audience- or user-basis, as well as an incremental learning process, are only a subset of the new challenges for scientists.

Additionally, future studies in the context of attribution, need to be more structured to be able to implement an approach which is able to ensure that gained knowledge is included in an incremental learning process. There is not only the call for an incremental learning process within a model, but there is also a need for an incremental learning process in this field by defining standards and consequential comparability. One of the most significant challenges is to derive a structured approach - a framework or a model - which aggregates gained knowledge and offers a process which does not neglect marketing insights already gained.

References

Abhishek, Vibhanshu; Fader, Peter; Hosanagar, Kartik (2012): Media Exposure Through the Funnel. A Model of Multi-Channel Attribution. Available online at http://dx.doi.org/10.2139/ssrn.2158421.

Ackermann, Sebastian; Wangenheim, Florian von (2014): Behavioral Consequences of Customer-Initiated Channel Migration. In Journal of Service Research 17 (3), pp. 262–277. DOI: 10.1177/1094670513519862.

Alon, Noga; Gamzu, Iftah; Tennenholtz, Moshe (2012): Optimizing Budget Allocation Among Channels and Influencers. In Alain Mille, Fabien Gandon, Jacques Misselis, Michael Rabinovich, Steffen Staab (Eds.): Proceedings of the 21st international conference on World Wide Web. the 21st international conference. Lyon, France. 21st World Wide Web Conference 2012; ACM Special Interest Group on Hypertext, Hypermedia, and Web. New York, NY: ACM, p. 381.

Anderl, Eva; Becker, Ingo; Wangenheim, Florian von; Schumann, Jan Hendrik (2016a): Mapping the Customer Journey. Lessons Learned from Graph-Based Online Attribution Modeling. In International Journal of Research in Marketing 33 (3), pp. 457–474. DOI: 10.1016/j.ijresmar.2016.03.001.

Anderl, Eva; Schumann, Jan Hendrik; Kunz, Werner (2016b): Helping Firms Reduce Complexity in Multichannel Online Data. A New Taxonomy-Based Approach for Customer Journeys. In Journal of Retailing 92 (2), pp. 185–203. DOI: 10.1016/j.jretai.2015.10.001.

Archak, Nikolay; Mirrokni, Vahab S.; Muthukrishnan, Si. (2012): Budget Optimization for Online Campaigns with Positive Carryover Effects. In Paul W. Goldberg (Ed.): Internet and network economics. 8th international workshop, WINE 2012, Liverpool, UK, December 10 - 12, 2012 ; proceedings, vol. 7695. Berlin: Springer (Lecture Notes in Computer Science, 7695), pp. 86–99.

Bacharach, Samuel B. (1989): Organizational Theories. Some Criteria for Evaluation. In The Academy of Management Review 14 (4), p. 496. DOI: 10.2307/258555.

Berman, Ron (2015): Beyond the Last Touch. Attribution in Online Advertising. Available online at http://ron-berman.com/papers/attribution.pdf, checked on 9/28/2016.

Bogner, Alexander (Ed.) (2005): Das Experteninterview. Theorie, Methode, Anwendung. 2nd ed. Wiesbaden: VS Verl. für Sozialwiss. Available online at http://www.socialnet.de/rezensionen/isbn.php?isbn=978-3-531-14447-4.

Bogner, Alexander; Littig, Beate; Menz, Wolfgang (2014): Interviews mit Experten. Wiesbaden: Springer Fachmedien Wiesbaden.

Camiade, Benoït A. J.-M. (2013): Multi-Channel, Cross-Channel, Omni-Channel Retailing: Business in All Its Forms (1/2). ATInternet.com. Available online at https://blog.atinternet.com/en/series-multi-channel-cross-channel-omni-channel-retailing-business-forms-12/, checked on 12/12/2017.

Cook, Glenn (2014): Customer Experience in the Omni-Channel World and the Challenges and Opportunities This Presents. In Journal of Direct, Data and Digital Marketing Practice 15 (4), pp. 262–266. DOI: 10.1057/dddmp.2014.16.

Creswell, John W. (2014): Research Design. Qualitative, Quantitative, and Mixed Methods Approaches. 4. ed., internat. student ed. Los Angeles Calif. u.a.: Sage.

Cronin, Patricia; Ryan, Frances; Coughlan, Michael (2008): Undertaking a Literature Review. A Step-by-Step Approach. In British journal of nursing 17 (1), pp. 38–43.

Dalessandro, Brian; Perlich, Claudia; Stitelman, Ori; Provost, Foster (2012): Causally Motivated Attribution for Online Advertising. In ADKDD '12 Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy, pp. 1–9. DOI: 10.1145/2351356.2351363.

Diaz-Morales, Roberto (2015): Cross-Device Tracking. Matching Devices and Cookies. In Peng Cui (Ed.): 15th IEEE International Conference on Data Mining Workshop (ICDMW). Atlantic City, NJ, USA. Institute of Electrical and Electronics Engineers; IEEE. Piscataway, NJ: IEEE, pp. 1699–1704.

Diekmann, Andreas (2007): Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen. 17. Ed. Reinbek: Rororo Rowohlt-Taschenbuch-Verl. (Rowohlts Enzyklopädie, 55551).

Dinner, Isaac M.; van Heerde, Harald J.; Neslin, Scott (2011): Driving Online and Offline Sales. The Cross-Channel Effects of Digital Versus Traditional Advertising. In SSRN Journal. DOI: 10.2139/ssrn.1955653.

Econsultancy (2015): Quarterly Digital Intelligence Briefing. Digital Trends 2015. Available online at https://www.marketingsociety.com/sites/default/files/QDIB%20Adobe%20Digital%20Trends%20Report%202015_EMEA_0.pdf, checked on December 12th, 2017.

eMarketer (2016): Primary Attribution Model Used by Their Marketing Team to Measure Performance According to US B2B Marketers. Available online at emarketer.com, checked on January 20th, 2016.

Feit, Eleanor Mcdonnell; Wang, Pengyuan; Bradlow, Eric T.; Fader, Peter S. (2013): Fusing Aggregate and Disaggregate Data with an Application to Multiplatform Media Consumption. In Journal of Marketing Research 50 (3), pp. 348–364. DOI: 10.1509/jmr.11.0431.

Flick, Uwe (2007): Qualitative Sozialforschung. Eine Einführung. 1st ed., revised and extened. Reinbek bei Hamburg: Rowohlt Taschenbuch Verl. (Rororo, 55694).

Früh, Werner (2015): Inhaltsanalyse. Theorie und Praxis. 8., überarbeitete Auflage. Konstanz, München: UVK Verlagsgesellschaft mbH; UVK / Lucius (UTB, 2501).

Gallino, Santiago; Moreno, Antonio (2014): Integration of Online and Offline Channels in Retail. The Impact of Sharing Reliable Inventory Availability Information. In Management Science 60 (6), pp. 1434–1451. DOI: 10.1287/mnsc.2014.1951.

Geyik, Sahin Cem; Saxena, Abhishek; Dasdan, Ali (2014): Multi-Touch Attribution Based Budget Allocation in Online Advertising ADKDD’14, pp. 1–9. DOI: 10.1145/2648584.2648586.

Gläser, Jochen; Laudel, Grit (2010): Experteninterviews und qualitative Inhaltsanalyse. Als Instrumente rekonstruierender Untersuchungen. 4st ed. Wiesbaden: VS Verl. f. Sozialwiss (Lehrbuch).

Google Inc. (2017): Attribution Modeling Overview. Assign Credit for Sales and Conversions to Touchpoints in Conversion Paths. Available online at https://support.google.com/analytics/answer/1662518?hl=en, checked on January 21st, 2017.

Grady, Jeffrey O. (1995): System Engineering Planning and Enterprise Identity. 1st ed.: CRC Press.

Greenhalgh, Trisha; Peacock, Richard (2005): Effectiveness and Efficiency of Search Methods in Systematic Reviews of Complex Evidence: Audit of Primary Sources. In BMJ (Clinical research ed.) 331 (7524), pp. 1064–1065. DOI: 10.1136/bmj.38636.593461.68.

Grewal, Dhruv; Bart, Yakov; Spann, Martin; Zubcsek, Peter Pal (2016): Mobile Advertising. A Framework and Research Agenda. In Journal of Interactive Marketing 34, pp. 3–14. DOI: 10.1016/j.intmar.2016.03.003.

Guest, Greg; MacQueen, Kathleen M.; Namey, Emily E. (2012): Applied Thematic Analysis. Thousand Oaks CA u.a.: Sage.

Haan, Evert de; Wiesel, Thorsten; Pauwels, Koen (2016): The Effectiveness of Different Forms of Online Advertising for Purchase Conversion in a Multiple-Channel Attribution Framework. In International Journal of Research in Marketing 33, pp. 491–507.

Hammerschmidt, Maik; Falk, Tomas; Weijters, Bert (2015): Channels in the Mirror. An Alignable Model for Assessing Customer Satisfaction in Concurrent Channel Systems. In Journal of Service Research 19 (1), pp. 88–101. DOI: 10.1177/1094670515589084.

Hopf, Christel; Schmidt Christiane (1993): Zum Verhältnis von innerfamilialen sozialen Erfahrungen, Persönlichkeitsentwicklung und politischer Orientierungen. Dokumentation und Erörterung des methodischen Vorgehens in einer Studie zu diesem Thema. Unpublished Manuscipt.

IDC (2010): IDC Retail Insights. Multichennal Report 2010. Available online at http://info.hybris.com/rs/hybris/images/IDC-Multichannel-EN.pdf, checked on April 4th, 2015.

Jayawardane, Himani W.; Halgamuge, Sage. K.; Kayande, Uka. (2015): Attributing Conversion Credit in an Online Environment: An Analysis and Classification. In: 2015 3rd International Symposium on Computational and Business Intelligence (ISCBI). Bali, Indonesia, pp. 68–73.

Joo, Mingyu; Wilbur, Kenneth C.; Cowgill, Bo; Zhu, Yi (2014): Television Advertising and Online Search. In Management Science 60 (1), pp. 56–73. DOI: 10.1287/mnsc.2013.1741.

Kannan, P. K.; Reinartz, Werner; Verhoef, Peter C. (2016): The Path to Purchase and Attribution Modeling. Introduction to special section. In International Journal of Research in Marketing 33 (3), pp. 449–456. DOI: 10.1016/j.ijresmar.2016.07.001.

Kuckartz, Udo (2014): Mixed Methods. Methodologie, Forschungsdesigns und Analyseverfahren. Wiesbaden: Springer Fachmedien Wiesbaden. Available online at http://dx.doi.org/10.1007/978-3-531-93267-5.

Lee, Garry (2010): Death of ‘last click wins’. Media Attribution and the Expanding Use of Media Data. In J Direct Data Digit Mark Pract 12 (1), pp. 16–26. DOI: 10.1057/dddmp.2010.14.

Leone, Robert P. (1995): Generalizing What Is Known About Temporal Aggregation and Advertising Carryover. In Marketing Science 14 (3), 141-150. DOI: 10.1287/mksc.14.3.G141.

Lewis, Randall A.; Rao, Justin M. (2013): On the Near Impossibility of Measuring the Returns to Advertising. Available online at http://justinmrao.com/lewis_rao_nearimpossibility.pdf, checked on November 21st, 2015.

Li, Hongshuang; Kannan, P. K. (2014): Attributing Conversions in a Multichannel Online Marketing Environment. An Empirical Model and a Field Experiment. In Journal of Marketing Research 51 (1), pp. 40–56. DOI: 10.1509/jmr.13.0050.

Mayring, Philipp (2010): Qualitative Inhaltsanalyse. Grundlagen und Techniken. 11th ed. Weinheim u.a.: Beltz (Pädagogik).

Mayring, Philipp (2016): Einführung in die qualitative Sozialforschung. Eine Anleitung zu qualitativem Denken. 6th ed. (Pädagogik).

Moffett, Tina; Pilecki, Mary; McAdams, Rebecca (2014): The Forrester Wave: Cross-Channel Attribution Providers, Q4 2014. Available online at https://www.forrester.com/report/The+Forrester+Wave+CrossChannel+Attribution+Providers+Q4+2014/-/E-RES115221, checked on November 5th, 2017.

MSI, Marketing Science Institute (2016): Research Priorities 2O16–2O18. MSI Marketing Science institute. Available online at https://www.msi.org/uploads/articles/MSI_RP16-18.pdf, checked on November 3rd, 2017.

Naik, Prasad A.; Peters, Kay (2009): A Hierarchical Marketing Communications Model of Online and Offline Media Synergies. In Journal of Interactive Marketing 23 (4), pp. 288–299. DOI: 10.1016/j.intmar.2009.07.005.

Neslin, Scott A.; Grewal, Dhruv; Leghorn, Robert; Shankar, Venkatesh; Teerling, Marije L.; Thomas, Jacquelyn S.; Verhoef, Peter C. (2006): Challenges and Opportunities in Multichannel Customer Management. In Journal of Service Research 9 (2), pp. 95–112. DOI: 10.1177/1094670506293559.

Nottorf, Florian (2014): Multi-Channel Attribution Modeling on User Journeys. In: E-Business and Telecommunications, vol. 456, pp. 107–125.

Nottorf, Florian; Funk, Burkhardt (2013): A Cross-Industry Analysis of the Spillover Effect in Paid Search Advertising. In Electron Markets 23 (3), pp. 205–216. DOI: 10.1007/s12525-012-0117-z.

Olbrich, Rainer; Schultz, Carsten D. (2014): Multichannel Advertising. Does Print Advertising Affect Search Engine Advertising? In European Journal of Marketing 48 (9/10), pp. 1731–1756. DOI: 10.1108/EJM-10-2012-0569.

Paccard, Erwan (2017): Omnichannel vs Multichannel: Are they so different? Available online at http://multichannelmerchant.com/blog/omnichannel-vs-multichannel-different/, checked on June 11th, 2017.

Petersen, Andrew J.; McAlister, Leigh; Reibstein, David J.; Winer, Russell S.; Kumar, V.; Atkinson, Geoff (2009): Choosing the Right Metrics to Maximize Profitability and Shareholder Value. In Journal of Retailing 85 (1), pp. 95–111. DOI: 10.1016/j.jretai.2008.11.004.

Polo, Yolanda; Sese, Javier F. (2016): Does the Nature of the Interaction Matter? Understanding Customer Channel Choice for Purchases and Communications. In Journal of Service Research 19 (3), pp. 276–290. DOI: 10.1177/1094670516645189.

Rusko, Rauno (2015): Conflicts of Supply Chains in Multi-Channel Marketing. A Case From Northern Finland. In Technology Analysis & Strategic Management 28 (4), pp. 477–491. DOI: 10.1080/09537325.2015.1100294.

Rutz, Oliver J.; Bucklin, Randolph E. (2011): From Generic to Branded. A Model of Spillover in Paid Search Advertising. In Journal of Marketing Research 48 (1), pp. 87–102. DOI: 10.1509/jmkr.48.1.87.

Saldaña, Johnny (2009): The Coding Manual for Qualitative Researchers: Sage Pubn Inc.

Saldanha, Alexander; Berman, Ron; Vummarao Keshore (2013): Advertising Conversion Attribution. Applied for by Abakus, Inc., Emeryville, CA (US) on 3/14/2013. App. no. 13/830,618. Patent no. US 8,775,248 B1.

Salipante, Paul; Notz, William; Bigelow, John (1982): A Matrix Approach to Literature Reviews. In Research in Organizational Behavior. 4, pp. 321–348.

Schreier, Margrit (2012): Qualitative Content Analysis in Practice. 1st ed. London u.a.: Sage Publ.

Shao, Xuhui; Li, Lexin (2011): Data-Driven Multi-Touch Attribution Models. In Chid Apte, Joydeep Ghosh, Padhraic Smyth (Eds.): the 17th ACM SIGKDD international conference. San Diego, California, USA, pp. 258–264.

Shapley, Lloyd S. (1953): A Value for n-Person Games. Contributions to the Theory of Games, Volume II. H.W. Kuhn und A.W. Tucker: Contributions to the Theory of Games, volume II. (Princeton University Press).

Ulrich, Hans (1995): Von der Betriebswirtschaftslehre zur systemortientierten Managementlehre. Wunderer, R., Betriebswirtschaftslehre als Management-und Führungslehre. 3rd Edition. Stuttgart, Germany: Schäffer-Poeschel

Ulrich, Hans; Krieg, Walter; Mallik, Fredmund (1976): Zum Praxisbezug einer systemorientierten Betriebswirtschaftslehre. In: Zum Praxisbezug der Betriebswirtschaftslehre - in wissenschaftlicher Sicht. Bern / Stuttgart: Ulrich, Hans, pp. 135–151.

Uniquedigital (2012): Cross-Channel Management. Optimierung der Budgetallokation durch User-Journey Analyse und dynamisches Attributionmodelling. Available online at http://www.uniquedigital.de/fileadmin/content/file/whitepaper/uniquedigital_whitepaper_cross-channel-management.pdf, checked on January 20st, 2017.

Verhoef, Peter C.; Kannan, P. K.; Inman, Jeffrey J. (2015): From Multi-Channel Retailing to Omni-Channel Retailing. In Journal of Retailing 91 (2), pp. 174–181. DOI: 10.1016/j.jretai.2015.02.005.

Voorveld, Hilde A. M. (2011): Media Multitasking and the Effectiveness of Combining Online and Radio Advertising. In Computers in Human Behavior 27 (6), pp. 2200–2206. DOI: 10.1016/j.chb.2011.06.016.

Vuylsteke, Alexander; Wen, Zhong; Baesens, Bart; Poelmans, Jonas (2010): Consumers' Search for Information on the Internet. How and Why China Differs from Western Europe. In Journal of Interactive Marketing 24 (4), pp. 309–331. DOI: 10.1016/j.intmar.2010.02.010.

Webster, Jane; Watson, Richard T. (2002): Analyzing the past to prepare for the future: Writing a literature review. In MIS Quarterly 26 (2), pp. 13–23.

Whetten, David A. (1989): What Constitutes a Theoretical Contribution? In The Academy of Management Review 14 (4), p. 490. DOI: 10.2307/258554.

White, Tom (2012): Hadoop. The definitive guide; [storage and analysis at Internet scale]. 3. ed. Sebastopol Calif.: O'Reilly.

Whitener, Michael (2015): Cookies Are So Yesterday; Cross-Device Tracking Is In – SomeTips. Available online at https://iapp.org/news/a/cookies-are-so-yesterday-cross-device-tracking-is-insome-tips/, checked on January 11th, 2016.

Wiesel, Thorsten; Pauwels, Koen; Arts, Joep (2011): Practice Prize Paper —Marketing's Profit Impact. Quantifying Online and Off-line Funnel Progression. In Marketing Science 30 (4), pp. 604–611. DOI: 10.1287/mksc.1100.0612.

Woo, Jong Roul; Ahn, Joongha; Lee, Jongsu; Koo, Yoonmo (2015): Media channels and consumer purchasing decisions. In Industry Management & Data Systems 115 (8), pp. 1510–1528. DOI: 10.1108/IMDS-02-2015-0036.

Xu, Lizhen; Duan, Jason A.; Whinston, Andrew (2014): Path to Purchase. A Mutually Exciting Point Process Model for Online Advertising and Conversion. In Management Science 60 (6), pp. 1392–1412. DOI: 10.1287/mnsc.2014.1952.

Zhang, Ya; Wei, Yi; Ren, Jianbiao (2014): Multi-touch Attribution in Online Advertising with Survival Theory. In: 2014 IEEE International Conference on Data Mining (ICDM). Shenzhen, China, pp. 687–696.

6 Developing an Omni-Channel Ready Data Foundation

In the second publication data requirements are analyzed and identified. Building an omni-channel attribution approach enforces an omni-channel ready data foundation. The process of building such a data foundation, meeting all identified requirements by Nass et al. (2018) is described in the this chapter. The model development is presented in the following publication three in chapter 7. Cross-device, cross-platform user interaction data generated in different channels is utilized for the presented research. The raw data is provided by one of Germany's largest real-estate platforms.

6.1 Identifier Matching

All vendors providing the data for this research are independent. A link ability of all data sources is not given by default. Before the investigation begins and the data collection phase is executed, a linking attribute needs to be specified and implemented. Google Universal Analytics (GUA) is chosen to be the central data source. Within GUA so-called custom dimensions can be specified with individual content for different scopes on a hit-, user-, session-, etc. basis. The linking setup of all data sources is described in publication three (see chapter 7).

For the research two custom dimensions are defined: the first one holding the user's Tealium id, and the second one holding the corresponding intelliAd id. The data providing company already utilizes the tag management system Tealium iQ in different platforms such as website and mobile applications (apps) to manage their marketing tags. Therefore, the customer's device id within the Tealium context (Tealium id) is already present in the client context. This id is populated as a custom dimension in the GUA context.

The intelliAd id is not within the scope of the client. IntelliAd offers an id matching service which basically returns the id of the calling client. To set both, the Tealium id and the intelliAd id within the page view call, either of them need to be present at a very early state of the page load. A prior synchronic call to obtain the intelliAd id is not an option because such a setup will delay the tracking and can cause tracking errors. This may happen if a user exits the page too early. Since both custom dimensions are not bound to a *hit* but to a *user* scope the intelliAd id will be obtained during the first page visit, stored in a cookie and populated during the second call directly from the cookie. The script providing and populating both ids is placed in Appendix 4.

6.2 Data Transformation Process

The data requirement no. six (ability to calculate in real-time) enforces an at least semi-automated process to prepare the data on a regular basis. To satisfy this specification, an automated transformation process including extract, transform, and load steps (ETL) is developed to prepare the data for the attribution approach. This process is inspired by a data warehouse setup (Jordan et al. 2011).

An often-applied data warehouse structure (Jordan et al. 2011; Inmon 2005) consists of three different areas for loading data:

- Staging area (STG) – holding a copy of the raw data
- Cleansing area (CLS) – area for selecting, transformation and cleansing of data
- Core area (CO) – cleansed data for further processing

For each input source a separate workflow [01] to [05] is developed and implemented (see Figure 13). This enables a flexible handling of each individual data source. The applied suggested approach by Jordan et al. (2011) is a mixture of a bottom-up (the work-flow is built upon existing structures of external data sources) and a top-down where the target, the customer value and conversion probability (see chapter 7) is predefined. Data sources can be edited without influencing other sources, and this setup ensures a future-proof flexibility if new data sources need be added to, modified or removed from this setup.

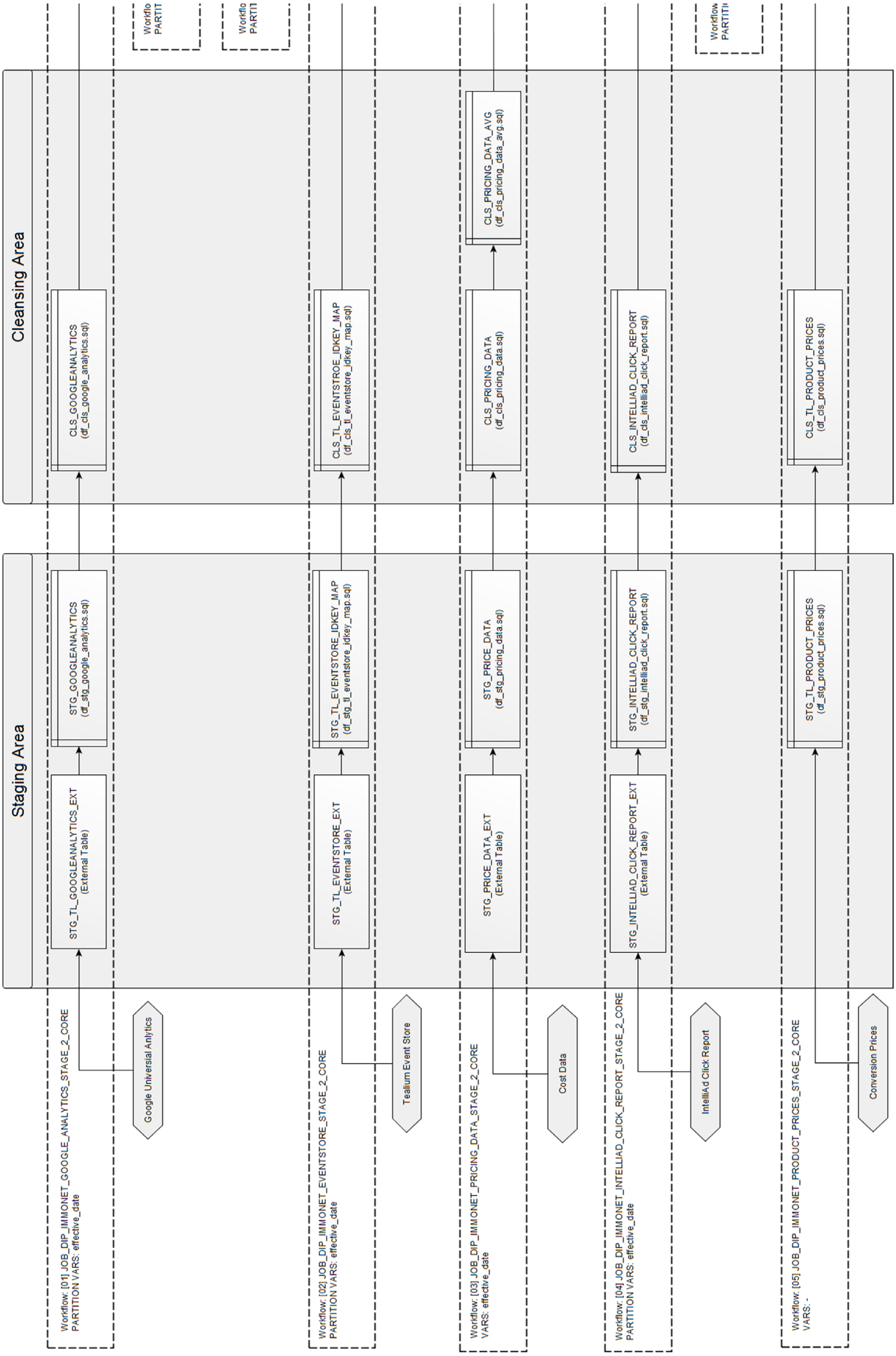

Staging Area
Cleansing Area
Workflow: [01] JOB_DIP_IMMONET_GOOGLE_ANALYTICS_STAGE_2_CORE
PARTITION VARS: effective_date
Google Universial Analytics
STG_TL_GOOGLEANALYTICS_EXT
(External Table)
STG_GOOGLEANALYTICS
(df_stg_google_analytics.sql)
CLS_GOOGLEANALYTICS
(df_cls_google_analytics.sql)
Workflow: [02] JOB_DIP_IMMONET_EVENTSTORE_STAGE_2_CORE
PARTITION VARS: effective_date
Tealium Event Store
STG_TL_EVENTSTORE_EXT
(External Table)
STG_TL_EVENTSTORE_IDKEY_MAP
(df_stg_tl_eventstore_idkey_map.sql)
CLS_TL_EVENTSTROE_IDKEY_MAP
(df_cls_tl_eventstore_idkey_map.sql)
Workflow: [03] JOB_DIP_IMMONET_PRICING_DATA_STAGE_2_CORE
VARS: effective_date
Cost Data
STG_PRICE_DATA_EXT
(External Table)
STG_PRICE_DATA
(df_stg_pricing_data.sql)
CLS_PRICING_DATA
(df_cls_pricing_data.sql)
CLS_PRICING_DATA_AVG
(df_cls_pricing_data_avg.sql)
Workflow: [04] JOB_DIP_IMMONET_INTELLIAD_CLICK_REPORT_STAGE_2_CORE
PARTITION VARS: effective_date
IntelliAd Click Report
STG_INTELLIAD_CLICK_REPORT_EXT
(External Table)
STG_INTELLIAD_CLICK_REPORT
(df_stg_intelliad_click_report.sql)
CLS_INTELLIAD_CLICK_REPORT
(df_cls_intelliad_click_report.sql)
Workflow: [05] JOB_DIP_IMMONET_PRODUCT_PRICES_STAGE_2_CORE
VARS: -
Conversion Prices
STG_TL_PRODUCT_PRICES
(df_stg_product_prices.sql)
CLS_TL_PRODUCT_PRICES
(df_cls_product_prices.sql)
Workflo
PARTIT
Workflo
PARTIT
Workflow
PARTITI

Figure 12: Top-down and bottom-up approach - import setup for all data sources divided up into workflows across all stages

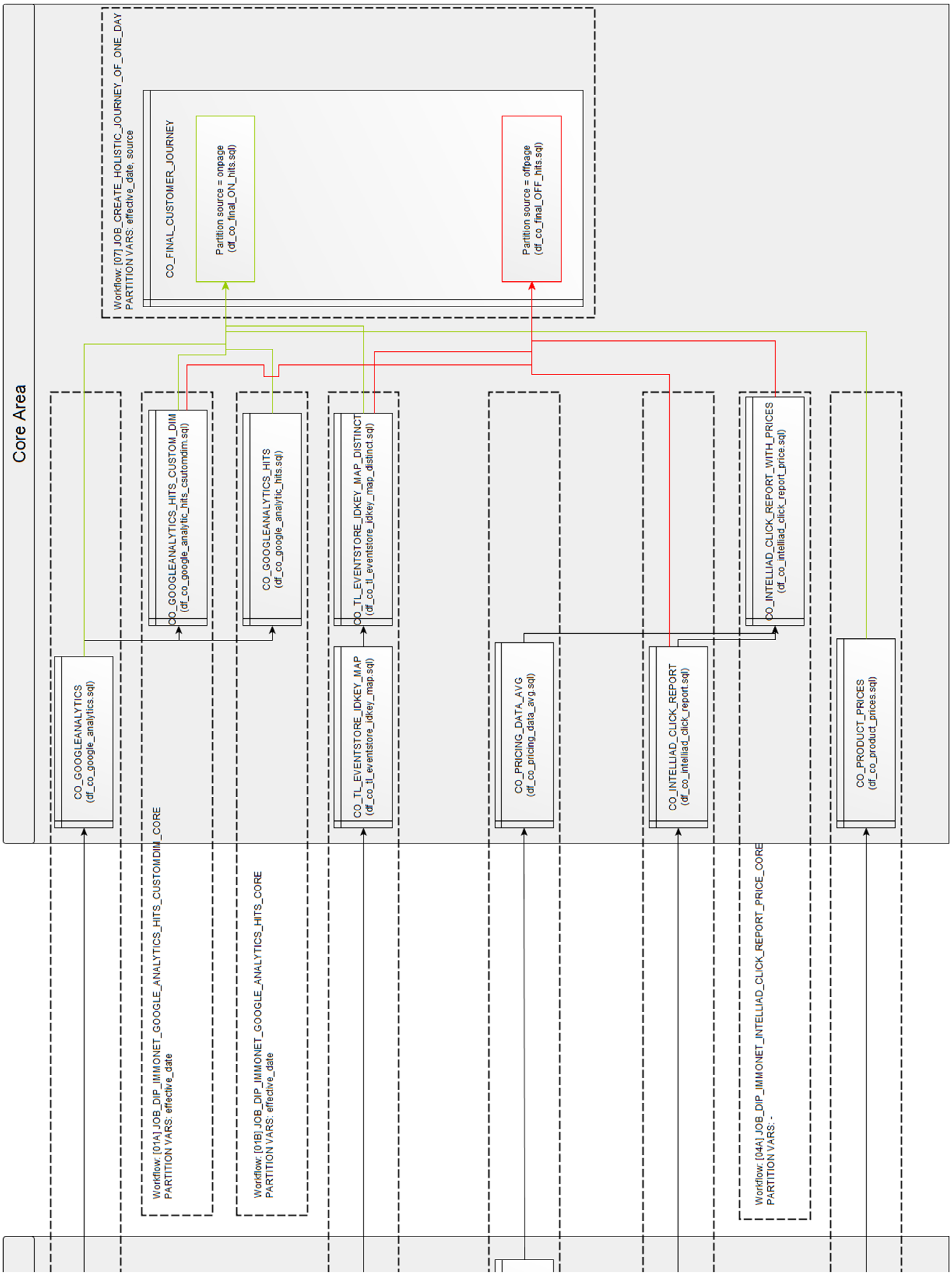

Core Area
Workflow: [07] JOB_CREATE_HOLISTIC_JOURNEY_OF_ONE_DAY
PARTITION VARS: effective_date, source
CO_FINAL_CUSTOMER_JOURNEY
Partition source = onpage
(df_co_final_ON_hits.sql)
Partition source = offpage
(df_co_final_OFF_hits.sql)
CO_GOOGLEANALYTICS
(df_co_google_analytics.sql)
CO_GOOGLEANALYTICS_HITS_CUSTOM_DIM
(df_co_google_analytic_hits_csutomdim.sql)
CO_GOOGLEANALYTICS_HITS
(df_co_google_analytic_hits.sql)
CO_TL_EVENTSTORE_IDKEY_MAP_DISTINCT
(df_co_tl_eventstore_idkey_map_distinct.sql)
CO_TL_EVENTSTORE_IDKEY_MAP
(df_co_tl_eventstore_idkey_map.sql)
CO_PRICING_DATA_AVG
(df_co_pricing_data_avg.sql)
CO_INTELLIAD_CLICK_REPORT
(df_co_intelliad_click_report.sql)
CO_INTELLIAD_CLICK_REPORT_WITH_PRICES
(df_co_intelliad_click_report_price.sql)
CO_PRODUCT_PRICES
(df_co_product_prices.sql)
Workflow: [01A] JOB_DIP_IMMONET_GOOGLE_ANALYTICS_HITS_CUSTOMDIM_CORE
PARTITION VARS: effective_date
Workflow: [01B] JOB_DIP_IMMONET_GOOGLE_ANALYTICS_HITS_CORE
PARTITION VARS: effective_date
Workflow: [04A] JOB_DIP_IMMONET_INTELLIAD_CLICK_REPORT_PRICE_CORE
PARTITION VARS: -

Figure 13: HCJ: Top-down and bottom-up approach - import setup for all data sources divided up into workflows across all stages (cont.)

The data transformation aims at a high-quality output. Information in terms of on-page hits or off-page touchpoints are neglected, if a corresponding data set is not available in other data sources. Only complete journeys are considered. It is not appropriate to calculate basic descriptive statistic values such as mean, mode, median, and corresponding quartiles on the individual fields of the raw data, since such numbers don’t add any value and are not target-oriented. Instead the derived journeys included in the holistic customer journey (HCJ) are described in publication 3 and in chapter 6.4. The HCJ is the name of the targeted data foundation including a holistic overview of the user’s behavior.

In the following, all workflows are briefly described across all areas (stage, cleanse, core). In the appendix all transformation scripts are attached as reference and detailed description. The entire data preparation and transformation process is designed for a daily run. Each workflow or sub-workflow requires the parameter *date* containing the date of the data to be processed. All workflows use MapReduce (Dean and Ghemawat 2004) as an engine.

6.2.1 Workflow [01]: GOOGLE_ANALYTICS_STAGE_2_CORE

General Description

Within workflow no. one, relevant information from the Google Universal Analytics’ (GUA) raw data is transformed onto the staging area for further processing. The raw data is being accessed through Google BigQuery (Google Inc. 2018c) and stored on a local server at the data providing company prior to this workflow.

This data source contains the on-page tracking and conversion data. Every 24 hours one compressed JSON-file (Json.org 2018; ECMA-262 Standard; ECMA-404 Standard) is being created containing the data from the previous day.

Raw Data Schema

The structure of the raw data is structured as a nested JSON-object per session of one user. The schema is documented by the data providing company Google (Google Inc. 2018a).

Areas

Staging area (run sql: Appendix 6)

The staging area's state is volatile. For every run of the workflow, the resulting table of the staging area is dropped before new data is being processed. An external table is created to access the raw data.

The JSON files containing the data is being selected and the whole json-object-string of one session is stored in the resulting table of this stage: `stg_googleanalytics`.

Cleansing area (run sql: Appendix 7)

The cleansing area's state is volatile, too. The pre-processed JSON-string is separated into individual columns. During this step the whole JSON-object itself is flattened and each information is stored as one column in the table `cls_googleanalytics`. Nested attributes get all their ancestors' attributes' name as a prefix separated by an underscore (_).

Core area (create sql: Appendix 8; run sql: Appendix 9)

The core area's state is, in contrast to the cleansing and staging area, persistent and not volatile. Processed data in this core area is kept. The table `co_googleanaytics` is partitioned by the date *date*. Therefore, for each date a separate partition is created. This setup enables a specific insert or update per day, not influencing data of other days.

6.2.2 Workflow [01A]: GOOGLE_ANALYTICS_HITS_CUSTOMDIM_CORE

General Description

Within the workflow 01A the Tealium *visitor_id* and the *intelliAd id* are extracted from the GUA custom dimensions string and attached to the corresponding GUA `fullVisitorId`. Since this workflow doesn't transform any raw input data, there is no action taken place in the staging area or in the cleansing area. The resulting table holds the linking information for all three data sources GUA, intelliAd click_report and Tealium EventStore.

Core area (create sql: Appendix 10; run sql: Appendix 11)

The resulting table `co_googleanalytics_hits_customdim` of this workflow is persistent and partitioned by date. Based on the data generated in the workflow [01], the GUA `fullVisitorId`, the intelliAd id and the tealium id are extracted. In the first step of the script the JSON string of the `hits_customdimensions` key is extracted and transformed into a sorted array. During this process only, the two custom dimensions holing the ids are utilized. Within the next steps the temporary data set is reduced to only one GUA `fullVisitorId` per device. Data sets with missing ids are neglected to meet the focus of the whole process. Only correct data sets are processed having both id populated.

6.2.3 Workflow [01B]: GOOGLE_ANALYTICS_HITS_CORE

General Description

The workflow [01B] extracts the on-page hits. The resulting table of workflow [01] is utilized to extract all conversion hits from all session at the pre-specified date *date*. As in workflow [01A] this workflow extracts no raw input data. Therefore, this workflow is only present in the core area.

Core area (create sql: Appendix 12; run sql: Appendix 13)

In this step the following information is extracted from the table **co_googleanaytics** into the table **co_googleanalytics_hits**:

- timing information,
- path information, and
- product information of conversion events.

All resulting hits are complemented by the GUA **fullVisitorId** and the GUA **visit_id**, where the latter one represents the session.

6.2.4 Workflow [02]: EVENTSTORE_STAGE_2_CORE

General Description

This workflow extracts the cross-device information from the raw data of the EventStore provided by Tealium Inc.

Raw Data Schema

The raw data is provided every 24 hours as a compressed JSON-file (Json.org 2018; ECMA-262 Standard; ECMA-404 Standard) in Amazon S3 cloud storage (Amazon Inc. 2006). In contrast to the GUA JSON-object the data structure provided by Tealium is flat and not nested. The object only consists of key-value-pairs. The schema of the file can be found in the Tealium Learning Community (Tealium Inc. 2015). The, for the presented research, relevant value the **customer_id_key_map** holds the data providing company's specific user unique key. This value is populated if the user enters the email address for a login or while performing other conversions. This value is the same for one customer on every used device. Based on this value the cross device / cross platform stitching is realized.

Staging area (run sql: Appendix 14)

From the raw data only the `event_id`, `visitor_id`, `eventtime` and `customer_idkey_map` are extracted. All information is stored into the table `stg_tl_eventstore_idkey_map` based on the defined date *date*. The `visitor_id` is representing a used device. If the data holds more `visitor_id`s for one `customer_idkey_map` the corresponding user uses different devices.

Cleansing area (run sql: Appendix 15)

No transformation is being performed within this step. The table `stg_tl_eventstore_idkey_map` is copied into the table `cls_tl_eventstore_idkey_map`.

Core area (run sql: Appendix 16, Appendix 17)

Within the core there are two separate steps performed. During the first step the data from the cleansing area is copied into the earmarked partition into the core area. The data is stored permanently. Afterwards, in the second step all `customer_idkey_map`s are identified for each `visitor_id`. If there is more than one `customer_idkey_map` per `visitor_id` (multiple login on one device) only the one id is kept and attached to the `visitor_id`. By doing so all devices with or without the populated `customer_idkey_map` are identified.

6.2.5 Workflow [03]: PRICING_DATA_STAGE_2_CORE

General Description

Within this workflow all pricing information is processed. This information will be attached to the corresponding off-page touchpoints performed in various online marketing channels. The actual assembling with the off-page touchpoints is performed in workflow [04A].

Raw Data Schema

The pricing information is provided by the online marketing department of the data providing company. Pricing information is available for Microsoft Bing, Google Adwords and Criteo. Based on the vendor, the data is aggregated in a different degree. Microsoft Bing's and Google Adwords' data is available on a keyword level on a daily basis. Criteo's data is only avaliable on a daily basis.

Staging area (create sql: Appendix 18, run sql: Appendix 19)

Next to the definition of an external table to access the raw data all data is being imported into the table `stg_price_data` for further processing.

Cleansing area (run sql: Appendix 20, Appendix 21)

Available data from the table **`stg_price_data`** is processed into the table **`cls_price_data`**. A transformation is performed for the values of the two columns **`price_date`** and **`cpc`** (costs per click). The **`price_date`** is formatted and the **`cpc`** is rounded to three digits.

In the second step missing pricing information per vendor are filled with the average cpc within a date, keyword and vendor.

Core area (run sql: Appendix 22)

Within the core area the table **`cls_pricing_data_avg`** is copied into **`co_pricing_data_avg`** for further processing.

6.2.6 Workflow [04]: INTELLIAD_CLICK_REPORT_STAGE_2_CORE

General Description

This workflow processes all off-page touch-points provided in the click_report by intelliAd.

Raw Data Schema

IntelliAd provided an export containing all off-page touchpoints in a csv list.

Staging area (run sql: Appendix 23)

An external table for accessing the data is created. Furthermore, the data provided for the date *date* is being loaded into the table **`stg_intelliad_click_report`**.

Cleansing area (run sql: Appendix 24)

The table **`stg_intelliad_click_report`** is copied into the table **`cls_intelliad_click_report`**. No transformation is performed.

Core area (run sql: Appendix 25)

The table **`cls_intelliad_click_report`** is processed and inserted into the table **`co_intelliad_click_report`** into the earmarked partition defined by *date* for further processing in workflow [4A].

6.2.7 Workflow [04A]: INTELLIAD_CLICK_REPORT_PRICE_CORE

General Description

This workflow processes the results from the previous workflows [03] and workflow [04]. Summarized, this workflow appends the pricing information (cpc) to all off-page touchpoints.

Core area (run sql: Appendix 26)

The table **`co_intelliad_click_report_with_prices`** is being filled at the earmarked date *date*. The resulting data set consists of the clickid provided by intelliAd, the cpc provided by the online marketing department and the date. The concatenation is performed in four steps. In the first step a keyword match with the match type *exact* is processed and the cpc is assigned, followed by a non-exact match type. In this second case the keyword may also consist of other terms. Within these two steps the pricing information only from Google Adwords and Microsoft Bing is attached. During the third step all cpc offered by Criteo are assigned to the corresponding touch-points. All other touch-points performed in online-marketing channels such as SEO, direct mail etc. don't cause any costs and the cpc is set to 0,0 Euro.

6.2.8 Workflow [05]: PRODUCT_PRICES_STAGE_2_CORE

General Description

This workflow processes the conversion prices provided by the company.

Staging area (run sql: Appendix 27)

Since there are only 20 products the prices are set manually in the script. The sql script creates the table **`stg_tl_product_prices`**.

Cleansing area (run sql: Appendix 28)

The table **`stg_tl_prodcut_prices`** is copied into **`cls_tl_product_prices`**.

Core area (run sql: Appendix 29)

The table **`cls_tl_product_prices`** is copied into **`co_tl_product_prices`** for further processing.

6.2.9 Workflow [06]: IMPORT_ONE_DAY_2_CORE

General Description

This workflow is a combination of previous workflows to load data from one pre-defined date *date* into the core for further processing.

6.2.10 Workflow [07]: CREATE_HOLISTIC_CUSTOMER_JOURNEY_OF_ONE_DAY

General Description

Workflow [07] processes the results of all previous workflows to build the *holistic customer journey* (HCJ) including off-page touchpoints with cost information attached and on-page conversions with prices on a hit basis. All hits/touchpoints can be ordered correctly by the column **`click_time`** which is extracted from both GUA and the intelliAd click_report. The cross-device information is attached to every single hit and touchpoint. The resulting data set represents the targeted *holistic customer journey*.

<u>Core area</u> (run sql: Appendix 30 and Appendix 31)

This workflow is split into two processes: *on-page hit transformation* and *off-page touch-point transformation*. During the *on-page hit transformation* the GUA hits are inserted into the table **`co_final_customer_journey`**. This table consists of two partitions. The first partition is the *source.* Possible values for *source* are *on-page* and *off-page*. The second partition is the already utilized *date*-partition. This setup enables a specific calculation based on *source* and *date*, not influencing any other information in this table. The resulting table **`co_final_customer_journey`**‘s columns can be separated into three different categories *combine*, *on-page* and *off-page*.

All stitching and timing information are included in the category *combine*. The *on-page* category consists of information from GUA holding conversion hits with attached price information.

IntelliAd touch-points with attached cost information are placed in the third category *off-page*. While processing off-page touch-points the category on-page is filled with **`Null`**-values and vice versa. This is because no on-page information is available in an off-page touchpoint.

The table **`co_final_customer_journey`** holds the HCJ and represents the basis for the following feature generating process.

6.3 Feature Generation

The holistic customer journey (HCJ) is the data foundation utilized for the following feature generating process. Table 17 lists all generated features including a definition. The feature generation script is placed in Appendix 32. The script consists of the calculation. Domain knowledge is the key driver for the feature definition process including feature extraction and feature selection (Meyer and Whateley Brendon 2004; Menkov et al. 2006). An automatic feature generating approach is not applied.

Table 17: Generated features for the machine learning approach

Id	Name of feature	Range of values	Definition
01	total_earnings	double	Sum of all conversions
02	total_spendings	double	Sum of all spendings
03	customer_value	double	Earnings – spendings
04	first_touch	timestamp	Timestamp of first entry in the journey
05	last_touch	timestamp	Timestamp of last entry in the journey
06	age_of_journey	integer	Days between first touch and last touch
07	customer_value_journey	double	Customer_value / age_of_journey
08	session_cnt	integer	Count sessions of journey (across all used devices)
09	is_logged_in	boolean	If device_customer_idkey_map is populated in journey
10	is_cross_device_user	boolean	Gua_device_devicecategory and off_hits_devicetype have two or more different entries. (*Annotation: if one user uses only two different mobile devices this value will not be set to true*)
11	avg_events_per_session	double	Number of all hits divided by the number of sessions. (*Annotation: off-page touch-points are included in the total number of hits, although they don't belong to a session*)
12	total_hit_cnt	integer	Count all on-page and off-page hits divided by all touch-points
13	overall_journey_cnt	integer	Count all journeys
14	overall_avg_earning _per_journey	double	Sum of all conversions divided by the number of journeys

15	overall_avg_spendings_per_journey	double	Sum of all spendings divided by the number of journeys
16	percentage_of_overall_mean_device_cnt_per_journey	double	Number of different devices of current journey divided by the number of devices of all journeys
17	percentage_of_overall_mean_session_cnt_per_journey	double	Number of sessions of current journey divided by the number of sessions of all journeys
18	device_array	Array<string>	Array of all used device categories in journey (e.g. mobile, tablet, desktop)
19	uses_desktop	boolean	Ture if hits accessed by a desktop device in journey exist
20	uses_mobile	boolean	Ture if hits accessed by a mobile device in journey exist
21	uses_tablet	boolean	Ture if hits accessed by a tablet device in journey exist
22	desktop_usage	double	Amount of desktop hits divided by all hits of journey
23	mobile_usage	double	Amount of mobile hits divided by all hits of journey
24	tablet_usage	double	Amount of tablet hits divided by all hits of journey
25	channel_array	Array<string>	Array with all used marketing channels from on-page source
26	cnt_channel	integer	Number of different channels used
27	cnt_earnings_events	integer	Number of conversions events
28	cnt_spendings_events	integer	Number of touchpoints with costs > 0
29	total_ratio_touchpoint_onsite	double	Number of on-page events divided by all hits

30	total_ratio _touchpoint_offsite	double	Number of off-page events divided by all hits
31	hits_1_2d	integer	Number of hits within the last two days before last hit
32	hits_3_4d	integer	Number of hits between three and four days before the last hit
33	hits_5_8d	integer	Number of hits between five and eight days before the last hit
34	hits_9_16d	integer	Number of hits between nine and 16 days before the last hit
35	hits_1_2s	integer	Number of sessions within the last two days before last hit
36	hits_3_4s	integer	Number of sessions between three and four days before the last hit
37	hits_5_8s	integer	Number of sessions between five and eight days before the last hit
38	hits_9_16s	integer	Number of sessions between nine and 16 days before the last hit
39	earnings_1_2d	double	Sum of earnings within the last two days before last hit
40	earnings_3_4d	double	Sum of earnings between three and four days before the last hit
41	earnings_5_8d	double	Sum of earnings between five and eight days before the last hit
42	earnings_9_16d	double	Sum of earnings between nine and 16 days before the last hit
43	earnings_1_2s	double	Sum of earnings within the last two sessions before last hit
44	earnings_3_4s	double	Sum of earnings between three and four sessions before the last hit

45	earnings_5_8s	double	Sum of earnings between five and eight sessions before the last hit
46	earnings_9_16s	double	Sum of earnings between nine and 16 sessions before the last hit
47	spendings_1_2d	double	Sum of spendings within the last two days before last hit
48	spendings_3_4d	double	Sum of spendings between three and four days before the last hit
49	spendings_5_8d	double	Sum of spendings between five and eight days before the last hit
50	spendings_9_16d	double	Sum of spendings between nine and 16 days before the last hit
51	customer_value_latest	double	earnings_1_2s – spendings_1_2d
52	last_device_category	double	Last device used
53	product_percent _blickfang	double	Earnings of product divided by all earnings
54	product_percent _brokercontact	double	Earnings of product divided by all earnings
55	product_percent _maklerempfehlung	double	Earnings of product divided by all earnings
56	product_percent _suchagent	double	Earnings of product divided by all earnings
57	product_percent _neubauanfrage	double	Earnings of product divided by all earnings
58	product_percent _phonecontact	double	Earnings of product divided by all earnings
59	product_percent _contact	double	Earnings of product divided by all earnings
60	product_percent _kataloghausbau	double	Earnings of product divided by all earnings
61	product_percent _tir	double	Earnings of product divided by all earnings

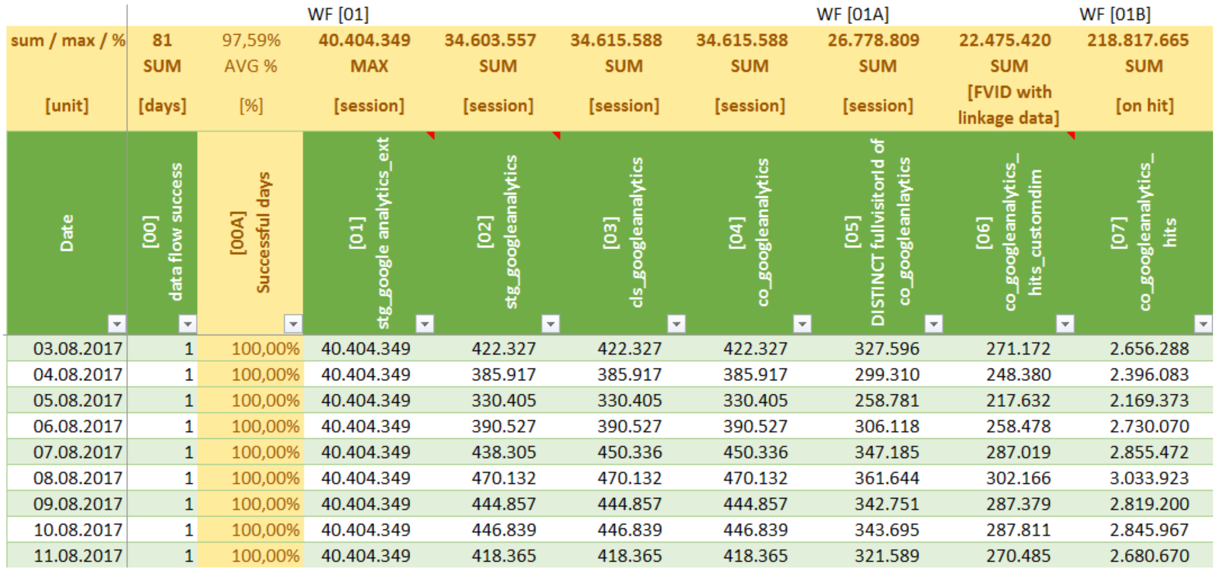
62	product_percent_gesuchcontact	double	Earnings of product divided by all earnings
63	product_percent_isa	double	Earnings of product divided by all earnings
64	product_percent_call	double	Earnings of product divided by all earnings
65	product_percent_immobewertung	double	Earnings of product divided by all earnings
66	product_percent_pia	double	Earnings of product divided by all earnings
67	product_percent_mailcontact	double	Earnings of product divided by all earnings
68	used_channels	Array<string>	Array with all used marketing channels from off-page source
69	used_channels_cleaned	Array<string>	Array with all used marketing channels with campaign id from off-page source
70	used_markets	Array<string>	Array of used markets

6.4 Descriptive Statistics of the HCJ

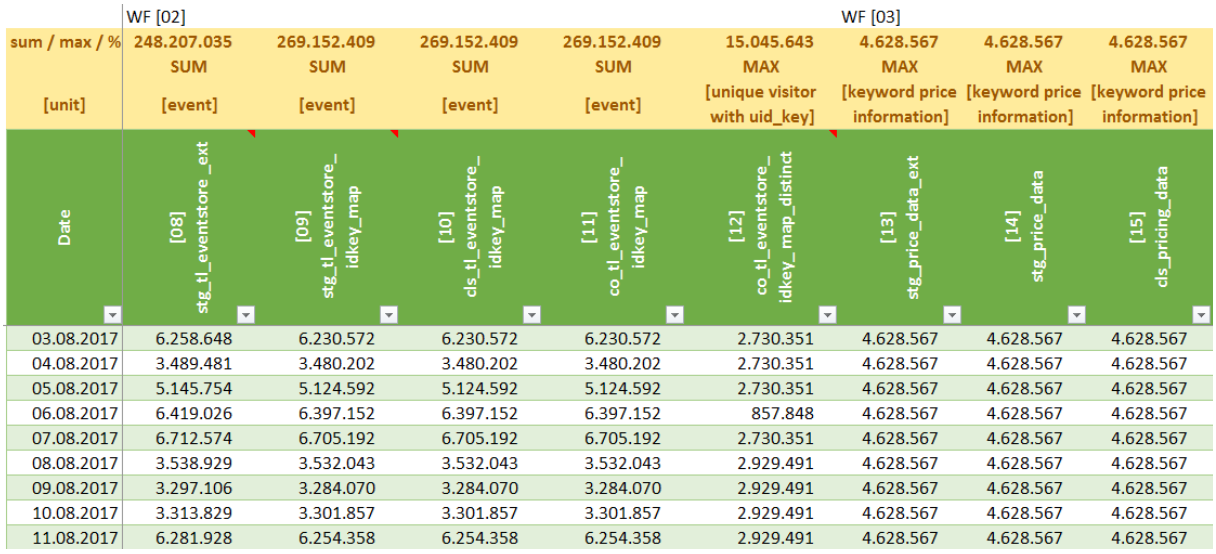
In total, about 3.5 TB interaction data is collected within a time-range of three months in 2017. All data sets aggregated consist of almost 240.000.000 hits/touchpoints from which over 225.000.000 hits are placed in more than 9.700.000 journeys. It is not purposeful to present values of descriptive statistics from the raw data. A combination of interaction information from various users in different channels from different platforms is meaningless.

An extract of the performance of the transformation process is illustrated in Table 18. All performance data is attached in Appendix 5. Table 18 contains the total numbers of successfully transformed data sets or a corresponding percentage from each workflow step. There is one column per workflow step. All transformation steps are ordered from the first workflow (left) to the last workflow on the right. Every day is represented by one row.

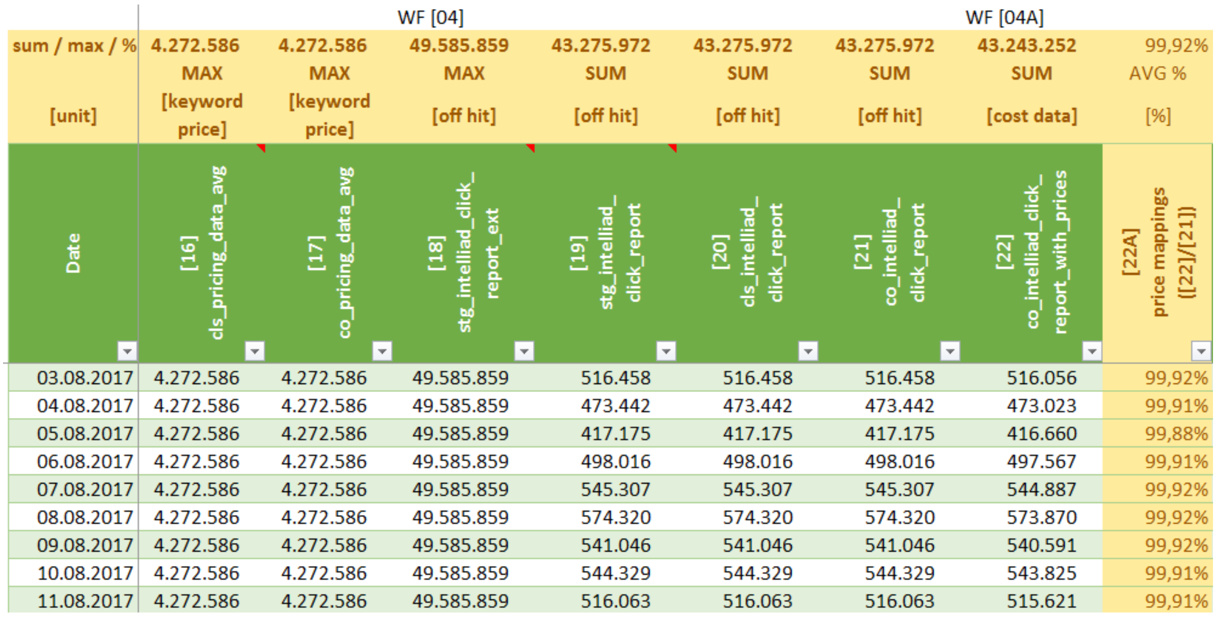
Due to corrupted files, two days have not been considered in this research. In total, 97,59% of the available data is utilized. Because of the transformation focus which only considers complete journeys (linkable data sets), 85,89% of all hits/touch points is transformed into journeys; 14,11% of the hits/touch points is neglected.



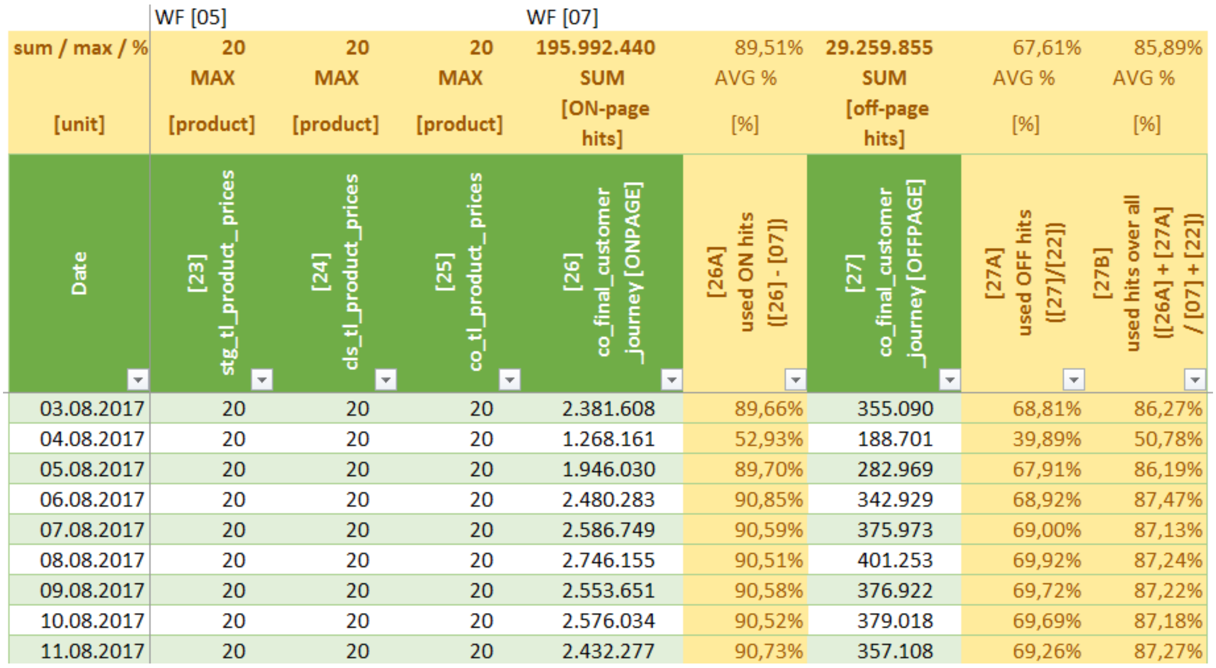
			WF [01]				WF [01A]		WF [01B]
sum / max / %	81	97,59%	40.404.349	34.603.557	34.615.588	34.615.588	26.778.809	22.475.420	218.817.665
	SUM	AVG %	MAX	SUM	SUM	SUM	SUM	SUM	SUM
[unit]	[days]	[%]	[session]	[session]	[session]	[session]	[session]	[FVID with linkage data]	[on hit]
Date	[00] data flow success	[00A] Successful days	[01] stg_google analytics_ext	[02] stg_googleanalytics	[03] cls_googleanalytics	[04] co_googleanalytics	[05] DISTINCT fullvisitorId of co_googleanlaytics	[06] co_googleanalytics_ hits_customdim	[07] co_googleanalytics_ hits
03.08.2017	1	100,00%	40.404.349	422.327	422.327	422.327	327.596	271.172	2.656.288
04.08.2017	1	100,00%	40.404.349	385.917	385.917	385.917	299.310	248.380	2.396.083
05.08.2017	1	100,00%	40.404.349	330.405	330.405	330.405	258.781	217.632	2.169.373
06.08.2017	1	100,00%	40.404.349	390.527	390.527	390.527	306.118	258.478	2.730.070
07.08.2017	1	100,00%	40.404.349	438.305	450.336	450.336	347.185	287.019	2.855.472
08.08.2017	1	100,00%	40.404.349	470.132	470.132	470.132	361.644	302.166	3.033.923
09.08.2017	1	100,00%	40.404.349	444.857	444.857	444.857	342.751	287.379	2.819.200
10.08.2017	1	100,00%	40.404.349	446.839	446.839	446.839	343.695	287.811	2.845.967
11.08.2017	1	100,00%	40.404.349	418.365	418.365	418.365	321.589	270.485	2.680.670



	WF [02]					WF [03]		
sum / max / %	248.207.035	269.152.409	269.152.409	269.152.409	15.045.643	4.628.567	4.628.567	4.628.567
	SUM	SUM	SUM	SUM	MAX	MAX	MAX	MAX
[unit]	[event]	[event]	[event]	[event]	[unique visitor with uid_key]	[keyword price information]	[keyword price information]	[keyword price information]
Date	[08] stg_tl_eventstore _ext	[09] stg_tl_eventstore_ idkey_map	[10] cls_tl_eventstore_ idkey_map	[11] co_tl_eventstore_ idkey_map	[12] co_tl_eventstore_ idkey_ map_distinct	[13] stg_price_data_ext	[14] stg_price_data	[15] cls_pricing_data
03.08.2017	6.258.648	6.230.572	6.230.572	6.230.572	2.730.351	4.628.567	4.628.567	4.628.567
04.08.2017	3.489.481	3.480.202	3.480.202	3.480.202	2.730.351	4.628.567	4.628.567	4.628.567
05.08.2017	5.145.754	5.124.592	5.124.592	5.124.592	2.730.351	4.628.567	4.628.567	4.628.567
06.08.2017	6.419.026	6.397.152	6.397.152	6.397.152	857.848	4.628.567	4.628.567	4.628.567
07.08.2017	6.712.574	6.705.192	6.705.192	6.705.192	2.730.351	4.628.567	4.628.567	4.628.567
08.08.2017	3.538.929	3.532.043	3.532.043	3.532.043	2.929.491	4.628.567	4.628.567	4.628.567
09.08.2017	3.297.106	3.284.070	3.284.070	3.284.070	2.929.491	4.628.567	4.628.567	4.628.567
10.08.2017	3.313.829	3.301.857	3.301.857	3.301.857	2.929.491	4.628.567	4.628.567	4.628.567
11.08.2017	6.281.928	6.254.358	6.254.358	6.254.358	2.929.491	4.628.567	4.628.567	4.628.567



			WF [04]				WF [04A]	
sum / max / %	4.272.586	4.272.586	49.585.859	43.275.972	43.275.972	43.275.972	43.243.252	99,92%
	MAX	MAX	MAX	SUM	SUM	SUM	SUM	AVG %
[unit]	[keyword price]	[keyword price]	[off hit]	[off hit]	[off hit]	[off hit]	[cost data]	[%]
Date	[16] cls_pricing_data_avg	[17] co_pricing_data_avg	[18] stg_intelliad_click_ report_ext	[19] stg_intelliad_ click_report	[20] cls_intelliad_ click_report	[21] co_intelliad_ click_report	[22] co_intelliad_click_ report_with_prices	[22A] price mappings ([22]/[21])
03.08.2017	4.272.586	4.272.586	49.585.859	516.458	516.458	516.458	516.056	99,92%
04.08.2017	4.272.586	4.272.586	49.585.859	473.442	473.442	473.442	473.023	99,91%
05.08.2017	4.272.586	4.272.586	49.585.859	417.175	417.175	417.175	416.660	99,88%
06.08.2017	4.272.586	4.272.586	49.585.859	498.016	498.016	498.016	497.567	99,91%
07.08.2017	4.272.586	4.272.586	49.585.859	545.307	545.307	545.307	544.887	99,92%
08.08.2017	4.272.586	4.272.586	49.585.859	574.320	574.320	574.320	573.870	99,92%
09.08.2017	4.272.586	4.272.586	49.585.859	541.046	541.046	541.046	540.591	99,92%
10.08.2017	4.272.586	4.272.586	49.585.859	544.329	544.329	544.329	543.825	99,91%
11.08.2017	4.272.586	4.272.586	49.585.859	516.063	516.063	516.063	515.621	99,91%



	WF [05]			WF [07]				
sum / max / %	20	20	20	195.992.440	89,51%	29.259.855	67,61%	85,89%
	MAX	MAX	MAX	SUM	AVG %	SUM	AVG %	AVG %
[unit]	[product]	[product]	[product]	[ON-page hits]	[%]	[off-page hits]	[%]	[%]
Date	[23] stg_tl_product_prices	[24] dls_tl_product_prices	[25] co_tl_product_prices	[26] co_final_customer journey [ONPAGE]	[26A] used ON hits ([26] - [07])	[27] co_final_customer journey [OFFPAGE]	[27A] used OFF hits ([27]/[22])	[27B] used hits over all ([26A] + [27A] / [07] + [22])
03.08.2017	20	20	20	2.381.608	89,66%	355.090	68,81%	86,27%
04.08.2017	20	20	20	1.268.161	52,93%	188.701	39,89%	50,78%
05.08.2017	20	20	20	1.946.030	89,70%	282.969	67,91%	86,19%
06.08.2017	20	20	20	2.480.283	90,85%	342.929	68,92%	87,47%
07.08.2017	20	20	20	2.586.749	90,59%	375.973	69,00%	87,13%
08.08.2017	20	20	20	2.746.155	90,51%	401.253	69,92%	87,24%
09.08.2017	20	20	20	2.553.651	90,58%	376.922	69,72%	87,22%
10.08.2017	20	20	20	2.576.034	90,52%	379.018	69,69%	87,18%
11.08.2017	20	20	20	2.432.277	90,73%	357.108	69,26%	87,27%

Table 18: Description of the transformation process towards the HCJ

The transformation process results in filling in the data into the final table **co_final_customer_journey**. This table holds all available interaction information per user of the given time range. Based on this data the journeys are calculated (see the script in Appendix 32), consisting of all features listed in Table 17. A statistical description of selected features is presented in Table 19.

Table 19 consists of the mean (mean), standard deviation (std), minimun (min), the 25th, 50th, 75th percentiles and maximum (max) of a selection of features. For binary values on a nominal level a percentage is declared representing the percentage of positive values in the data set. Only 6,46% (**is_cross_device_user**) of all users use different devices. This number is congruent with company internal records. Most customers prefer desktop computers (**uses_desktop** 48,506%) and mobile devices (**uses_mobile** 44,960%). Only 13,128% (**uses_tablet**) of the customers use tablets. All three percentages sum up to over 100% since some customers (6,46% **is_cross_device_user**) use more than one device. The actual percentage of used device classes is represented by **desktop_usage**, **mobile_usage** and **tablet_usage**. Most users consult only one marketing channel in the provided time period. The maximum number of used marketing channels is 8 (**cnt_channel**).

The mean of the total amount of earnings per journey is 25,54 Euro (**total_earnings**). The maximum is listed with 44.778 Euro. Journeys with extreme high values are analyzed separately. The HCJ consists of only few outliers. Such outliers represent customers converting with mainly *phone_contact* or *mail_contact* on several real-estate objects on almost every day within the recorded time period. For those conversions the user is not charged (see the description of the business model in publication 3 in chapter 7). Such a special behavior is rare, but possible. These journeys are kept since the data represents real behavior. Similar behavior can be identified on the other side of the spendings as well. The mean costs are stated with 8,3 Eurocent (`total_spendings`) per journey. The outlier journeys with over 81 Euros represent a possible behavior as well. Table 19 consists of the most important features and does not contain any timing features and product usage information features for reasons of clarity. A full list is attached in Appendix 33.

Table 19: Descriptive statistic values and distribution of selected features

feature name	mean	std	min	25%	50%	75%	max	percentage
total_earnings	25,540	78,781	0	3	9	22	44778	
total_spendings	0,083	0,317	0	0	0	0,082	81,509	
customer_value	25,456	78,684	-11,999	3	8,769	22	44778	
age_of_journey	11,366	21,644	0	0	0	11	91	
session_cnt	3,550	8,184	1	1	1	3	1115	
is_logged_in	0,063	0,243	0	0	0	0	1	
is_cross_device _user			0				1	6,459%
avg_events_per _session	5,629	6,804	1	2	4	7	1423	
total_hit_cnt	28,773	78,898	1	5	10	25	17156	
uses_desktop			0				1	48,506%
uses_mobile			0				1	44,960%
uses_tablet			0				1	13,128%
desktop_usage			0				1	41,946%
mobile_usage			0				1	45,344%
tablet_usage			0				1	12,710%
cnt_channel	1,275	0,571	1	1	1	1	8	
cnt_earnings _events	8,546	26,312	0	1	3	8	14926	
cnt_spendings _events	0,773	3,469	0	0	0	1	956	
total_ratio _touchpoint_onsite			0				1	81,580%
total_ratio _touchpoint_offsite			0				1	18,420%

7 Publication 3 Ready for Omni-Channel: Cross Device and Cross Platform Machine Learning Attribution Approach – A Field Experiment

The corresponding jupyter-notebook (content of the ipynb-file) including all scripts and instructions for the statistical course of research is attached in Appendix 33 and Appendix 34.

The process of identifying the optimal hyperparameter combination is documented in Appendix 35.

Ready for Omni-Channel: Cross Device and Cross Platform Machine Learning Attribution Approach – A Field Experiment	
DOI	Not assigned yet
Format	Journal article
Journal	Journal of Theoretical and Applied Electronic Commerce Research
Language	English
Status	Submitted (September 16th, 2018)
Abstract	**How much is a customer currently worth to a company?** The authors present an omni-channel attribution approach based on a cross device and cross platform data foundation utilizing machine learning. The approach enables attribution on user-level by providing the current customer value and a conversion probability. This attribution approach is designated to be omni-channel applicable as it fulfills the identified requirements and specifications by Nass et al. (2018).
Keywords	Omni-channel attribution, cross device cross platform attribution, dynamic attribution, machine learning attribution, omni-channel marketing

Introduction and Objectives

Is it reasonable to invest more money into the current customer by presenting advertisements or performing other marketing activities, or not? How much is this customer currently worth? These questions arise within the area of attribution in the field of (online-) marketing – one of the most important research priorities in the years 2016 to 2018 defined by the Marketing Science Institute (MSI 2016).

The objective of this article is to develop an omni-channel-ready attribution model utilizing machine learning (ML). As a novel approach the presented research is the first one developing an ML-attribution approach built on a cross-device and cross-platform data basis.

This article aims to provide the following contributions to the academic science community:

1. To present the first omni-channel ML-attribution approach based on a cross-device and cross-platform data foundation.
2. To analyze the practicability of the identified requirements and specifications for attribution modelling in an omni-channel environment (Nass et al. 2018).

This article is organized in the following way. Within the introduction the objectives and the theoretical background are defined. The latter focuses on two topics: the development of attribution in marketing and the importance of machine-learning in marketing. After describing the research methodology, the business model and pre-requirements defined by the data providing company are described and outlined to understand the field experiment. Since this investigation applies the identified requirements and specifications by Nass et al. (2018) the main investigation is split into *data foundation* and *attribution model*, focusing on the attribution model. The research is complemented by a presentation and discussion of the results including recommendations for practitioners and a conclusion.

Theoretical Background

On his way to purchase, the customer leaves various information about his behavior in different data sources. Tracing information is generated on a firm's website, within different mobile applications (apps) or other services provided by a company. Furthermore, data is collected about the users' behavior within different marketing channels, such as display advertisement, paid search, direct mailings and social networks. Econsultancy (2015) describes the widely applied strategic multi-channel approach whereby the communication and interaction with one's customers is realized within different independent marketing channels. Neslin et al. (2006) report that independent departments use mainly their own data to perform actions which are not synchronized with other channels.

How much is a customer currently worth to a company? This question can only be answered if information from different channels is linked within a company or institution. Verhoef et al.

(2015) describe the necessary shift from multi-channel to omni-channel in a retail context. A seamless experience (Carroll and Guzmán 2013) and the customer's value (Cook 2014) across all channels needs to be provided.

Towards Attribution in an Omni-Channel Environment

So-called static attribution approaches such as *last click/last interaction* are still widely applied (eMarketer 2016), although these rule-based models assigning credit to a certain source (channel) for conversions, sales, or leads are inaccurate (Petersen et al. 2009). These models assign credit based on static rules and neglect individual user behavior. Whole user sessions which do not lead to a conversion are disregarded as well (Petersen et al. 2009). Heuristic models are easy to implement, and their results are, in some circumstances, sufficient.

In a marketing context attribution models are often distinguished into *static* attribution models such as the already mentioned last click/last interaction approach and *dynamic* attribution models (Anderl et al. 2016a; Li and Kannan 2014; Shao and Li 2011).

Jayawardane et al. (2015) differentiate static attribution models in *simplistic* models and *rule-based* models. As the authors have a mathematical perspective, dynamic attribution models are termed *algorithmic* models. The two categories of algorithmic and dynamic attribution approaches are congruent.

The attribution problem is analyzed by several authors by applying different statistical approaches. Those models are placed in the latter category of dynamic or algorithmic attribution models. Abhishek et al. (2012), for example, build a model using a hidden Markov model to perform attribution; Li and Kannan (2014) developed a conceptual framework analyzing carryover- and spillover-effects across online marketing channels; Nottorf (2014) developed a solution based on a binary logit model with a Bayesian mixture approach; Shao and Li (2011) applied a bagged logistic regression model to the attribution problem; and Zhang et al. (2014) developed an additive hazard model base on survival theory. All mentioned dynamic approaches are developed or tested in a cross-channel environment. Neither model is applied onto a cross-device data basis, nor uses cross-platform information, e.g. linkable tracking data of users within different mobile apps and online portals. Next to other requirements, such a data foundation is claimed for an efficient attribution approach in an omni-channel environment (Nass et al. 2018). Nass et al. (2018) present data requirements and model requirements for an attribution approach in an omni-channel environment. Furthermore, based on a structured literature research conducted by the authors, an omni-channel attribution approach, meeting the identified requirements, has not been published (Nass et al. 2018). This research gap – the lack of an omni-channel attribution approach – is filled by the presented research.

Importance of Machine Learning for Marketing in Businesses

The available amount of relevant data in a marketing context increased rapidly within the last decade and is still growing. Large amounts of data need to be analyzed by different mathematical/statistical approaches. Already in 1995, Chen (1995) describe machine learning (ML) as a method for information retrieval in a business context. The increase of significance of ML for business (and marketing) in general is emphasized by Bose and Mahapatra (2001). In a marketing and business context ML became a powerful tool to gain insights within large and noisy data sources. Cui et al. (2006) evaluated the endogeneity bias in a RFM (Recency, Frequency, Monetary) model applying different ML approaches. Within the past few years the application of ML started growing rapidly in marketing research because of the availability of large scale data sources and low-cost cloud computing. More often generated models are applied within decision-making processes (Jordan and Mitchell 2015; Yousafzai et al. 2016).

Define Objective

The current research is guided by the two hypotheses listed in Table 20.

Table 20: Formulated hypotheses for the current research

Hypotheses		References
H1	It is possible to build a required data foundation and attribution model to work efficient in an omni-channel environment.	As a further development of static attribution approaches such as last click or last non-direct click (Google Inc. 2017; Jayawardane et al. 2015), there is a research focus on developing dynamic attribution approaches utilizing different statistical approaches (Abhishek et al. 2012; Anderl et al. 2016a; Dalessandro et al. 2012; Geyik et al. 2014; Li and Kannan 2014; Nottorf 2014; Shao and Li 2011; Xu et al. 2014; Zhang et al. 2014). Based on experts' interviews Nass et al. (2018) formulated data and model requirements / specifications for attribution in an omni channel environment. Furthermore, Nass et al. (2018) identified the current research gap – the lack of an omni-channel attribution approach. From here the first hypothesis is derived.
H2	If such a model can be developed, savings from at least 10% can be achieved for e.g. a company or an institution.	Performing an explorative study, one of Germany's largest real-estate platforms provided different data sources to verify hypothesis one. Furthermore, the company defined business model specific requirements (explained in detail within this publication). Taking those business model specific requirements into the research process enables a statement about possible savings for the specific case.

Nass et al. (2018) analyzed criteria for an attribution model to work efficiently in an omni-channel environment. The authors performed a structured literature research to identify relevant attribution models and applied the criteria. Based on their results there is no attribution model fulfilling the identified criteria. This identified research gap is filled by providing an omni-channel attribution approach within the current research. Lazaris and Vrechopoulos (2014) describe the necessity of research from multi-channel towards omni-channel in a retailing context. They call for research initiatives "that should investigate this topic [continuous change in retail practices] through multiple perspectives and approaches." (Lazaris and Vrechopoulos 2014).

To the best of the authors' knowledge the presented research is the first one applying a ML algorithm onto the marketing attribution problem fulfilling omni-channel requirements. Furthermore, this is the first approach using cross-device and cross-platform information in the research area of attribution in an online-marketing context.

Research Methodology

Academic research is conceived as a teleological process of knowledge acquisition (Köhler 1977). The attribution problem has a strong practical interest (Nass et al. 2018; Nottorf 2014; Anderl et al. 2016a). According to Ulrich (1995) the applied research design begins with practical problems which are unresolved. Those problems are analyzed by utilizing available literature and theories. His scientific approach conceives business studies as part of the application-orientated social science, which does not act in a static way, but considers change and alteration as instruments for creating design concepts of the future social reality (Ulrich 1981). Ulrich claims that business studies understood as applied science should be adjusted by problems related to the practice of corporation management (Ulrich et al. 1976; Ulrich 1981, 1985). By also considering the aforementioned research priorities defined by the MSI (MSI 2016), the attribution problem within the presented research is both, a scientific problem as well as a practice relevant problem which enables the course of the current examination.

The attribution problem is treated as a data mining problem, trying "[to mine] knowledge […] from data" (Han et al. 2012) and to solve it with ML algorithms.

The methodological approach for the current research is inspired by the in 1996 conceived Cross-Industry Standard Process for Data Mining (CRISP-DM) (Shearer 2000) and the Marketing-Analytics-Process (MAP) (Schoeneberg et al. (2017), see chapter 4) which is a specification of the CRISP-DM for (online-) marketing problems.

The applied CRISP-DM approach is widely spread in practice and science (Piatetsky 2014). The last poll about the applied methodology for analytics, data mining or data science projects indicates that almost half of the surveyed institutes decide to utilize the CRISP-DM. "The 6 high-level phases of CRISP-DM are still a good description for the analytics process, but the details and specifics need to be updated" (Piatetsky 2014)

Since attribution modelling is one research stream in the research field of (online-) marketing, the combination of CRISP-DM and the MAP approach, which focusses on marketing specific problems is chosen. In contrast to the more general CRISP-DM approach the MAP approach specifies six tangible phases. The six phases serve as orientation for the current research.

1. Problem statement and goal definition
2. Selection of data sources
3. Data preparation
4. Modeling
5. Model evaluation
6. Recommendation for action

In phase one a definition of a problem and the target of the research are required.

The problem: The research gap identified by Nass et al. (2018). The authors object to the lack of proper attribution approaches for an omni-channel marketing environment.

The goal: The development of an omni-channel attribution model. The model is supposed to calculate the customer value and predict if it is reasonable to invest more money into the customer, or not.

Business model of the data providing company

The data foundation underlying the research is provided by one of Germany's largest real estate platforms. The data set itself and the utilized data sources are described within the next chapters. To comprehend some of the decisions made during the research process, it is mandatory to understand the business model of the data providing company. In general, the company is providing different platforms such as a web-application and different mobile apps for people searching for various kinds of real-estate and corresponding products. Such corresponding products are: different lead services such as movement services, different kinds of artisans' firms, gardeners and other services helping people to get to relocate. Contact inquiries (contacting an agent indicating interest in a real-estate) via contact form or by telephone are a very important conversion type as well. If a user places a contact inquiry he/she doesn't pay for it. For these conversion types an internal value is defined to be able to handle those conversions within the marketing context in relation to other lead products. The actual money for those contact inquiries is earned from e.g. agents inserting their real-estate into the system, making it available to the public.

The data providing company defined the following requirements:

1. When predicting whether a customer has potential or not, it needs to be considered, that contact inquiries conversions do not provide earnings.
2. The prediction accuracy should be greater than 90%

These specifications were defined before the investigation began and are considered in the following steps two to six of the MAP methodology.

We treat the prediction aspect of the attribution problem as a statistical learning problem and attempt to solve it with ML algorithms. According to James et al. (2013) there are two main categories splitting up most of the statistical learning problems: *Supervised* problems and *unsupervised* problems. Supervised learning algorithms train a function which maps a feature set (input) onto a target (output) based on input-output-pairs (Russell and Norvig 2016). Such a data foundation is also termed as *labeled training data* which consists of training examples

(Mohri et al. 2012). In a ML context its common to say that a classifier is being created, used to generalize the problem for new instances (Kotsiantis 2007).

To solve unsupervised problems, algorithms are applied to find patterns (Bishop 2006) within the provided data, which in contrast to supervised learning problems does not consist of a pre-defined target.

Next to supervised and unsupervised problems there is a third category holding so-called *semi-supervised* problems (Chapelle et al. 2006). Problems belonging into this rather small group are characterized as followed: To build a model to solve a semi-supervised problem there are n observations available. During the data collection phase for m observations, where $m < n$, there are trainings information and a target collected. For $m – n$ observations, there is only training information (observations) without a target available (James et al. 2013).

Problems belonging to both, supervised and unsupervised problems, can be solved with different algorithms which can be separated per category into *categorical/discrete* and *continuous* algorithms.

For the current research both an unsupervised approach and a supervised approach are applied. A Principal Components Analysis (PCA) is utilized to reduce the dimensionality of the feature set. The goal of this step is to speed up the learning process by removing the least relevant information before training the model. Afterwards a supervised classification approach is utilized to predict whether an investment is reasonable, or not.

Data Foundation

Based on semi-structured interviews, Nass et al. (2018) identify different data requirements and specifications for attribution modelling in an omni-channel environment required by practitioners. Those identified requirements are applied as a basis for the utilized data foundation, which has been developed prior to this research.

Nass et al. (2018) identify the following twelve general data requirements for a data foundation to build on an attribution model within an omni-channel environment.

1. Data sources contain hard facts
2. Data sources contain soft facts
3. Highest possible data granularity
4. Stitch ability of a single user cross-device
5. Linkable data sources
6. Ability to calculate in real-time
7. Value calculation on user level
8. Value calculation on audience level
9. High-quality output
10. Ability to connect (third party) vendors directly
11. Interface driven design
12. Interface definition / standards
13. Plug and play

The above list consists of general data requirements which can be split up into two categories:

I. *concrete data requirements*, which have a direct impact on the data or data attributes and
II. *general data requirements*.

The latter category holds meta requirements which do not have a direct impact on the data itself. Requirements ten to 13 are placed in the second category. These requirements deal with the access ability and connectivity into existing systems.

The following Table 21 consists of requirements placed in category I) *concrete data requirements* and derived formulated requirements for the current analysis. All derived requirements (D1 to D6) refer to the targeted data foundation which represents the basis for the attribution approach.

Table 21: Conceived data requirements for the current analysis

No.	Requirement	Detailed requirements for the analysis
D1	[1] Data source contains hard facts	The data needs to consist of hard facts. Based on Nass et al. (2018) hard facts are defined as "real facts" such as tracking data, CRM data, DWH data etc.
D2	[2] Data source contains soft facts	The data needs to consist of soft facts. Based on Nass et al. (2018) "[soft facts] represent a meta-level of information which is derived from a user's behavior or situation. The assumption of a user's feelings, attitude or position is determined to be a soft fact".
D3	[3] Highest possible data granularity [7] Value calculation on user level [8] Value calculation on audience level [9] High-quality output	All utilized data sources need to be present in the highest degree of granularity [3] and quality [9] available. The data needs to consist of data on user level [7, 8].
D4	[4] Stitch ability of a single user cross-devices	The targeted data foundation needs to enable a stitch ability across different devices.
D5	[5] Linkable data sources	All utilized data sources need to have linking information.
D6	[6] Ability to calculate in real-time	In combination with the attribution model a calculation in real-time needs to be facilitated

Within the second phase specified by the MAP methodology for the research relevant, data sources need to be selected. Available data sources were analyzed. For this research all data sources listed in Table 22 are analyzed and selected for further processing.

Table 22: Used data source, loading frequency and access type

Data source	Data source provider	Frequency	Access Type
Google BiqQuery (Google Universal Analytics)	Google	Daily	Automatic pull
Click report	IntelliAd	Export	Manual (automatic pull possible)
Event store DB	Tealium	Daily	Automatic pull
Various vendors	Google AdWords, Bing, Criteo	Export	Manual
Pricing data	Internal	Export	Manual

The resulting data set is termed *holistic customer journey* (HCJ) as it includes all available information about one customer in a *holistic* way. The HCJ meets all six derived requirements listed in Table 21, except requirement D2. The data providing company does not hold any soft facts which are available for the investigation. The HCJ-ETL (extract, transform, load) process is inspired by a data warehousing setup schema. The process is separated into different areas (stage, cleanse, core) to ensure a regular run (Jordan et al. 2011).

Collecting Data

In the third phase of the MAP methodology, data needs to be prepared in terms of linkage and cleansing. Before the data was surveyed it had been ensured that the data sources are linkable. This was not possible by default. For each user within Google Universal Analytics (GUA) (Google Inc. 2018b) two custom dimensions are utilized to store a corresponding Tealium-Id (Tealium Inc. 2018) and a corresponding intelliAd-Id (intelliAd Media GmbH 2018) on a user-level scope.

Figure 14 illustrates the linkage implementation of how all used data sources are linked to each other.

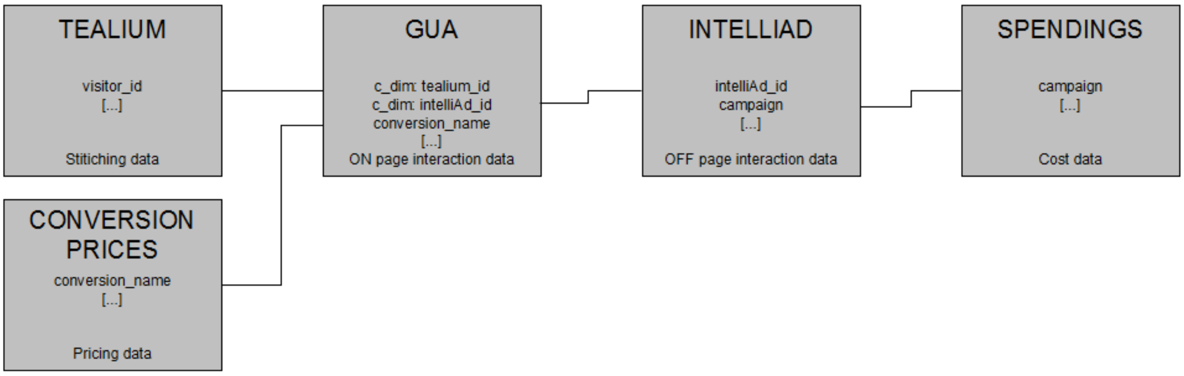


Figure 14: Linkage of all utilized data sources

Cross-Device / Cross-Platform

A cross-device stitching is achieved by a reliable method. Based on a login or a populated email address, the cross-device stitching is implemented. Having a user stitched across different devices is important to strengthen advertisement (Varan et al. 2013). Other methods of identifying a user across different devices such as statistical approaches or machine learning approaches as presented by Diaz-Morales (2015) are not applied.

The data foundation is also extended across different platforms. For the investigation usage information from the web portal and a mobile application is available. The stitching is achieved the same way as the cross-device stitching is implemented.

The HCJ Data Foundation in Numbers

In total about 3.5 TB of interaction data were collected in a time-range of three months. This raw data is processed towards the HCJ data foundation. All data sets aggregated consist of almost 240.000.000 hits/touchpoints from which over 225.000.000 hits were placed in more than 9.700.000 journeys.

For the HCJ-ETL-process a medium sized Hadoop-server (White 2015) managed by Amazon providing Hadoop User Experience (HUE) version 3.12.0 (Apache 2017), Apache Hive version 2.1.1 (Capriolo et al. 2012), Apache Oozie version 4.3.0 (Islam and Srinivasan 2015) and other applications is utilized for this research.

HCJ Summarized

The HCJ data foundation is characterized as follows:

On a user-level…

- on-site conversions are included.
- off-site interaction data, such as affiliate touchpoints, seo touchpoints, sea touchpoints banner clicks etc. are enclosed.
- cross-device stitching information are present. (Different devices used by one customer are connected.)
- cross-platform information is available.
- pricing information for off-site touchpoints are available
- earning information for on-site conversions are available

One customer journey includes all the above-mentioned information as a set of hits/ touchpoints which is utilized for the following feature generating process.

It is to be noted that all information relies on the correctness and completeness of the data providing company and corresponding third party vendors. Further research in this direction is beyond the scope of this research.

Feature Generation

To process the data into a ML approach the HCJ data needs to be transformed. One holistic customer journey, currently consisting of many hits/touchpoints (rows), needs to be transformed into a single row of data. Furthermore, relevant timing information of the HCJ needs to be transformed into features.

For the current research domain knowledge is the key driver for the feature definition process. This includes feature extraction and feature selection (Meyer and Whateley Brendon 2004; Menkov et al. 2006). An automatic feature generating approach is not applied.

In total 70 features and a target are selected, extracted and generated. All developed features can be separated into the following categories:

- **General journey information** such as first hit, age of journey, session count, hit count, etc.
- **Value information** such as long-term customer value, short-term customer value, total earning, total spendings, conversions, etc.
- **Used device information** such as the usage of mobile, tablet or desktop devices, cross-device user, etc.
- **Timing information** such as the development of earnings/spendings/hits within the last sessions/days.
- **Marketing touchpoint information** such as used channels etc.

Target Specification

Two pieces of information are requested by the (online-) marketing department for an attribution approach in an omni-channel environment, defined prior to the investigation.

1. customer value (conversion incomes minus the amount spent through marketing activities)
2. prediction: is it reasonable to invest more into the customer or not?

The current customer value is already available within the HCJ data foundation. Both, *total_earnings* and the *total_spendings* are already present. The *customer_value* is defined as followed:

$$customer\_value \ = total\_earnings - \ total\_spendings \qquad (1)$$

The second piece of information – whether it is reasonable to invest into a user or not – will be attained within the second part of the attribution model by utilizing a machine learning approach.

The company provided unlabeled data. This means no target is available. To develop a machine learning approach, the target *conversion_probability* needs to be defined. This is a binary classification problem. The range of possible values is either *True* - do invest in the customer or *False* - do not invest in the customer.

To model the *conversion_probability* the formula (2) is applied. Within the HCJ data foundation a short-term customer value (ST_CV) which basically consists of earnings and spendings which were performed within the last two days / sessions and the actual customer value (CV), which can be considered as a long-term customer value. If the ST_CV is negative the *conversion_probabilty_amount* will be multiplied by *-1* to make the result negative as it is explained in Table 23.

$$conversion\_probability\_amount\ (cpa) = \left|\frac{STCV}{CV}\right| \begin{cases} IF \quad ST\_CV < 0\ THEN * (-1) \\ ELSE \end{cases} \qquad (2)$$

Table 23: Conversion probability cases

No.	Condition	Explanation	Probability
1	ST_SC < 0 AND CV > 0	In short-term more money is spent than earned	↘
2	ST_SC < 0 AND CV < 0	Both, in long-term and short-term more money is spent than earned	↘
3	ST_SC > 0 AND CV < 0	In long-term more money is spent than earned in short-term.	↗
4	ST_SC > 0 AND CV > 0	Both, in long-term and short-term more money is earned than spent	↗

Example 1 (positive ST_CV)

ST_CV = 5,23€

CV = 7,56€

$$cpa = \left|\frac{5{,}23\ \text{€}}{7{,}56\ \text{€}}\right| = 0{,}692$$

Example 2 (negative ST_CV)

ST_CV = -0,83€

CV = 12,67

$$cpa = \left|\frac{-0{,}83\ \text{€}}{12{,}67\ \text{€}}\right| = 0{,}066 * (-1) = -0{,}066$$

Because of the business model the split between *True* (invest) and *False* (do not invest) for the *conversion_probability* is set at 0.50 of the *conversion_probability_amount* value. For other business models this split probably would be at 0.00 to separate the shortly positive developed customers correctly from the shortly negative developed ones. For the data providing company this is not an optimal split, because of the contact inquiry conversions which do not represent real income. A split at 0.00 would draw customers which are not relevant for the company into the group of positive customers. Runkler (2015) enables such a decision because domain experts are needed for the evaluation of features and the feature generation process to identify patterns.

Attribution Model

In the following the model development is described. These steps are placed within phase four of the MAP. This includes its training and testing.

Model Requirements

In previous research Nass et al. (2018) identified the following requirements for an attribution approach for an omni-channel environment.

1. Ability to handle hard facts
2. Ability to handle soft facts
3. Ability to add/remove data sources
4. Stitch ability cross-device
5. Calculation in real-time
6. Incremental learning process
7. Ability to predict future actions
8. Value calculation on user level
9. Value calculation on audience level
10. Machine learning / artificial intelligence approach
11. Data-driven calculation
12. High-quality output
13. Ability to connect third party vendors
14. Performance test of the model
15. Intuitive interface
16. Plug and play

Based on these collected requirements, 13 model requirements (see Table 24) are conceived for the presented investigation.

Table 24: Conceived model requirements for the current analysis

No.	Requirement	Detailed requirements for the analysis
M1	[01] Ability to handle hard facts	The whole process needs to be able to handle hard and soft facts, if present.
M2	[02] Ability to handle soft facts	
M3	[03] Ability to add/remove data sources	Data sources need to be exchangeable.
M4	[04] Stitch ability cross-device	The model needs to handle data from multiple devices used by one customer (stitching across different devices).
M5	[05] Calculation in real-time	The results of the attribution model need to be provided in real-time.
M6	[06] Incremental learning process	The model needs to become better over time.
M7	[07] Ability to predict future actions	The model needs to predict whether an investment is reasonable or not.
M8	[08] Value calculation on user level [09] Value calculation on audience level	The calculation needs to be on executed a user-level.
M9	[10] Machine learning / artificial intelligence approach [11] Data-driven calculation [14] Performance test of the model	The realization of the model needs to include a machine learning / artificial intelligence approach. Such an approach implies a data-driven calculation [11]. The model needs to be tested and validated [14]
M10	[12] High-quality output	The model prediction accuracy needs to be greater than 90%. This is a pre-requirement of the marketing department.

M11	[13] Ability to connect third party vendors [16] Plug and play	The setup needs to enable an integration of third party vendors. By integrating the model via a tag-management system this requirements is already fulfilled (see Nass et al. (2018)).
-	[15] Intuitive interface	An intuitive interface is not within the scope of the current research.

The model is implemented in python (Lutz 2017) within a jupyter-notebook (Toomey 2017). This technology enables a local development and a productive use in a cloud-based service such as Amazons AWS (Ryan 2018) or other.

Prepare Features and Target

All relevant features are stored in a matrix *X* and the raw target, the probability of investment, in a vector *y*. The distribution of the target *y* consists of about 60% *True* (positive) and 40% *False* (negative) samples.

Within the first step all features need to be standardized and categorial values need to be one-hot-encoded (Strand 2016; Richert and Coelho 2013; Müller and Guido 2017). Afterwards, a principal component analysis (PCA) is applied onto the feature matrix *X*. The resulting principal components (PC) are standardized and independent. A plot of the first two principal components is displayed in Figure 15. It points out, that the two target groups (red = *True* and blue = *False*) can probably be effectively separated. This indicates a high accuracy of the prediction from a given feature set *X* towards the target *y*.

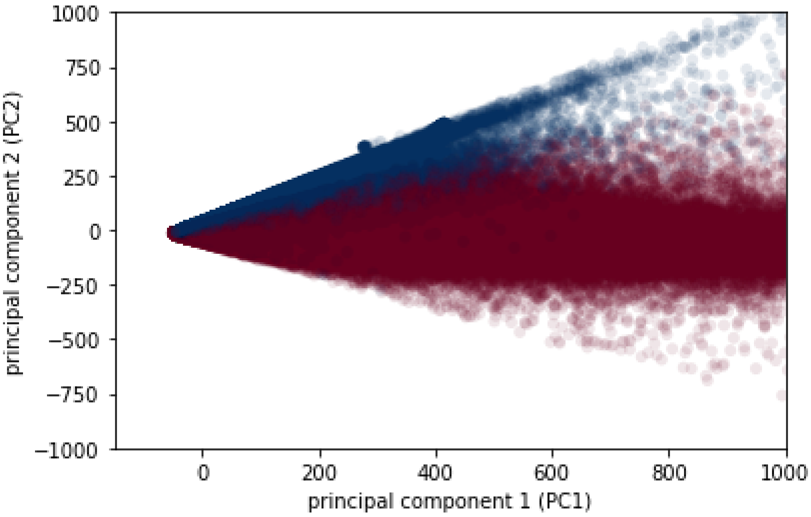


Figure 15: Plot of the first two principal components (PC). The color distinguishes between the two categories: invest / not invest.

Dimension Reduction – Increasing Performance

Figure 16 displays the cumulative variance of the components. This plot indicates that about four PCs cover already over 95% of the variance of the data. All other PCs hold only 5% of the information and are not relevant for further processing because they only add little variance (information value).

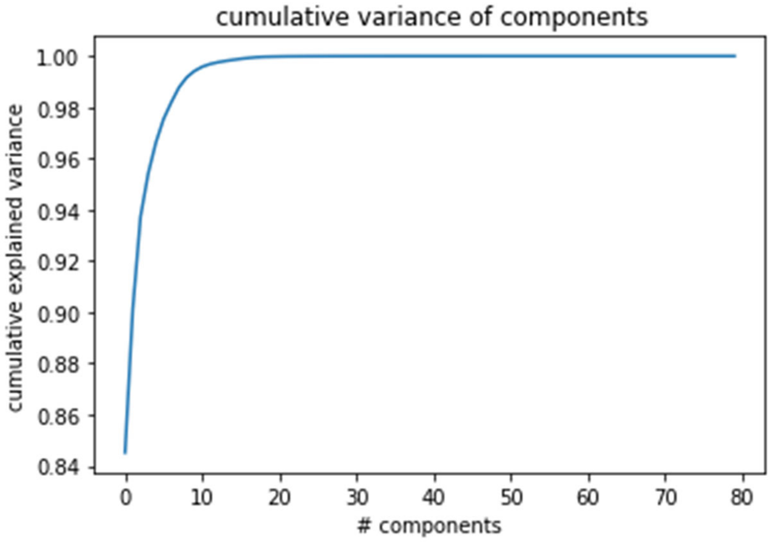


Figure 16: Cumulative variance of all principal components (PC)

These PCs add mainly noise which decreases the quality of the model (Jolliffe 2002). Removing non-relevant PCs will speed up the model training because the model is trained only with the most important PC. For the current research a cut after PC four is reasonable. A general cumulative value of variance is in between 70% - 90% of variance (Jolliffe 2002; Cangelosi and Goriely 2007). The decision of where to cut off the PC is domain-specific. Due to the demanded accuracy of more than 90%, a cut at 95% (four PCs) is made.

Since it is not required to analyze the impact degree of each feature with respect to the results, PCs can be utilized as an input for the machine learning model. Therefore, a dimension reduction PCA is applied onto the initial feature matrix *X*. Only the first four most important PCs are kept.

To develop the best performing classifier the following four tree-based algorithms are selected.

- Random Forest
- Extra Trees
- Ada Boost
- Gradient Boosts

Random Forest and Extra Tree can utilize multiple CPU cores during training, Ada Boost and Gradient Boost are single-threaded. For the current research the accuracy and the training speed is relevant. Only tree-based algorithms are selected because such algorithms are robust and accurate.

Define Classification Metric

A suitable metric for the classification task to be selected. For the present research a metric is required which takes both the correctness of the model and the quality of the output into consideration. The F1-score (3) is a harmonic mean of *precision* and *recall* developed for this requirement (Sasaki 2007; Chicco 2017) and defined as follows:

$$F1 = 2 * \frac{precision*recall}{precision+recall} \tag{3}$$

By valuing both correctness of the model and the true positive rate, the F1-score is a meaningful metric for this problem. The F1-score accuracy is limited to numbers between 0 (worst) to 1 (best).

Data Split

Splitting the data correctly is the foundation to ensure that the results of the model are reliable. “The accuracy of a classier C is the probability of correctly classifying a randomly selected instance” (Kohavi 1995). To ensure a reliable result the data needs to be split multiple times. The matrix *X* and the target vector *y* are split randomly into a *training* set and a *test* set (see Figure 17 Split 1). The test set consists of 20% of the whole data set (Chicco 2017; Boulesteix 2015). The *training* set is split again (see Figure 17 Split 2) into a *training-training* set and a *training-validation* set. The latter one consists of 20% of the *training* split. Only data within the *training* split is utilized for the following model training. The *test* split is not used until the final accuracy test of the developed model.

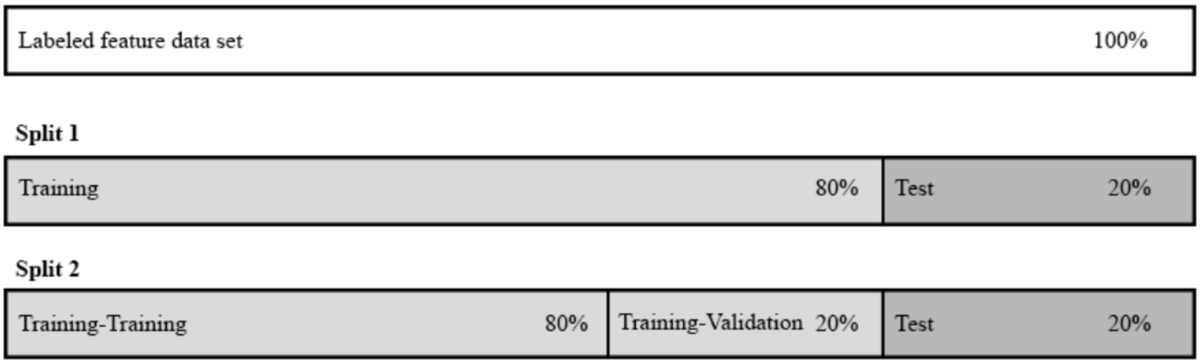


Figure 17: Data split into training set, validation set and test set

Identify Best Classifying Algorithm

The above-mentioned algorithms Random Forest, Extra Trees, Ada Boost and Gradient Boost are applied onto the *training* split. A grid-search is performed for optimizing the hyperparameter *n_estimators*, representing the number of internal trees, to identify the optimal algorithm. In a range from 10 to 190 in steps of 90 all algorithms are trained and tested to identify which algorithm outputs the best performing classifier. During this training process a cross-validation with two folds is performed (Hsu et al. 2003; Liu 2009; Breiman et al. 1984). Since the mean of all F1-scores is used as a performance indicator a small amount of folds is sufficient.

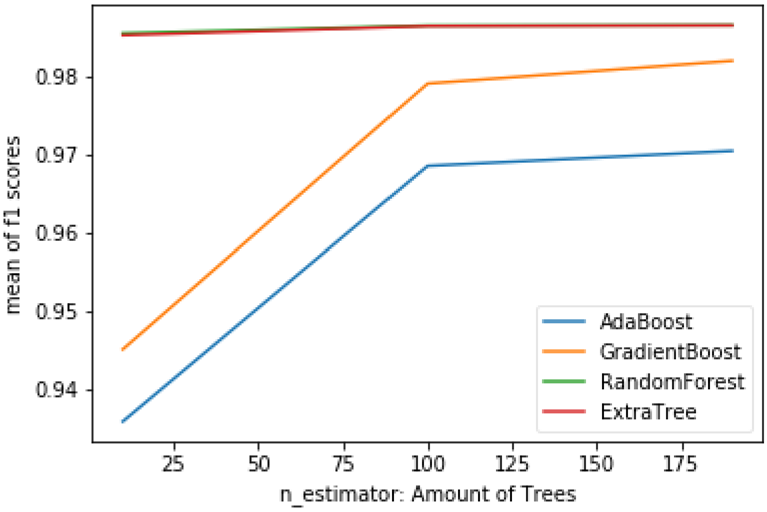


Figure 18: Accuracy of the applied algorithms Random Forest, Extra Tree, Ada Boost, and Gradient Boost

The tuning process is illustrated in Figure 18. As a second evaluation criterion the calculation time has been defined. The detailed F1-scores and the corresponding calculation time for each algorithm is presented in Table 25.

Table 25: Accuracy and trainings duration of the applied algorithms Random Forest, Extra Tree, Ada Boost and Gradient Boost

Algorithm	**Duration** **(three runs with two folds)**	**Number of estimators**	**Mean of accuracy**
Ada Boost	42 minutes 43 seconds	10	0.93576633
		100	0.96855738
		190	0.97048073
Gradient Boost	53 minutes 3 seconds	10	0.94503966
		100	0.97913457
		190	0.98206418
Random Forest	**7 minutes 41 seconds***	10	0.98559905
		100	0.98659783
		190	**0.98664157***
Extra Tree	10 minutes 9 seconds	10	0.98535537
		100	0.98653484
		190	0.9866157

In terms of prediction accuracy, Random Forest and Extra Tree are performing best. Random Forest is about 20% faster than the Extra Tree algorithm. The Random Forest algorithm is selected since it performs best in terms of accuracy and speed for the current research.

Identify Best Hyperparameter Configuration

With the next step the optimal hyperparameter configuration needs to be determined. To obtain reliable results the second split (see Figure 17) is performed. This split divides the data of the *training* set into a *training-training* set (80%) and a *training-validation* set (20%). All hyperparameters need to be defined before the training starts, because those higher-level properties cannot be learned by the algorithm directly from the training phase (Chicco 2017).

For this research no automatic hyperparameter learning approach such as Auto-Sklearn (Auto-Sklearn Development Team 2018), Auto-Weka (Kotthoff et al. 2017), TPOT (Olson et al. 2016), or PennAI (Olson et al. 2017) is applied, because it does not serve the verification of the hypotheses H1 and H2.

Depending upon the selection of the hyperparameters, time for training and testing can strongly vary (Claesen and Moor 2015). Most of the performance variation can be achieved by tuning only a few hyperparameters (Rijn and Hutter 2018; Claesen and Moor 2015; Hutter et al. 2014). The three hyperparameters listed in Table 26 are used to tune the model during the training-process. *n_estimators* and *max_depth* are chosen because they have a direct impact on the accuracy. *max_depth* has a strong impact on the training speed. Large trees are more accurate but slower. Finally, *min_samples_split* has a direct impact on how the tree evolves during training.

Table 26: Hyperparameters to tune with value ranges

No.	Hyperparameter	Value Range	Number of values	Description
1	`min_samples_split`	1, 2, 4, 8, 10	5	Representing the minimum number of samples required to split an internal node.
2	`max_depth`	10, 12, 14, 16, 20, 25, 50	7	Maximum depth of the tree.
3	`n_estimators`	10, 50, 90, 130, 170, 210, 250, 290, 330	9	The number of estimators (internal trees).

During the tuning process an iteration over all permutations of hyperparameter combinations (5 * 7 * 9 = 315) is executed to identify the optimal configuration. Within each iteration a classifier is trained, and the accuracy is tested by having the model predict all samples of the *training-training* data set (same data the classifier is trained with) and the *training-validation* data set. The following hyperparameter configuration producing the lowest *test_error* on the *training-validation* set is selected to be the configuration producing the classifier with the highest accuracy:

Algorithm:	Random Forest
min_samples_split:	8
max_depth:	20
n_estimators:	290

Model Accuracy

The evaluation of the model is placed in the fifth phase of the MAP methodology. During this phase the classifier is trained with the above-mentioned configuration. The classifier's accuracy is obtained by having the classifier predict the samples within the *test* set from the first split (see Figure 17). This *test* set has not been used for the development or training of the model. This ensures a reliable accuracy of the model.

Results and Discussion

The **main objective** of the current research consists of developing an omni-channel ready attribution approach based on a cross-device and cross-platform data foundation. We successfully developed an attribution approach calculating the customer value and predicting if a future investment is reasonable, or not. Both the accuracy of the customer value and the prediction depends on correct data relying on average tracking challenges (Nottorf and Funk 2013; Varan et al. 2013; Whitener 2015) and login behavior. Provided a user has entered the credentials on all used devices and the tracking is properly executed, the accuracy of the customer value is very precise, since all earnings and spendings are included in the HCJ. If the user uses only one device, the population of credentials is not necessary since one device is treated as one journey by default.

To analyze **hypotheses 1** all data requirements *D_* (see Table 21) and model requirements *M_* (see Table 24) are combined and listed in Table 27. Table 27 consists of a column *Implemented / Realized* which indicates whether a requirement is implemented or realized in this research (✓ **OK**), or not (× ERR).

Table 27: Verification of the implementation of derived data requirements for the current analysis

No.	**Requirement**	**Implemented/ Realized**	**Description**
D1	[1] Data source contains hard facts	✓ **OK.**	GUA contains hard facts such as conversion data.
D2	[2] Data source contains soft facts	× ERR.	The data providing company doesn't provide such information.
D3	[3] Highest possible data granularity [7] Value calculation on user level [8] Value calculation on audience level [9] High-quality output	✓ **OK.**	All utilized data sources are present in the highest degree of granularity [3] and quality [9] available. The data is on a user level [7, 8].
D4	[4] Ability to stitch a user cross-device	✓ **OK.**	
D5	[5] Linkable data sources	✓ **OK.**	
D6	[6] Ability to calculate in real-time	✓ **OK.**	This is realized in combination with the model integration.
M1	[01] Ability to handle hard facts	✓ **OK**.	Soft facts were not available. The model would be able to handle soft facts by modifying or extending the feature set. The ETL process could be extended as well.
M2	[02] Ability to handle soft facts	✓ **OK.**	The model can process such information. Due to the lack of soft facts, this has not been verified.
M3	[03] Ability to add/remove data sources	✓ **OK.**	The ETL process allows an exchange of data sources.
M4	[04] Stitch ability cross-device	✓ **OK.**	The ETL process and feature generation process includes usage data from different devices belonging to one customer.

M5	[05] Calculation in real-time	✓ **OK**.	An integration of the model directly through the tag-manager allows a call towards the attribution application responding with the required information in real-time.
M6	[06] Incremental learning process	✓ **OK**.	Short-term perspective: The more input data the more journeys to train the model. Long-term perspective: Based on the estimation of the model, future conversions need to be weighted by the prediction. If a prediction is not correct, this information needs to be processed by extending the feature set. The model uses this information for training as a new feature.
M7	[07] Predict future actions	✓ **OK**.	The model predicts whether an investment is reasonable or not.
M8	[08] Value calculation on user level [09] Value calculation on audience level	✓ **OK**.	The calculation is on user-level; an aggregation on audience-level is possible.
M9	[10] Machine learning / artificial intelligence approach [11] Data-driven calculation [14] Performance test of the model	✓ **OK**.	A machine-learning approach is applied [10] and trained, based on dynamic data [11] and correctly trained, verified and tested [14].

M10	[12] High-quality output	✓ **OK.**	The model predicts better than the pre-required 90%.
M11	[13] Ability to connect third party vendors [16] Plug and play	✓ **OK**.	The prediction information is available in the scope of the client and can be processed to perform actions based on the prediction towards third party vendors. By integrating the model via a tag-management system these requirements are already fulfilled (see Nass et al. (2018)).

With the attribution approach developed, the first objective of the current research can be studied and the first hypotheses H1 can be verified. Although the data providing company didn't provide any soft facts the model and the corresponding ETL process are able to handle such information.

The **second hypotheses** H2 can be verified as well. The prediction accuracy of the developed prediction model-component, being part of the whole attribution model, is 98,4%. The corresponding error (1 – accuracy) is 1,6%. Assuming 30% (in the given data set there were about 40%) of the less relevant users are removed from marketing campaigns already 40% of the invested money can be used more efficiently to correspond with relevant users or simply saved by not serving any activities to those 40%.

Recommendations for Practitioners

Since the defined problem in phase one of the MAP methodology is solved, the methodology foresees recommendations for actions which are placed in the final phase.

The **second objective** of the research is to analyze the practicability of the identified requirements. The data providing company has an adequate setup to collect data about the customer in different channels. Although the existing data collection setup can be rated as advanced, several changes were necessary before the data collecting phase for this research could be executed. For instance, a linking information had to be added into the different data sources to enable this research. Building an attribution approach upon company internal data sources still is a challenging task. Bell et al. (2014) describe necessary steps to be successful in an omni-channel world. Based on the experiences and findings of the presented research the requirements and specifications defined by Nass et al. (2018) turned out to be a helpful general guideline for practitioners on their way to an omni-channel setup. Those high-level requirements combined with a specifying methodology of the CRISP-DM such as the utilized MAP methodology (Schoeneberg et al. (2017), see chapter 4) is a helpful guide for improving the data setup of a company or an institution to improve marketing outcomes.

Attribution Model Extension and Integration

Associated with one customer, the current developed attribution model consists of the two values *customer_value* and *conversion_probability.* Other possible values are *next_best_channel,* holding the customers preferred marketing channel, or *next_best_product,* analyzing which product is relevant to the customer. There is more information which can be used to optimize the attribution process. The build ETL-process in combination with the prediction model can be extended easily by adding different values by further models. All models will be integrated in one attribution service which can be called by different applications such as a mobile app or the web application. Each application posts a request holding a user identifier which is being processed within the service to identify the correct journey. This piece of information is utilized to call all other models and the prediction approach. All calculated values will be returned into the application and can be processed for further marketing decisions directly in the client.

Conclusion

Although it turned out to be more complex and challenging than expected to create the required data foundation to build the attribution model, the results indicate a successful research. All objectives have been met and both hypotheses have been verified.

The developed attribution approach allows savings by removing irrelevant customers from marketing activities or by shifting the user towards cheaper channels such as direct mail. A meaningful budget-shift from not relevant users to company-relevant users is feasible as well.

The current attribution approach is a further development of previously presented dynamic attribution approaches (Abhishek et al. 2012; Anderl et al. 2016a; Dalessandro et al. 2012; Geyik et al. 2014; Li and Kannan 2014; Nottorf 2014; Shao and Li 2011; Xu et al. 2014; Zhang et al. 2014) focusing on an application in an omni-channel environment. As the shift towards omni-channel marketing is already in progress, further attribution approaches will be developed to increase attribution quality. By providing the first omni-channel ready attribution approach, new research areas are identified. What user-specific information is relevant for attribution? In the current research two values *customer_value* and *conversion_proability* are implemented. Other values such as *next_best_channel* or *next_best_product* could be relevant as well. What information about a customer, his behavior or his attitudes are relevant for attribution in an omni-channel environment? What impact do the individual values have? The existing research stream of targeting and attribution should grow together. Both need detailed information about the user to serve the user's needs.

The key aspect of developing a successful omni-channel ready attribution approach is placed in the used data foundation. The developed data foundation (HCJ) and model should be an encouragement for practitioners and science experts to start analyzing additional information about the current customer to be more precise in knowing what the customer expects, wants and needs. A future-proofed marketing setup can interact on an individual basis with the customer across different channels considering the customer's intention and his value. This enables meaningful actions from an attributional (financial) and marketing perspective.

References

Abhishek, Vibhanshu; Fader, Peter; Hosanagar, Kartik (2012): Media Exposure Through the Funnel. A Model of Multi-Channel Attribution. Available online at http://dx.doi.org/10.2139/ssrn.2158421.

Anderl, Eva; Becker, Ingo; Wangenheim, Florian von; Schumann, Jan Hendrik (2016a): Mapping the Customer Journey. Lessons Learned from Graph-Based Online Attribution Modeling. In International Journal of Research in Marketing 33 (3), pp. 457–474. DOI: 10.1016/j.ijresmar.2016.03.001.

Apache (2017): Apache Porject: Hue. Apache. Available online at http://gethue.com/, checked on August 2nd, 2018.

Auto-Sklearn Development Team (2018): Auto-Sklearn. Available online at https://github.com/automl/auto-sklearn, updated on June 18th, 2018, checked on August 1st, 2018.

Bell, David R.; Gallino, Santiago; Moreno, Antonio (2014): How to Win in an Omnichannel World. Fall 2014, Vol. 56 No.1. MITSloan (MITSloan Management Review, 1).

Bishop, Christopher M. (2006): Pattern Recognition and Machine Learning. New York, NY: Springer Science+Business Media LLC (Information Science and Statistics). Available online at http://dx.doi.org/10.1007/978-0-387-45528-0.

Bose, Indranil; Mahapatra, Radha K. (2001): Business Data Mining — A Machine Learning Perspective. In Information & Management 39 (3), pp. 211–225. DOI: 10.1016/S0378-7206(01)00091-X.

Boulesteix, Anne-Laure (2015): Ten Simple Rules for Reducing Overoptimistic Reporting in Methodological Computational Research. In PLoS computational biology 11 (4), e1004191. DOI: 10.1371/journal.pcbi.1004191.

Breiman, Leo; Friedman, Jerome; Stone, Charles J.; Olshen, Richard A. (1984): Classification and Regression Trees. 1st ed. Boca Raton: CRC Press. Available online at https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=5109080.

Cangelosi, Richard; Goriely, Alain (2007): Component Retention in Principal Component Analysis with Application to cDNA Microarray Data. In Biology direct 2 (2). DOI: 10.1186/1745-6150-2-2.

Capriolo, Edward; Nash, Courtney; Wampler, Dean; Rutherglen, Jason; Loukides, Michael Kosta; Demarest, Rebecca (Eds.) (2012): Programming Hive. 1st ed. Sebastopol, CA: O'Reilly & Associates. Available online at http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10759003.

Carroll, Donald; Guzmán, Inés (2013): The New Omni-Channel Approach to Serving Customers. Strategy Implications for Communications Service Providers. Available online at https://www.accenture.com/be-en/~/media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Industries_2/accenture-new-omni-channel-approach-serving-customers.pdf, checked on 8/2/2018.

Chapelle, Olivier; Schölkopf, Bernhard; Zien, Alexander (Eds.) (2006): Semi-Supervised Learning. Cambridge, Mass.: MIT Press (Adaptive computation and machine learning).

Chen, Hsinchun (1995): Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms. In J. Am. Soc. Inf. Sci. 46 (3), pp. 194–216. Available online at https://doi.org/10.1002/(SICI)1097-4571(199504)46:3<194::AID-ASI4>3.0.CO;2-S.

Chicco, Davide (2017): Ten Quick Tips for Machine Learning in Computational Biology. In BioData mining 10, p. 35. DOI: 10.1186/s13040-017-0155-3.

Claesen, Marc; Moor, Bart De (2015): Hyperparameter Search in Machine Learning. Available online at http://arxiv.org/pdf/1502.02127v2, checked on August 3rd, 2018.

Cook, Glenn (2014): Customer Experience in the Omni-Channel World and the Challenges and Opportunities This Presents. In Journal of Direct, Data and Digital Marketing Practice 15 (4), pp. 262–266. DOI: 10.1057/dddmp.2014.16.

Cui, Geng; Wong, Man Leung; Lui, Hon-Kwong (2006): Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. In Management Science 52 (4), pp. 597–612. DOI: 10.1287/mnsc.1060.0514.

Dalessandro, Brian; Perlich, Claudia; Stitelman, Ori; Provost, Foster (2012): Causally Motivated Attribution for Online Advertising. In ADKDD '12 Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy, pp. 1–9. DOI: 10.1145/2351356.2351363.

Diaz-Morales, Roberto (2015): Cross-Device Tracking. Matching Devices and Cookies. In Peng Cui (Ed.): 15th IEEE International Conference on Data Mining Workshop (ICDMW). Atlantic City, NJ, USA. Institute of Electrical and Electronics Engineers; IEEE. Piscataway, NJ: IEEE, pp. 1699–1704.

Econsultancy (2015): Quarterly Digital Intelligence Briefing. Digital Trends 2015. Available online at https://www.marketingsociety.com/sites/default/files/QDIB%20Adobe%20Digital%20Trends%20Report%202015_EMEA_0.pdf, checked on December 12th, 2017.

eMarketer (2016): Primary Attribution Model Used by Their Marketing Team to Measure Performance According to US B2B Marketers. Available online at emarketer.com, checked on January 20th, 2016.

Geyik, Sahin Cem; Saxena, Abhishek; Dasdan, Ali (2014): Multi-Touch Attribution Based Budget Allocation in Online Advertising ADKDD'14, pp. 1–9. DOI: 10.1145/2648584.2648586.

Google Inc. (2017): Attribution Modeling Overview. Assign Credit for Sales and Conversions to Touchpoints in Conversion Paths. Available online at https://support.google.com/analytics/answer/1662518?hl=en, checked on January 21st, 2017.

Google Inc. (2018b): Google Analytics. Available online at https://analytics.google.com/analytics/web/, checked on August 10th, 2018.

Han, Jiawei; Kamber, Micheline; Pei, Jian (2012): Data mining. Concepts and techniques. 3. ed. Amsterdam: Elsevier/Morgan Kaufmann (The Morgan Kaufmann series in data management systems). Available online at https://ebookcentral.proquest.com/lib/subhh/detail.action?docID=729031.

Hsu, Chih-Wei; Chang, Chih-Chung; Lin, Chih-Jen (2003): A Practical Guide to Support Vector Classification. Available online at https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, updated on May 19th, 2016, checked on August 3rd, 2018.

Hutter, Frank; Hoos, Holger; Leyton-Brown, Kevin (2014): An Efficient Approach for Assessing Hyperparameter Importance. In Eric P. Xing, Tony Jebara (Eds.): Proceedings of the 31st International Conference on Machine Learning, vol. 32. Bejing, China: PMLR (Proceedings of Machine Learning Research), pp. 754–762. Available online at http://proceedings.mlr.press/v32/hutter14.html.

intelliAd Media GmbH (2018): intelliAd. Available online at https://www.intelliad.de/impressum/, checked on August 10th, 2018.

Islam, Mohammad Kamrul; Srinivasan, Aravind (2015): Apache Oozie. [the workflow scheduler for hadoop]. Beijing: O´Reilly.

James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2013): An Introduction to Statistical Learning. With Applications in R. New York, NY: Springer (Springer Texts in Statistics, 103). Available online at http://dx.doi.org/10.1007/978-1-4614-7138-7.

Jayawardane, Himani W.; Halgamuge, Sage. K.; Kayande, Uka. (2015): Attributing Conversion Credit in an Online Environment: An Analysis and Classification. In: 2015 3rd International Symposium on Computational and Business Intelligence (ISCBI). Bali, Indonesia, pp. 68–73.

Jolliffe, Ian T. (2002): Principal Component Analysis. 2. ed. New York, NY: Springer (Springer series in statistics). Available online at http://www.loc.gov/catdir/enhancements/fy0817/2002019560-d.html.

Jordan, Claus; Schnider, Dani; Wehner, Joachim; Welker, Peter (2011): Data Warehousing mit Oracle. Business Intelligence in der Praxis. Munich: Hanser.

Jordan, Michael I.; Mitchell, Tom M. (2015): Machine Learning: Trends, Perspectives, and Prospects. In Science (New York, N.Y.) 349 (6245), pp. 255–260. DOI: 10.1126/science.aaa8415.

Kohavi, Ron (1995): A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In International Joint Conference on Articial Intelligence 14 (1).

Köhler, Richard (Ed.) (1977): Empirische und handlungstheoretische Forschungskonzeptionen in der Betriebswirtschaftslehre. Bericht über d. Tagung in Aachen, März 1976. Verband der Hochschullehrer für Betriebswirtschaft. Stuttgart: Poeschel.

Kotsiantis, Sotiris (2007): Supervised Machine Learning: A Review of Classification Techniques. In Informatica (Ljubljana) 31.

Kotthoff, Lars; Thornton, Chris; Hoos, Holger H.; Hutter, Frank; Leyton-Brown, Kevin (2017): Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. Journal of Machine Learning Research 18 (25), pp. 1–5. Available online at http://jmlr.org/papers/v18/16-261.html.

Lazaris, Christos; Vrechopoulos, Adam (2014): From Multichannel to "Omnichannel" Retailing: Review of the Literature and Calls for Research. In 2nd International Conference on Contemporary Marketing Issues, (ICCMI), 18-20 June 2014, At Athens, Greece. DOI: 10.13140/2.1.1802.4967.

Li, Hongshuang; Kannan, P. K. (2014): Attributing Conversions in a Multichannel Online Marketing Environment. An Empirical Model and a Field Experiment. In Journal of Marketing Research 51 (1), pp. 40–56. DOI: 10.1509/jmr.13.0050.

Liu, Ling (Ed.) (2009): Encyclopedia of Database Systems. New York, NY: Springer (Springer reference). Available online at http://deposit.dnb.de/cgi-bin/dokserv?id=2820703&prov=M&dok_var=1&dok_ext=htm.

Lutz, Mark (2017): Learning Python. 5th ed. Beijing, Boston, Farnham, Sebastopol, Tokyo: O'Reilly.

Menkov, Vladimir; Dayanik, Aynur; Lewis, David D.; Madigan, David; Genkin, Alexander (2006): Constructing Informative Prior Distributions from Domain Knowledge in Text Classification, pp. 493–500. DOI: 10.1145/1148170.1148255.

Meyer, Tony A.; Whateley Brendon (2004): SpamBayes: Effective Open-Source, Bayesian Based, Email Classification System., pp. 493–500.

Mohri, Mehryar; Rostamizadeh, Afshin; Talwalkar, Ameet (2012): Foundations of Machine Learning. Cambridge, Mass., London: The MIT Press (Adaptive computation and machine learning).

MSI, Marketing Science Institute (2016): Research Priorities 2O16–2O18. MSI Marketing Science institute. Available online at https://www.msi.org/uploads/articles/MSI_RP16-18.pdf, checked on November 3rd, 2017.

Müller, Andreas C.; Guido, Sarah (2017): Introduction to Machine Learning with Python. A Guide for Data Scientists. First edition. Sebastopol, CA: O'Reilly Media. Available online at http://proquest.tech.safaribooksonline.de/9781449369880.

Nass, Ole; Garrigós, José A.; Gómez, Hermengildo G.; Schoeneberg, Klaus-Peter (2018): Attribution modelling in an omni-channel environment. new requirements and specifications from a practical perspective. In Int. J. Electronic Marketing and Retailing.

Neslin, Scott A.; Grewal, Dhruv; Leghorn, Robert; Shankar, Venkatesh; Teerling, Marije L.; Thomas, Jacquelyn S.; Verhoef, Peter C. (2006): Challenges and Opportunities in Multichannel Customer Management. In Journal of Service Research 9 (2), pp. 95–112. DOI: 10.1177/1094670506293559.

Nottorf, Florian (2014): Multi-Channel Attribution Modeling on User Journeys. In: E-Business and Telecommunications, vol. 456, pp. 107–125.

Nottorf, Florian; Funk, Burkhardt (2013): A Cross-Industry Analysis of the Spillover Effect in Paid Search Advertising. In Electron Markets 23 (3), pp. 205–216. DOI: 10.1007/s12525-012-0117-z.

Olson, Randal S.; Sipper, Moshe; La Cava, William; Tartarone, Sharon; Vitale, Steven; Fu, Weixuan et al. (2017): A System for Accessible Artificial Intelligence. Available online at http://arxiv.org/pdf/1705.00594v2.

Olson, Randal S.; Urbanowicz, Ryan J.; Andrews, Peter C.; Lavender, Nicole A.; La Kidd, Creis; Moore, Jason H. (2016): Automating biomedical data science through tree-based pipeline optimization. Available online at http://arxiv.org/pdf/1601.07925v1.

Petersen, Andrew J.; McAlister, Leigh; Reibstein, David J.; Winer, Russell S.; Kumar, V.; Atkinson, Geoff (2009): Choosing the Right Metrics to Maximize Profitability and Shareholder Value. In Journal of Retailing 85 (1), pp. 95–111. DOI: 10.1016/j.jretai.2008.11.004.

Piatetsky, Gregory (2014): CRISP-DM, Still the Top Methodology for Analytics, Data Mining, or Data Science Projects. KDnuggets. Available online at https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html, checked on August 3rd, 2018.

Richert, Willi; Coelho, Luis Pedro (2013): Building Machine Learning Systems with Python. Master the Art of Machine Learning with Python and Build Effective Machine Learning Systems with this Intensive Hands-on Guide. Birmingham: Packt Publ (Community experience distilled).

Rijn, Jan N. van.; Hutter, Frank (2018): Hyperparameter Importance Across Datasets. In Yike Guo, Faisal Farooq (Eds.): Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18. the 24th ACM SIGKDD International Conference. London, United Kingdom, 8/19/2018 - 8/23/2018. New York, New York, USA: ACM Press, pp. 2367–2376.

Runkler, Thomas A. (2015): Data Mining. Modelle und Algorithmen intelligenter Datenanalyse. 2nd ed. Wiesbaden: Springer Vieweg (Lehrbuch).

Russell, Stuart J.; Norvig, Peter (2016): Artificial Intelligence. A Modern Approach. With assistance of Ernest Davis, Douglas Edwards. 3rd ed. Boston, Columbus, Indianapolis, New York, San Francisco: Pearson (Always learning).

Ryan, Mike (2018): AWS System Administration: Best Practices for Sysadmins in the Amazon Cloud: O'Reilly Media, Incorporated.

Sasaki, Yutaka (2007): The Truth of the F-Measure. With assistance of School of Computer Science, University of Manchester. Available online at https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf, checked on August, 1st 2018.

Schoeneberg, Klaus-Peter; Nass, Ole; Schmitt, Lennart (2017): Marketing-Analytics-Process (MAP). Data-Driven-Marketing-Projekte erfolgriech durchführen. In Christopher Zerres (Ed.): Handbuch Marketing-Controlling. Grundlagen - Methoden - Umsetzung, 4th Ed. 4., vollständig überarbeitete Auflage. Berlin, Germany: Springer Gabler, pp. 15–39.

Shao, Xuhui; Li, Lexin (2011): Data-Driven Multi-Touch Attribution Models. In Chid Apte, Joydeep Ghosh, Padhraic Smyth (Eds.): the 17th ACM SIGKDD international conference. San Diego, California, USA, pp. 258–264.

Shearer, Colin (2000): The CRISP-DM Model: The New Blueprint for Data Mining. In Journal of data warehousing 5 (4), pp. 13–22.

Strand, Håkon H. (2016): What is One-hot Encoding and When is it Used in Data Science? Available online at https://www.quora.com/What-is-one-hot-encoding-and-when-is-it-used-in-data-science, updated on December 20th, 2016, checked on August 2nd, 2018.

Tealium Inc. (2018): Tealium Inc. Available online at https://tealium.com/, checked on August 10th, 2018.

Toomey, Dan (2017): Jupyter for Data Science. Exploratory Analysis, Statistical Modeling, Machine Learning, and Data Visualization with Jupyter. Birmingham, Mumbai: Packt.

Ulrich, Hans (1981): Die Betriebswirtschaftslehre als Anwendungsorientierte Sozialwissenschaft. In M. Geist, R. Köhler (Eds.): Die Führung des Betriebes. Stuttgart: Kohlhammer, pp. 1–25.

Ulrich, Hans (1985): Von der Betriebswirtschaftslehre zur Systemorientierten Managementlehre. In R. Wunderer (Ed.): Betriebswirtschaftslehre als Management- und Führungslehre. Stuttgart: Kohlhammer, pp. 3–32.

Ulrich, Hans (1995): Von der Betriebswirtschaftslehre zur systemortientierten Managementlehre. Wunderer, R., Betriebswirtschaftslehre als Management-und Führungslehre. 3rd Edition. Stuttgart, Germany: Schäffer-Poeschel.

Ulrich, Hans; Krieg, Walter; Mallik, Fredmund (1976): Zum Praxisbezug einer systemorientierten Betriebswirtschaftslehre. In: Zum Praxisbezug der Betriebswirtschaftslehre - in wissenschaftlicher Sicht. Bern / Stuttgart: Ulrich, Hans, pp. 135–151.

Varan, Duane; Murphy, Jamie; Hofacker, Charles F.; Robinson, Jennifer A.; Potter, Robert F.; Bellman, Steven (2013): What Works Best When Combining Television Sets, PCs, Tablets, or Mobile Phones? In JAR 53 (2), pp. 212–220. DOI: 10.2501/JAR-53-2-212-220.

Verhoef, Peter C.; Kannan, P. K.; Inman, Jeffrey J. (2015): From Multi-Channel Retailing to Omni-Channel Retailing. In Journal of Retailing 91 (2), pp. 174–181. DOI: 10.1016/j.jretai.2015.02.005.

White, Tom (2015): Hadoop: The Definitive Guide. Storage and Analysis at Internet Scale. 4. ed., rev. & updated. Beijing: O'Reilly.

Whitener, Michael (2015): Cookies Are So Yesterday; Cross-Device Tracking Is In – SomeTips. Available online at https://iapp.org/news/a/cookies-are-so-yesterday-cross-device-tracking-is-insome-tips/, checked on January 11th, 2016.

Xu, Lizhen; Duan, Jason A.; Whinston, Andrew (2014): Path to Purchase. A Mutually Exciting Point Process Model for Online Advertising and Conversion. In Management Science 60 (6), pp. 1392–1412. DOI: 10.1287/mnsc.2014.1952.

Yousafzai, Abdullah; Chang, Victor; Gani, Abdullah; Noor, Rafidah Md (2016): Multimedia augmented m-learning: Issues, trends and open challenges. In Int J Inf Manage 36 (5), pp. 784–792. DOI: 10.1016/j.ijinfomgt.2016.05.010.

Zhang, Ya; Wei, Yi; Ren, Jianbiao (2014): Multi-touch Attribution in Online Advertising with Survival Theory. In: 2014 IEEE International Conference on Data Mining (ICDM). Shenzhen, China, pp. 687–696.

8 Results

In the presented research different results were generated. All relevant results are presented in this chapter in its own sub-chapter.

8.1 Marketing Analytics Process (MAP)

The Marketing Analytics Process (MAP) methodology is developed and successfully utilized in the quantitative analysis of this research. As a specification of the CRISP-DM, the MAP methodology offers a guideline for bigdata projects placed in an (online-) marketing environment.

8.2 What does Efficient Attribution in an Omni-Channel Environment Look Like?

The main research question has been analyzed in the presented research. From a practical point of view, as analyzed in publication two, a major change regarding the requirements towards attribution modelling in an omni-channel environment can be identified. Based on experts' interviews, model requirements and specification and data requirements are identified. The applicability of the identified requirements is successfully implemented in a field experiment.

More available data sources containing granular information about the user and the user's behavior enable an attribution on a user level, as proofed in publication three. This research introduces the first omni-channel ready attribution approach in science which utilizes a cross-device and cross-platform data foundation. The presented approach positively attributes on a user-level by providing two user attributes: the *customer value* and the customer's *conversions probability*. The customer value consists of the current customer value and the conversion probability indicates whether a user is likely to perform another conversion. The user's conversion probability is predicted with an accuracy of 98,4%.

The derived hypotheses (H1 to H4) of this research can be verified.

8.2.1 Verifying Hypotheses H1 and H2

Hypothesis H1 "*New requirements are requested for attribution modelling from a practical point of view in an omni-channel environment*." is verified within publication two, by conducting semi-structured expert interviews and identifying the new requirements and specifications.

The resulting requirements were applied onto identified attribution approaches to analyze their applicability in an omni-channel environment. There is no attribution approach which meets a majority of the identified requirements. Therefore, no existing attribution approach performs attribution efficiently in omni-channel environment. The second hypothesis H2 "*Existing attribution models are not effectively applicable in an omni-channel environment from a practical perspective.*" was verified as well.

By proving that no efficient attribution approach for an omni-channel context exists, a research gap was clearly identified. The research gap indicates the necessity of developing an omni-channel attribution model approach meeting those requirements.

8.2.2 Verifying Hypotheses H3 and H4

The third hypothesis H3 "*It is possible to build a required data foundation and attribution model to work efficiently in an omni-channel environment.*" aimed at filling the research gap by developing such an approach and analyzing the feasibility of the development.

The research gap was filled by presenting the HCJ data foundation and the corresponding attribution approach. Although the provided data used for this research does not contain any soft facts, the data transformation process indicates that such information can be processed, if available. The limitation of the presented research is rooted in the lack of such information in the provided data sources. By presenting the HCJ data foundation and the attribution approach which meets the pre-identified requirements, the third hypothesis H3 has been verified.

Since this research consists of a successfully developed attribution approach, the fourth hypothesis H4 "*If such a model can be developed, savings from at least 10% can be achieved for e.g. a company or an institution.*" could be analyzed. The results of the third publication proofed that significant savings (>10%) and/or a budget shift can be achieved.

Since all hypotheses were analyzed, the main research question can be studied. The main objective of this research is to identify what attribution in an omni-channel environment looks like. Based on the presented research attribution in an omni-channel environment consists of the following characteristics.

An efficient attribution approach in an omni-channel environment should…

- meet the identified model requirements,
- meet the identified data requirements and
- populate the attribution information in real-time back into the user's client for further processing.

9 Conclusion

In the beginning of this chapter, a summary of the investigation is presented. The main research question is discussed, and limitations are indicated. This chapter finishes with the contribution to the scientific community including implications for theory and practitioners.

9.1 Summary of the Investigation

Within the presented research the main research question "*What does efficient attribution in an omni-channel environment look like?"* is examined. By presenting the MAP methodology, a specification of the CRISP-DM is introduced which was successfully applied in the current research to study the main research question. Within a sequential mixed-method inspired approach, the main research question has been examined. By executing a structured literature review and conducting expert interviews, the research gap – the lack of omni-channel attribution approaches – is clearly identified, which is examined and filled by the presented research. An omni-channel ready attribution approach is presented. The analysis of the main research question is guided by four hypotheses which all were verified. The main characteristics of an efficient attribution approach in an omni-channel environment are listed in the results of this research and will be discussed in this chapter.

9.2 What does Efficient Attribution in an Omni-Channel Environment Look Like?

In general, attribution changes fundamentally in an omni-channel environment compared to attribution approaches in a multi-channel setup. A holistic perspective of the user's journey enables attribution on a user's level. Boundaries created by providing different channels were removed due to a central data foundation which holds and/or connects data from different sources. At this point it is to mention that the accessibility of third party raw data is a challenging task and enforces ETL-knowledge. Depending on the vendor and the amount of data, accessing the raw data can be cost-intensive. If the data in its entirety is accessible the challenges and limitations of data silos are removed. By performing attributing on a user level, attribution becomes a complex marketing bigdata problem. As Illustrated in this research, data sources need to be extended by linking information to connect and combine data from multiple marketing channels and sources. In the presented research the targeted HCJ data foundation consists of available information of users in a holistic way.

As conceived in this research, attribution providing a customer value and a conversion probability, harbors potential in using marketing budgets more efficiently. The process of budget splitting is supported by real-time information on a user-level. The information about the user's behavior, derived attitudes and interests can be used to influence the split of the marketing budget. These two user attributes can be populated for attribution purposes which can be utilized to automate marketing decisions in real-time back into the user's client.

9.3 Critical Appraisal, Limitations and Opportunities for Further Research

The focus of attribution modelling has shifted from a channel perspective to a user centric approach. Based on the requirements, an attribution needs to be influenced by the customer's behavior. The need of a budget split across provided channels remains relevant, but the decision whether a customer needs to be addressed through certain channels needs to be made in real-time, based on the user's characteristics.

Shifting towards an omni-channel setup enables optimization potential for a company or institution in terms of using budgets more efficiently based on customer's needs or budget savings. Such a change is having an impact on what marketers need to do for their daily business. Budget allocation will be strongly influenced by the results of future attribution systems, which enable an attribution on a user level. This ensures a dynamic, more realistic, budget split onto the different channels. The budget split is no longer performed mainly on appraisals from marketing experts but is dynamically indicated by the user's behavior.

Generalizability

Within the presented research, a saving potential has been shown based on the data from one real-estate platform located in Germany. The provided approach needs to be adjusted and applied onto data from companies in different industries to proof generalizability.

The structure of the ETL process can be utilized without modification if the same data sources are provided at a different company or institution. The feature engineering process within the qualitative analysis is business model specific and needs to be altered towards the needs of other business models. The ML approach can be applied onto the modified features to identify the hyperparameter configuration. The provided approach can be utilized. The setup of populating the calculated values into the scope of the user's client to perform marketing decisions can remain the same. ML is utilized to provide a problem-specific solution, for the business model specifically used in this research. Of course, the generalizability needs to be analyzed for similar industries in other countries.

Data Transformation / Model Input Data

The presented data transformation process aimed at a high-quality output in terms of data correctness, considering only complete journeys. By aiming at a high-quality output pieces of information, which were not linkable to other data sources, were neglected. In total, almost 86% of the available data has been considered. More than 14% have been neglected. This focus, of course, spawns a very high data quality which is an indication of the good prediction results. For example, a different transformation approach focusing on using all available data could result in a different outcome. Such an approach produces inconsistent journeys including more information. These two – or any other focus – need to be benchmarked against the provided approach to analyze whether the chosen approach is the optimal choice. This aspect opens up a new research area in the field of attribution.

This presented research relies on the correct and complete data provided by different third-party vendors such as Tealium and intelliAd utilized by the data providing company. In an omni-channel environment a critical analysis of utilized data sources becomes more important than in a multi-channel environment. If one data source holds incomplete and/or wrong information about a user, the quality of the whole linking process results in an insufficient output quality. The results of a prediction algorithm can be strongly influenced by incorrect training data. For the presented research, the correctness of the provided data sources has been inspected before the start of the investigation by testing the environment.

Model Output

In the introduction two definitions of attribution modelling were presented. Anderl et al. (2016a) describe the attribution problem as an iterative process optimizing the budget allocation onto provided channels. The presented attribution approach offers two new pieces of information: a *customer value* and a prediction of the *conversion probability*, helping to adjust budget allocation. Of course, budget allocation optimization itself remains as an iterative process. Moreover, as already discussed in the third publication, it is necessary to analyze which information about a user is relevant and to what degree. Based on the identified requirements the provided pieces of information are relevant from a practical point of view. Other possibilities could be information such as *next best product* or *next best channel*. This reveals a new research gap for future research. The presented flexible data transformation setup is extendable to add data which needs to be taken into consideration for new values. Future research needs to analyze the value of attribution modeling of the presented values and the new values.

Algorithm for Conversion Probability

Within the presented research only tree-based algorithms are applied to solve the classification problem of the conversion probability. According to the current state of research, different well-established algorithms were chosen to solve the classification problem. Since machine learning is a much-researched area, new algorithms or further development of existing algorithms will be available in the future. Further development of algorithms needs to be analyzed regularly. For example, new boosting approaches can be relevant for the analyzed context.

Applying an ML approach requires the consideration of ethical concerns: understandable machine learning/ artificial intelligence ethics. The presented predication of the conversion probability is not 100% understandable since the transformation, by applying a PCA and a Random Forrest approach, is not completely transparent. Future research in this field needs to enable a transparency and an understanding of why the result manifested in this way.

9.4 Contribution to Knowledge

Consistent with the results of other research in the general field of omni-channel marketing, such as Anderl et al. (2016a) and in retailing Verhoef et al. (2015), an omni-channel environment enables new opportunities due to the presence of granular data. The aforementioned research analyzed among other aspects how the problem of attribution modelling evolves in an omni-channel environment. The here provided research consists of the first omni-channel ready attribution approach using a cross-device and cross-platform data foundation. The contributed model extends the list of existing attribution approaches provided by different authors such as Anderl et al. (2016a), Abhishek et al. (2012) and Li and Kannan (2014) by providing an omni-channel ready attribution approach.

The provided data transformation process and the corresponding attribution approach are the first omni-channel attribution approach. The main contribution to the scientific community is the definition of how attribution looks like in an omni-channel environment. Furthermore, the following contributions are made:

1. A specification of the Cross-Industry Standard for Data Mining (CRISP-DM) process for (online-) marketing specific bigdata problems, the MAP methodology.
2. Identification of existing dynamic attribution models in the science community based on a structured literature research process.
3. Requirements and specifications for dynamic attribution models in an omni-channel environment based on expert interviews.
4. Evaluation of the identified models based on the requirements and specifications from the expert interviews.
5. An omni-channel ready attribution approach built onto a cross-device and cross-platform data foundation.
6. Proof of practicability of the implantation of the pre-identified requirements and specification for efficient attribution in an omni-channel environment.
7. Definition of exigencies, research fields and research questions for further research.

9.4.1 Implications for Theory

The presented research opens up different research areas and research fields. At the end of the second publication, a list of research fields is presented. The main research questions and research fields deriving are listed below. Already mentioned research fields and research questions are not re-stated.

One important question for future research should deal with the provided output of the attribution model and answer the question: *What user attributes have what impact for efficient attribution in an omni-channel environment?* Moreover, the setup of how to return the calculated information to the decision engine needs to be analyzed. For the presented research a direct enrichment of the client’s user-profile is pursued. Another option is a non-client integration. This setup needs to be analyzed scientifically, implemented and tested by practitioners to return learnings and insights back to science.

To sum up, requirements for efficient attribution modelling have increased and change the attribution in an omni-channel environment significantly. Requirements, such as considering dynamically calculated information about the customer in real-time, need to be present and considered by the attribution model. New opportunities arise by providing the relevant information for the attribution decision. The presented approach can be utilized by future research to identify such relevant pieces of information.

9.4.2 Implications for Practitioners

Establishing a valid data basis meeting the presented data requirements is a challenging task and requires management support. Implementing an omni-channel attribution approach enforces a data driven culture within the company led by the management. Such a data driven culture enables the exchange of information between different departments. Such a change will raise other advantages as well. Due to a better understanding of the customers for product development (Lilien et al. 2002), (online-) marketing (Blattberg and Deighton 1991), sales and customer support, the user can be treated more advantageously.

As described in publication two, besides data requirements and model requirements three aspects are placed in the category *other criteria*. Marketing experts clearly express a change by postulating different skills for marketing experts. Next to fundamental skills from the business intelligence (BI) sector and an understanding of technical aspects in this field, an understanding of the raw data and the data sources is required to identify new potentials. These postulated skills underline the influences within a marketing department spawned by a shift towards an omni-channel setup.

The provided MAP methodology is successfully utilized in this presented research. Other projects in practice or in theory need to utilize the MAP to proof its added value.

In conclusion, in an omni-channel environment a significant change towards attribution in a multi-channel environment is detected. An efficient attribution approach in an omni-channel environment should meet the identified model requirements, meet the identified data requirements and populate the attribution information in real-time into the user’s client for further processing. Existing multi-channel attribution approaches are rather an evaluation of companies, without proof. Future omni-channel attribution models need to focus on the value of the customer.

10 Publication bibliography

Abhishek, Vibhanshu; Fader, Peter; Hosanagar, Kartik (2012): Media Exposure Through the Funnel. A Model of Multi-Channel Attribution. Available online at http://dx.doi.org/10.2139/ssrn.2158421.

Ackermann, Sebastian; Wangenheim, Florian von (2014): Behavioral Consequences of Customer-Initiated Channel Migration. In *Journal of Service Research* 17 (3), pp. 262–277. DOI: 10.1177/1094670513519862.

Alon, Noga; Gamzu, Iftah; Tennenholtz, Moshe (2012): Optimizing Budget Allocation Among Channels and Influencers. In Alain Mille, Fabien Gandon, Jacques Misselis, Michael Rabinovich, Steffen Staab (Eds.): Proceedings of the 21st international conference on World Wide Web. the 21st international conference. Lyon, France. 21st World Wide Web Conference 2012; ACM Special Interest Group on Hypertext, Hypermedia, and Web. New York, NY: ACM, p. 381.

Amazon Inc. (2006): Amazon S3. Amazon Inc. Available online at https://aws.amazon.com/s3/?nc1=h_ls, checked on August 20th, 2018.

Amelang, Manfred; Zielinski, Werner; Fydrich, Thomas (2004): Psychologische Diagnostik und Intervention. 3rd ed. Berlin: Springer (Springer-Lehrbuch).

Anderl, Eva; Becker, Ingo; Wangenheim, Florian von; Schumann, Jan Hendrik (2016a): Mapping the Customer Journey. Lessons Learned from Graph-Based Online Attribution Modeling. In *International Journal of Research in Marketing* 33 (3), pp. 457–474. DOI: 10.1016/j.ijresmar.2016.03.001.

Anderl, Eva; Schumann, Jan Hendrik; Kunz, Werner (2016b): Helping Firms Reduce Complexity in Multichannel Online Data. A New Taxonomy-Based Approach for Customer Journeys. In *Journal of Retailing* 92 (2), pp. 185–203. DOI: 10.1016/j.jretai.2015.10.001.

Apache (2017): Apache Porject: Hue. Apache. Available online at http://gethue.com/, checked on August 2nd, 2018.

Archak, Nikolay; Mirrokni, Vahab S.; Muthukrishnan, Si. (2012): Budget Optimization for Online Campaigns with Positive Carryover Effects. In Paul W. Goldberg (Ed.): Internet and network economics. 8th international workshop, WINE 2012, Liverpool, UK, December 10 - 12, 2012 ; proceedings, vol. 7695. Berlin: Springer (Lecture Notes in Computer Science, 7695), pp. 86–99.

Auto-Sklearn Development Team (2018): Auto-Sklearn. Available online at https://github.com/automl/auto-sklearn, updated on June 18th, 2018, checked on August 1st, 2018.

Bacharach, Samuel B. (1989): Organizational Theories. Some Criteria for Evaluation. In *The Academy of Management Review* 14 (4), p. 496. DOI: 10.2307/258555.

Bell, David R.; Gallino, Santiago; Moreno, Antonio (2014): How to Win in an Omnichannel World. Fall 2014, Vol. 56 No.1. MITSloan (MITSloan Management Review, 1).

Bello-Orgaz, Gema; Jung, Jason J.; Camacho, David (2016): Social Big Data: Recent Achievements and New Challenges. In *Inf. Fusion* 28 (1), pp. 45–59. DOI: 10.1016/j.inffus.2015.08.005.

Berman, Ron (2015): Beyond the Last Touch. Attribution in Online Advertising. Available online at http://ron-berman.com/papers/attribution.pdf, checked on 9/28/2016.

Bishop, Christopher M. (2006): Pattern Recognition and Machine Learning. New York, NY: Springer Science+Business Media LLC (Information Science and Statistics). Available online at http://dx.doi.org/10.1007/978-0-387-45528-0.

Blattberg, Robert C.; Deighton, John (1991): Interactive Marketing: Exploiting the Age of Adressability. In *MIT Sloan Management Review* 33 (1), pp. 5–14.

Bogner, Alexander (Ed.) (2005): Das Experteninterview. Theorie, Methode, Anwendung. 2nd ed. Wiesbaden: VS Verl. für Sozialwiss. Available online at http://www.socialnet.de/rezensionen/isbn.php?isbn=978-3-531-14447-4.

Bogner, Alexander; Littig, Beate; Menz, Wolfgang (2014): Interviews mit Experten. Wiesbaden: Springer Fachmedien Wiesbaden.

Borden, Neil H. (1964): The Concept of the Marketing Mix. In *Marketing management and administrative action*, pp. 31–40.

Bose, Indranil; Mahapatra, Radha K. (2001): Business Data Mining — A Machine Learning Perspective. In *Information & Management* 39 (3), pp. 211–225. DOI: 10.1016/S0378-7206(01)00091-X.

Boulesteix, Anne-Laure (2015): Ten Simple Rules for Reducing Overoptimistic Reporting in Methodological Computational Research. In *PLoS computational biology* 11 (4), e1004191. DOI: 10.1371/journal.pcbi.1004191.

Breiman, Leo; Friedman, Jerome; Stone, Charles J.; Olshen, Richard A. (1984): Classification and Regression Trees. 1st ed. Boca Raton: CRC Press. Available online at https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=5109080.

Camiade, Benoït A. J.-M. (2013): Multi-Channel, Cross-Channel, Omni-Channel Retailing: Business in All Its Forms (1/2). ATInternet.com. Available online at https://blog.atinternet.com/en/series-multi-channel-cross-channel-omni-channel-retailing-business-forms-12/, checked on 12/12/2017.

Cangelosi, Richard; Goriely, Alain (2007): Component Retention in Principal Component Analysis with Application to cDNA Microarray Data. In *Biology direct* 2 (2). DOI: 10.1186/1745-6150-2-2.

Capriolo, Edward; Nash, Courtney; Wampler, Dean; Rutherglen, Jason; Loukides, Michael Kosta; Demarest, Rebecca (Eds.) (2012): Programming Hive. 1st ed. Sebastopol, CA: O'Reilly & Associates. Available online at http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10759003.

Carroll, Donald; Guzmán, Inés (2013): The New Omni-Channel Approach to Serving Customers. Strategy Implications for Communications Service Providers. Available online at https://www.accenture.com/be-en/~/media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Industries_2/accenture-new-omni-channel-approach-serving-customers.pdf, checked on 8/2/2018.

Chapelle, Olivier; Schölkopf, Bernhard; Zien, Alexander (Eds.) (2006): Semi-Supervised Learning. Cambridge, Mass.: MIT Press (Adaptive computation and machine learning).

Chen, Hsinchun (1995): Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms. In *J. Am. Soc. Inf. Sci.* 46 (3), pp. 194–216. Available online at https://doi.org/10.1002/(SICI)1097-4571(199504)46:3<194::AID-ASI4>3.0.CO;2-S.

Chen, Yanpei; Alspaugh, Sara; Katz, Randy (2012): Interactive Analytical Processing in Big Data Systems. A Cross-Industry Study of MapReduce Workloads. In *Proc. VLDB Endow.* 5 (12), pp. 1802–1813. DOI: 10.14778/2367502.2367519.

Chicco, Davide (2017): Ten Quick Tips for Machine Learning in Computational Biology. In *BioData mining* 10, p. 35. DOI: 10.1186/s13040-017-0155-3.

Claesen, Marc; Moor, Bart De (2015): Hyperparameter Search in Machine Learning. Available online at http://arxiv.org/pdf/1502.02127v2, checked on August 3rd, 2018.

Cook, Glenn (2014): Customer Experience in the Omni-Channel World and the Challenges and Opportunities This Presents. In *Journal of Direct, Data and Digital Marketing Practice* 15 (4), pp. 262–266. DOI: 10.1057/dddmp.2014.16.

Creswell, John W. (2014): Research Design. Qualitative, Quantitative, and Mixed Methods Approaches. 4. ed., internat. student ed. Los Angeles Calif. u.a.: Sage.

Creswell, John W.; Clark, Vickie L. P.; Gutmann, Michelle; Hanson, William E. (2003): Advanced Mixed Methods Research Designs. In *Handbook of mixed methods in social and behavioral research*, pp. 209–240.

Cronin, Patricia; Ryan, Frances; Coughlan, Michael (2008): Undertaking a Literature Review. A Step-by-Step Approach. In *British journal of nursing* 17 (1), pp. 38–43.

Cui, Geng; Wong, Man Leung; Lui, Hon-Kwong (2006): Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. In *Management Science* 52 (4), pp. 597–612. DOI: 10.1287/mnsc.1060.0514.

Dalessandro, Brian; Perlich, Claudia; Stitelman, Ori; Provost, Foster (2012): Causally Motivated Attribution for Online Advertising. In *ADKDD '12 Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, pp. 1–9. DOI: 10.1145/2351356.2351363.

Dasu, Tamraparni; Johnson, Theodore (2003): Exploratory Data Mining and Data Cleaning. Hoboken, NJ: Wiley-Interscience (Wiley series in probability and statistics). Available online at http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10299281.

Dean, Jeffrey; Ghemawat, Sanjay (2004): MapReduce: Simplified Data Processing on Large Clusters, pp. 137–150. Available online at https://storage.googleapis.com/pub-tools-public-publication-data/pdf/16cb30b4b92fd4989b8619a61752a2387c6dd474.pdf, checked on 8/17/2018.

Diaz-Morales, Roberto (2015): Cross-Device Tracking. Matching Devices and Cookies. In Peng Cui (Ed.): 15th IEEE International Conference on Data Mining Workshop (ICDMW). Atlantic City, NJ, USA. Institute of Electrical and Electronics Engineers; IEEE. Piscataway, NJ: IEEE, pp. 1699–1704.

Diekmann, Andreas (2007): Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen. 17. Ed. Reinbek: Rororo Rowohlt-Taschenbuch-Verl. (Rowohlts Enzyklopädie, 55551).

Diener, Ed; Crandall, William R. (1978): Ethics in Social and Behavioral Research. Chicago: Univ. of Chicago Press.

Dinner, Isaac M.; van Heerde, Harald J.; Neslin, Scott (2011): Driving Online and Offline Sales. The Cross-Channel Effects of Digital Versus Traditional Advertising. In *SSRN Journal*. DOI: 10.2139/ssrn.1955653.

Domingos, Pedro (2012): A Few Useful Things to Know About Machine Learning. In *Commun. ACM* 55 (10), p. 78. DOI: 10.1145/2347736.2347755.

ECMA-262 Standard, December 1999: ECMA-262 ECMAScript Language Specification.

ECMA-404 Standard, December 2017: ECMA-404 The Json Data Interchange Syntax.

Econsultancy (2015): Quarterly Digital Intelligence Briefing. Digital Trends 2015. Available online at https://www.marketingsociety.com/sites/default/files/QDIB%20Adobe%20Digital%20Trends%20Report%202015_EMEA_0.pdf, checked on December 12th, 2017.

eMarketer (2016): Primary Attribution Model Used by Their Marketing Team to Measure Performance According to US B2B Marketers. Available online at emarketer.com, checked on January 20th, 2016.

Fan, Shaokun; Lau, Raymond Y. K.; Zhao, Leon J. (2015): Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix. In *Big Data Research* 2 (1), pp. 28–32. DOI: 10.1016/j.bdr.2015.02.006.

Feature Tools Development Group (2018): Feature Tools 0.3.0. Available online at https://docs.featuretools.com, checked on August 31st, 2018.

Feit, Eleanor Mcdonnell; Wang, Pengyuan; Bradlow, Eric T.; Fader, Peter S. (2013): Fusing Aggregate and Disaggregate Data with an Application to Multiplatform Media Consumption. In *Journal of Marketing Research* 50 (3), pp. 348–364. DOI: 10.1509/jmr.11.0431.

Felden, Carsten (2012): Datenqualitätsmanagement. Enzyklopädie der Wirtschaftsinformatik. Edited by Norbert Gronau, Jörg Becker, Karl Kurbel, Elmar Sinz, Leena Suhl. Available online at http://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/daten-wissen/Datenmanagement/Datenmanagement--Konzepte-des/Datenqualitatsmanagement, updated on 10/31/2012, checked on April 5th, 2016.

Flick, Uwe (2007): Qualitative Sozialforschung. Eine Einführung. 1st ed., revised and extened. Reinbek bei Hamburg: Rowohlt Taschenbuch Verl. (Rororo, 55694).

Früh, Werner (2015): Inhaltsanalyse. Theorie und Praxis. 8., überarbeitete Auflage. Konstanz, München: UVK Verlagsgesellschaft mbH; UVK / Lucius (UTB, 2501).

Gallino, Santiago; Moreno, Antonio (2014): Integration of Online and Offline Channels in Retail. The Impact of Sharing Reliable Inventory Availability Information. In *Management Science* 60 (6), pp. 1434–1451. DOI: 10.1287/mnsc.2014.1951.

Geyik, Sahin Cem; Saxena, Abhishek; Dasdan, Ali (2014): Multi-Touch Attribution Based Budget Allocation in Online Advertising ADKDD’14, pp. 1–9. DOI: 10.1145/2648584.2648586.

Gläser, Jochen; Laudel, Grit (2010): Experteninterviews und qualitative Inhaltsanalyse. Als Instrumente rekonstruierender Untersuchungen. 4st ed. Wiesbaden: VS Verl. f. Sozialwiss (Lehrbuch).

Goi, Chai L. (2009): A Review of Marketing Mix. 4Ps or More? In *IJMS* 1 (1). DOI: 10.5539/ijms.v1n1p2.

Google Inc. (2017): Attribution Modeling Overview. Assign Credit for Sales and Conversions to Touchpoints in Conversion Paths. Available online at https://support.google.com/analytics/answer/1662518?hl=en, checked on January 21st, 2017.

Google Inc. (2018a): BigQuery Export Schema. Edited by Google Inc. Available online at https://support.google.com/analytics/answer/3437719?hl=en, checked on July, 28th 2018.

Google Inc. (2018b): Google Analytics. Available online at https://analytics.google.com/analytics/web/, checked on August 10th, 2018.

Google Inc. (2018c): Google BigQuery. Edited by Google Inc. Available online at https://cloud.google.com/bigquery/, checked on July, 28th 2018.

Grady, Jeffrey O. (1995): System Engineering Planning and Enterprise Identity. 1st ed.: CRC Press.

Granville, Vincent (2015): Reducing Data Cleansing Time to Get Actionable Insights Faster. Available online at https://www.datasciencecentral.com/profiles/blogs/reducing-data-cleansing-time-to-get-actionable-insights-faster, updated on June 15th, 2015, checked on June 2nd, 2016.

Greenhalgh, Trisha; Peacock, Richard (2005): Effectiveness and Efficiency of Search Methods in Systematic Reviews of Complex Evidence: Audit of Primary Sources. In *BMJ (Clinical research ed.)* 331 (7524), pp. 1064–1065. DOI: 10.1136/bmj.38636.593461.68.

Grewal, Dhruv; Bart, Yakov; Spann, Martin; Zubcsek, Peter Pal (2016): Mobile Advertising. A Framework and Research Agenda. In *Journal of Interactive Marketing* 34, pp. 3–14. DOI: 10.1016/j.intmar.2016.03.003.

Guest, Greg; MacQueen, Kathleen M.; Namey, Emily E. (2012): Applied Thematic Analysis. Thousand Oaks CA u.a.: Sage.

Haan, Evert de; Wiesel, Thorsten; Pauwels, Koen (2016): The Effectiveness of Different Forms of Online Advertising for Purchase Conversion in a Multiple-Channel Attribution Framework. In *International Journal of Research in Marketing* 33, pp. 491–507.

Hammerschmidt, Maik; Falk, Tomas; Weijters, Bert (2015): Channels in the Mirror. An Alignable Model for Assessing Customer Satisfaction in Concurrent Channel Systems. In *Journal of Service Research* 19 (1), pp. 88–101. DOI: 10.1177/1094670515589084.

Han, Jiawei; Kamber, Micheline; Pei, Jian (2012): Data mining. Concepts and techniques. 3. ed. Amsterdam: Elsevier/Morgan Kaufmann (The Morgan Kaufmann series in data management systems). Available online at https://ebookcentral.proquest.com/lib/subhh/detail.action?docID=729031.

Haumer, Florian (2015): Was Marketingcontrolling und Big Data Analytics gemeinsam haben. Available online at http://www.sputnika.de/dresden/magazin/details/article/was-marketingcontrolling-und-big-data-analytics-gemeinsam-haben/, updated on Mai 5th, 2015, checked on June 6th, 2016.

Hopf, Christel; Schmidt, Christiane (pub.) (1993): Zum Verhältnis von innerfamilialen sozialen Erfahrungen, Persönlichkeitsentwicklung und politischer Orientierungen. Dokumentation und Erörterung des methodischen Vorgehens in einer Studie zu diesem Thema. Unpublished Manuscipt.

Hsu, Chih-Wei; Chang, Chih-Chung; Lin, Chih-Jen (2003): A Practical Guide to Support Vector Classification. Available online at https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, updated on May 19th, 2016, checked on August 3rd, 2018.

Hutter, Frank; Hoos, Holger; Leyton-Brown, Kevin (2014): An Efficient Approach for Assessing Hyperparameter Importance. In Eric P. Xing, Tony Jebara (Eds.): Proceedings of the 31st International Conference on Machine Learning, vol. 32. Bejing, China: PMLR (Proceedings of Machine Learning Research), pp. 754–762. Available online at http://proceedings.mlr.press/v32/hutter14.html.

IDC (2010): IDC Retail Insights. Multichennal Report 2010. Available online at http://info.hybris.com/rs/hybris/images/IDC-Multichannel-EN.pdf, checked on April 4th, 2015.

Inmon, William H. (2005): Building the Data Warehouse. 4th Ed. Indianapolis, IN: Wiley Pub. Available online at http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=140444.

intelliAd Media GmbH (2018): intelliAd. Available online at https://www.intelliad.de/impressum/, checked on August 10th, 2018.

Islam, Mohammad Kamrul; Srinivasan, Aravind (2015): Apache Oozie. [the workflow scheduler for hadoop]. Beijing: O´Reilly.

Jacobs, Adam (2009): The Pathologies of Big Data. In *Commun. ACM* 52 (8), p. 36. DOI: 10.1145/1536616.1536632.

James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2013): An Introduction to Statistical Learning. With Applications in R. New York, NY: Springer (Springer Texts in Statistics, 103). Available online at http://dx.doi.org/10.1007/978-1-4614-7138-7.

Jayawardane, Himani W.; Halgamuge, Sage. K.; Kayande, Uka (2015): Attributing Conversion Credit in an Online Environment: An Analysis and Classification. In: 2015 3rd International Symposium on Computational and Business Intelligence (ISCBI). Bali, Indonesia, pp. 68–73.

Jolliffe, Ian T. (2002): Principal Component Analysis. 2. ed. New York, NY: Springer (Springer series in statistics). Available online at http://www.loc.gov/catdir/enhancements/fy0817/2002019560-d.html.

Joo, Mingyu; Wilbur, Kenneth C.; Cowgill, Bo; Zhu, Yi (2014): Television Advertising and Online Search. In *Management Science* 60 (1), pp. 56–73. DOI: 10.1287/mnsc.2013.1741.

Jordan, Claus; Schnider, Dani; Wehner, Joachim; Welker, Peter (2011): Data Warehousing mit Oracle. Business Intelligence in der Praxis. Munich: Hanser.

Jordan, Michael I.; Mitchell, Tom M. (2015): Machine Learning: Trends, Perspectives, and Prospects. In *Science (New York, N.Y.)* 349 (6245), pp. 255–260. DOI: 10.1126/science.aaa8415.

Json.org (2018): Introducing JSON. Available online at https://www.json.org/index.html, checked on July, 28th 2018.

Judah, Saul; Selvage, Mei; Jain, Ankush (2017): Critical Capabilities for Data Quality Tools. In *Gartner Report*. Available online at https://www.gartner.com/doc/3835263/critical-capabilities-data-quality-tools, checked on August 11th, 2018.

Judd, Vaughan C. (1987): Differentiate with the 5th P: People. In *Industrial Marketing Management* 16 (4), pp. 241–247. DOI: 10.1016/0019-8501(87)90032-0.

Kannan, P. K.; Reinartz, Werner; Verhoef, Peter C. (2016): The Path to Purchase and Attribution Modeling. Introduction to special section. In *International Journal of Research in Marketing* 33 (3), pp. 449–456. DOI: 10.1016/j.ijresmar.2016.07.001.

Kimball, Ralph; Merz, Richard (2000): The Data Webhouse Toolkit. Building the Web-Enabled Data Warehouse. New York: Wiley. Available online at http://www.loc.gov/catdir/bios/wiley042/99055652.html.

Kohavi, Ron (1995): A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *International Joint Conference on Articial Intelligence* 14 (1).

Köhler, Richard (Ed.) (1977): Empirische und handlungstheoretische Forschungskonzeptionen in der Betriebswirtschaftslehre. Bericht über d. Tagung in Aachen, März 1976. Verband der Hochschullehrer für Betriebswirtschaft. Stuttgart: Poeschel.

Kotsiantis, Sotiris (2007): Supervised Machine Learning: A Review of Classification Techniques. In *Informatica (Ljubljana)* 31.

Kotthoff, Lars; Thornton, Chris; Hoos, Holger H.; Hutter, Frank; Leyton-Brown, Kevin (2017): Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. Journal of Machine Learning Research 18 (25), pp. 1–5. Available online at http://jmlr.org/papers/v18/16-261.html.

Krauth, Joachim (1995): Testkonstruktion und Testtheorie. Weinheim: Beltz Psychologie Verl.-Union.

Kuckartz, Udo (2014): Mixed Methods. Methodologie, Forschungsdesigns und Analyseverfahren. Wiesbaden: Springer Fachmedien Wiesbaden. Available online at http://dx.doi.org/10.1007/978-3-531-93267-5.

Lazaris, Christos; Vrechopoulos, Adam (2014): From Multichannel to "Omnichannel" Retailing: Review of the Literature and Calls for Research. In *2nd International Conference on Contemporary Marketing Issues, (ICCMI), 18-20 June 2014, At Athens, Greece*. DOI: 10.13140/2.1.1802.4967.

Lee, Garry (2010): Death of 'last click wins'. Media Attribution and the Expanding Use of Media Data. In *J Direct Data Digit Mark Pract* 12 (1), pp. 16–26. DOI: 10.1057/dddmp.2010.14.

Leone, Robert P. (1995): Generalizing What Is Known About Temporal Aggregation and Advertising Carryover. In *Marketing Science* 14 (3), 141-150. DOI: 10.1287/mksc.14.3.G141.

Levy, Michael; Weitz, Barton A.; Grewal, Dhruv (2014): Retailing Management. 9th ed. New York, NY: McGraw-Hill Education.

Lewis, Randall A.; Rao, Justin M. (2013): On the Near Impossibility of Measuring the Returns to Advertising. Available online at http://justinmrao.com/lewis_rao_nearimpossibility.pdf, checked on November 21st, 2015.

Li, Hongshuang; Kannan, P. K. (2014): Attributing Conversions in a Multichannel Online Marketing Environment. An Empirical Model and a Field Experiment. In *Journal of Marketing Research* 51 (1), pp. 40–56. DOI: 10.1509/jmr.13.0050.

Liao, Shu-Hsien; Chu, Pei-Hui; Hsiao, Pei-Yuan (2012): Data Mining Techniques and Applications – A Decade Review from 2000 to 2011. In *Expert Systems with Applications* 39 (12), pp. 11303–11311. DOI: 10.1016/j.eswa.2012.02.063.

Lilien, Gary L.; Morrison, Pamela D.; Searls, Kathleen; Sonnack, Mary; Hippel, Eric von (2002): Performance Assessment of the Lead User Idea-Generation Process for New Product Development. In *Management Science* 48 (8), pp. 1042–1059. DOI: 10.1287/mnsc.48.8.1042.171.

Liu, Ling (Ed.) (2009): Encyclopedia of Database Systems. New York, NY: Springer (Springer reference). Available online at http://deposit.dnb.de/cgi-bin/dokserv?id=2820703&prov=M&dok_var=1&dok_ext=htm.

Lutz, Mark (2017): Learning Python. 5th ed. Beijing, Boston, Farnham, Sebastopol, Tokyo: O'Reilly.

Malgara, Andrea (2014): Big Data und Attribution Modelling: Total Marketing Controlling in Echtzeit (New Business, 44/2014). Available online at http://www.mediaplus.com/de/presse-detail/big-data-und-attribution-modelling-total-marketing-controlling-in-echtzeit.html, checked on March 24th, 2016.

Mayring, Philipp (2001): Kombination und Integration qualitativer und quantitativer Analyse. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, Vol 2, No 1 (2001):

Qualitative and Quantitative Research: Conjunctions and Divergences. DOI: 10.17169/FQS-2.1.967.

Mayring, Philipp (2010): Qualitative Inhaltsanalyse. Grundlagen und Techniken. 11th ed. Weinheim u.a.: Beltz (Pädagogik).

Mayring, Philipp (2016): Einführung in die qualitative Sozialforschung. Eine Anleitung zu qualitativem Denken. 6th ed. (Pädagogik).

McCarthy, Jerome E. (1964): Basic Marketing. Homewood, IL, USA: Irwin.

Menkov, Vladimir; Dayanik, Aynur; Lewis, David D.; Madigan, David; Genkin, Alexander (2006): Constructing Informative Prior Distributions from Domain Knowledge in Text Classification, pp. 493–500. DOI: 10.1145/1148170.1148255.

Meyer, Tony A.; Whateley Brendon (2004): SpamBayes: Effective Open-Source, Bayesian Based, Email Classification System., pp. 493–500.

Mlodinow, Leonard (2008): The Drunkard's Walk. How Randomness Rules Our Lives: Knopf Doubleday Publishing Group.

Moffett, Tina; Pilecki, Mary; McAdams, Rebecca (2014): The Forrester Wave: Cross-Channel Attribution Providers, Q4 2014. Available online at https://www.forrester.com/report/The+Forrester+Wave+CrossChannel+Attribution+Providers+Q4+2014/-/E-RES115221, checked on November 5th, 2017.

Mohri, Mehryar; Rostamizadeh, Afshin; Talwalkar, Ameet (2012): Foundations of Machine Learning. Cambridge, Mass., London: The MIT Press (Adaptive computation and machine learning).

Morse, Janice M. (1991): Approaches to Qualitative-Quantitative Methodological Triangulation. In *Nursing Research* 40 (2), pp. 120–123.

MSI, Marketing Science Institute (2016): Research Priorities 2O16–2O18. MSI Marketing Science institute. Available online at https://www.msi.org/uploads/articles/MSI_RP16-18.pdf, checked on November 3rd, 2017.

Müller, Andreas C.; Guido, Sarah (2017): Introduction to Machine Learning with Python. A Guide for Data Scientists. First edition. Sebastopol, CA: O'Reilly Media. Available online at http://proquest.tech.safaribooksonline.de/9781449369880.

Naik, Prasad A.; Peters, Kay (2009): A Hierarchical Marketing Communications Model of Online and Offline Media Synergies. In *Journal of Interactive Marketing* 23 (4), pp. 288–299. DOI: 10.1016/j.intmar.2009.07.005.

Nass, Ole; Garrigós, José A.; Gómez, Hermengildo G.; Schoeneberg, Klaus-Peter (2018): Attribution modelling in an omni-channel environment. new requirements and specifications from a practical perspective. In *Int. J. Electronic Marketing and Retailing*.

Neslin, Scott A.; Grewal, Dhruv; Leghorn, Robert; Shankar, Venkatesh; Teerling, Marije L.; Thomas, Jacquelyn S.; Verhoef, Peter C. (2006): Challenges and Opportunities in Multichannel Customer Management. In *Journal of Service Research* 9 (2), pp. 95–112. DOI: 10.1177/1094670506293559.

Ngai, Eric W.T.; Xiu, Li; Chau, Dorothy C. K. (2009): Application of Data Mining Techniques in Customer Relationship Management. A Literature Review and Classification. In *Expert Systems with Applications* 36 (2), pp. 2592–2602. DOI: 10.1016/j.eswa.2008.02.021.

Nottorf, Florian (2014): Multi-Channel Attribution Modeling on User Journeys. In: E-Business and Telecommunications, vol. 456, pp. 107–125.

Nottorf, Florian; Funk, Burkhardt (2013): A Cross-Industry Analysis of the Spillover Effect in Paid Search Advertising. In *Electron Markets* 23 (3), pp. 205–216. DOI: 10.1007/s12525-012-0117-z.

Olbrich, Rainer; Schultz, Carsten D. (2014): Multichannel Advertising. Does Print Advertising Affect Search Engine Advertising? In *European Journal of Marketing* 48 (9/10), pp. 1731–1756. DOI: 10.1108/EJM-10-2012-0569.

Olson, Randal S.; Sipper, Moshe; La Cava, William; Tartarone, Sharon; Vitale, Steven; Fu, Weixuan et al. (2017): A System for Accessible Artificial Intelligence. Available online at http://arxiv.org/pdf/1705.00594v2.

Olson, Randal S.; Urbanowicz, Ryan J.; Andrews, Peter C.; Lavender, Nicole A.; La Kidd, Creis; Moore, Jason H. (2016): Automating biomedical data science through tree-based pipeline optimization. Available online at http://arxiv.org/pdf/1601.07925v1.

Paccard, Erwan (2017): Omnichannel vs Multichannel: Are they so different? Available online at http://multichannelmerchant.com/blog/omnichannel-vs-multichannel-different/, checked on June 11th, 2017.

Petersen, Andrew J.; McAlister, Leigh; Reibstein, David J.; Winer, Russell S.; Kumar, V.; Atkinson, Geoff (2009): Choosing the Right Metrics to Maximize Profitability and Shareholder Value. In *Journal of Retailing* 85 (1), pp. 95–111. DOI: 10.1016/j.jretai.2008.11.004.

Piatetsky, Gregory (2014): CRISP-DM, Still the Top Methodology for Analytics, Data Mining, or Data Science Projects. KDnuggets. Available online at https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html, checked on August 3rd, 2018.

Polo, Yolanda; Sese, Javier F. (2016): Does the Nature of the Interaction Matter? Understanding Customer Channel Choice for Purchases and Communications. In *Journal of Service Research* 19 (3), pp. 276–290. DOI: 10.1177/1094670516645189.

Richert, Willi; Coelho, Luis Pedro (2013): Building Machine Learning Systems with Python. Master the Art of Machine Learning with Python and Build Effective Machine Learning

Systems with this Intensive Hands-on Guide. Birmingham: Packt Publ (Community experience distilled).

Rijn, Jan N. van.; Hutter, Frank (2018): Hyperparameter Importance Across Datasets. In Yike Guo, Faisal Farooq (Eds.): Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18. the 24th ACM SIGKDD International Conference. London, United Kingdom, 8/19/2018 - 8/23/2018. New York, New York, USA: ACM Press, pp. 2367–2376.

Runkler, Thomas A. (2015): Data Mining. Modelle und Algorithmen intelligenter Datenanalyse. 2nd ed. Wiesbaden: Springer Vieweg (Lehrbuch).

Rusko, Rauno (2015): Conflicts of Supply Chains in Multi-Channel Marketing. A Case From Northern Finland. In *Technology Analysis & Strategic Management* 28 (4), pp. 477–491. DOI: 10.1080/09537325.2015.1100294.

Russell, Stuart J.; Norvig, Peter (2016): Artificial Intelligence. A Modern Approach. With assistance of Ernest Davis, Douglas Edwards. 3rd ed. Boston, Columbus, Indianapolis, New York, San Francisco: Pearson (Always learning).

Rutz, Oliver J.; Bucklin, Randolph E. (2011): From Generic to Branded. A Model of Spillover in Paid Search Advertising. In *Journal of Marketing Research* 48 (1), pp. 87–102. DOI: 10.1509/jmkr.48.1.87.

Ryan, Mike (2018): AWS System Administration: Best Practices for Sysadmins in the Amazon Cloud: O'Reilly Media, Incorporated.

Ryte (2016): Attribution Modelling. With assistance of Ryte Wiki.com. Available online at https://en.ryte.com/wiki/Attribution_Modelling, checked on August 21st, 2018.

Saldaña, Johnny (2009): The Coding Manual for Qualitative Researchers: Sage Pubn Inc.

Saldanha, Alexander; Berman, Ron; Vummarao Keshore (2013): Advertising Conversion Attribution. Applied for by Abakus, Inc., Emeryville, CA (US) on 3/14/2013. App. no. 13/830,618. Patent no. US 8,775,248 B1.

Salipante, Paul; Notz, William; Bigelow, John (1982): A Matrix Approach to Literature Reviews. In *Research in Organizational Behavior* 4, pp. 321–348.

Sasaki, Yutaka (2007): The Truth of the F-Measure. With assistance of School of Computer Science, University of Manchester. Available online at https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf, checked on August, 1st 2018.

Schoeneberg, Klaus-Peter; Nass, Ole; Schmitt, Lennart (2017): Marketing-Analytics-Process (MAP). Data-Driven-Marketing-Projekte erfolgriech durchführen. In Christopher Zerres (Ed.): Handbuch Marketing-Controlling. Grundlagen - Methoden - Umsetzung, 4th Ed. 4., vollständig überarbeitete Auflage. Berlin, Germany: Springer Gabler, pp. 15–39.

Schoeneberg, Klaus-Peter; Pein, Jennifer (2014): Entscheidungsfindung mit Big Data. Einsatz fortschrittlicher Visualisierungsmöglichkeiten zur Komplexitätsbeherrschung betriebswirtschaftlicher Sachverhalte im Unternehmen. In Klaus-Peter Schoeneberg (Ed.): Komplexitätsmanagement in Unternehmen. Herausforderungen im Umgang mit Dynamik, Unsicherheit und Komplexität meistern. Wiesbaden: Springer Gabler, pp. 309–354.

Schoeneberg, Klaus-Peter; Zerres, Christopher; Frass, Alexander; Igelbrink, Jörg (2016): Textmining. Markenführung mittels Social Media Analytics. In Michael Lang (Ed.): Business Intelligence erfolgreich umsetzen. Von der Technologie zum Geschäftserfolg. 1. Auflage 2016, neue Ausgabe. Düsseldorf: Symposion Publishing, pp. 75–99.

Schreier, Margrit (2012): Qualitative Content Analysis in Practice. 1st ed. London u.a.: Sage Publ.

Shao, Xuhui; Li, Lexin (2011): Data-Driven Multi-Touch Attribution Models. In Chid Apte, Joydeep Ghosh, Padhraic Smyth (Eds.): the 17th ACM SIGKDD international conference. San Diego, California, USA, pp. 258–264.

Shapley, Lloyd S. (1953): A Value for n-Person Games. Contributions to the Theory of Games, Volume II. H.W. Kuhn und A.W. Tucker: Contributions to the Theory of Games, volume II. (Princeton University Press).

Sharma, Sugam (2016): Expanded Cloud Plumes Hiding Big Data Ecosystem. In *Future Gener Comput Syst* 59, pp. 63–92. DOI: 10.1016/j.future.2016.01.003.

Shearer, Colin (2000): The CRISP-DM Model: The New Blueprint for Data Mining. In *Journal of data warehousing* 5 (4), pp. 13–22.

Shugan, Steven M. (2004): The Impact of Advancing Technology on Marketing and Academic Research. In *Marketing Science* 23 (4), pp. 469–475. DOI: 10.1287/mksc.1040.0096.

Strand, Håkon H. (2016): What is One-hot Encoding and When is it Used in Data Science? Available online at https://www.quora.com/What-is-one-hot-encoding-and-when-is-it-used-in-data-science, updated on December 20th, 2016, checked on August 2nd, 2018.

Tealium Inc. (2015): EventStore Data Guide. With assistance of Akshata Yerdoor. Tealium Inc. Available online at https://community.tealiumiq.com/t5/Universal-Data-Hub/EventStore-Data-Guide/ta-p/283, checked on July, 28th 2018.

Tealium Inc. (2018): Tealium Inc. Available online at https://tealium.com/, checked on August 10th, 2018.

Tendick, Patrick H.; Denby, Lorraine; Ju, Wen-Hua (2016): Statistical Methods for Complex Event Processing and Real Time Decision Making. In *Wiley Interdiscip. Rev. Comput. Stat.* 8 (1), pp. 5–26. DOI: 10.1002/wics.1372.

Toomey, Dan (2017): Jupyter for Data Science. Exploratory Analysis, Statistical Modeling, Machine Learning, and Data Visualization with Jupyter. Birmingham, Mumbai: Packt.

Ulrich, Hans (1981): Die Betriebswirtschaftslehre als Anwendungsorientierte Sozialwissenschaft. In M. Geist, R. Köhler (Eds.): Die Führung des Betriebes. Stuttgart: Kohlhammer, pp. 1–25.

Ulrich, Hans (1985): Von der Betriebswirtschaftslehre zur Systemorientierten Managementlehre. In R. Wunderer (Ed.): Betriebswirtschaftslehre als Management- und Führungslehre. Stuttgart: Kohlhammer, pp. 3–32.

Ulrich, Hans (1995): Von der Betriebswirtschaftslehre zur systemortientierten Managementlehre. Wunderer, R., Betriebswirtschaftslehre als Management-und Führungslehre. 3rd Edition. Stuttgart, Germany: Schäffer-Poeschel.

Ulrich, Hans; Krieg, Walter; Mallik, Fredmund (1976): Zum Praxisbezug einer systemorientierten Betriebswirtschaftslehre. In: Zum Praxisbezug der Betriebswirtschaftslehre - in wissenschaftlicher Sicht. Bern / Stuttgart: Ulrich, Hans, pp. 135–151.

Uniquedigital (2012): Cross-Channel Management. Optimierung der Budgetallokation durch User-Journey Analyse und dynamisches Attributionmodelling. Available online at http://www.uniquedigital.de/fileadmin/content/file/whitepaper/uniquedigital_whitepaper_cross-channel-management.pdf, checked on January 20st, 2017.

Varan, Duane; Murphy, Jamie; Hofacker, Charles F.; Robinson, Jennifer A.; Potter, Robert F.; Bellman, Steven (2013): What Works Best When Combining Television Sets, PCs, Tablets, or Mobile Phones? In *JAR* 53 (2), pp. 212–220. DOI: 10.2501/JAR-53-2-212-220.

Verhoef, Peter C.; Kannan, P. K.; Inman, Jeffrey J. (2015): From Multi-Channel Retailing to Omni-Channel Retailing. In *Journal of Retailing* 91 (2), pp. 174–181. DOI: 10.1016/j.jretai.2015.02.005.

Voorveld, Hilde A. M. (2011): Media Multitasking and the Effectiveness of Combining Online and Radio Advertising. In *Computers in Human Behavior* 27 (6), pp. 2200–2206. DOI: 10.1016/j.chb.2011.06.016.

Vuylsteke, Alexander; Wen, Zhong; Baesens, Bart; Poelmans, Jonas (2010): Consumers' Search for Information on the Internet. How and Why China Differs from Western Europe. In *Journal of Interactive Marketing* 24 (4), pp. 309–331. DOI: 10.1016/j.intmar.2010.02.010.

Warwick, Donald P. (1982): Types of Harm in Social Research. In Tom L. Beauchamp, Ruth R. Faden, Wallace JR., Walters Leroy (Eds.): Ethical Issues in Social Science Research: Baltimore: Johns Hopkins University Press.

Webster, Jane; Watson, Richard T. (2002): Analyzing the past to prepare for the future: Writing a literature review. In *MIS Quarterly* 26 (2), pp. 13–23.

Whetten, David A. (1989): What Constitutes a Theoretical Contribution? In *The Academy of Management Review* 14 (4), p. 490. DOI: 10.2307/258554.

White, Tom (2012): Hadoop: The Definitive Guide. Storage and Analysis at Internet Scale. 3. ed. Sebastopol Calif.: O'Reilly.

White, Tom (2015): Hadoop: The Definitive Guide. Storage and Analysis at Internet Scale. 4. ed., rev. & updated. Beijing: O'Reilly.

Whitener, Michael (2015): Cookies Are So Yesterday; Cross-Device Tracking Is In – SomeTips. Available online at https://iapp.org/news/a/cookies-are-so-yesterday-cross-device-tracking-is-insome-tips/, checked on January 11th, 2016.

Wiesel, Thorsten; Pauwels, Koen; Arts, Joep (2011): Practice Prize Paper —Marketing's Profit Impact. Quantifying Online and Off-line Funnel Progression. In *Marketing Science* 30 (4), pp. 604–611. DOI: 10.1287/mksc.1100.0612.

Woo, Jong Roul; Ahn, Joongha; Lee, Jongsu; Koo, Yoonmo (2015): Media channels and consumer purchasing decisions. In *Industry Management & Data Systems (Industrial Management & Data Systems)* 115 (8), pp. 1510–1528. DOI: 10.1108/IMDS-02-2015-0036.

Xu, Lizhen; Duan, Jason A.; Whinston, Andrew (2014): Path to Purchase. A Mutually Exciting Point Process Model for Online Advertising and Conversion. In *Management Science* 60 (6), pp. 1392–1412. DOI: 10.1287/mnsc.2014.1952.

Yousafzai, Abdullah; Chang, Victor; Gani, Abdullah; Noor, Rafidah Md (2016): Multimedia augmented m-learning: Issues, trends and open challenges. In *Int J Inf Manage* 36 (5), pp. 784–792. DOI: 10.1016/j.ijinfomgt.2016.05.010.

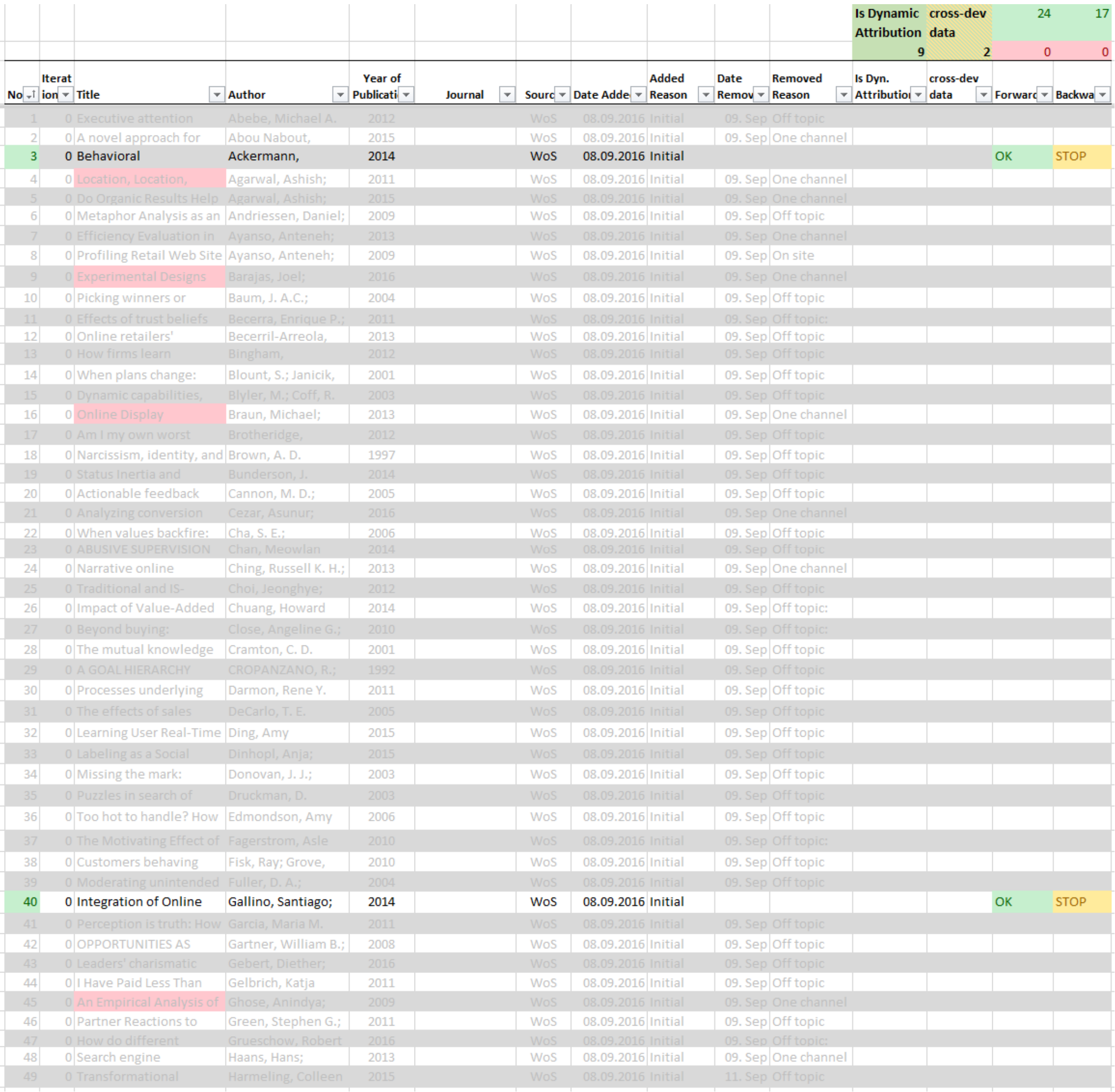
Zhang, Ya; Wei, Yi; Ren, Jianbiao (2014): Multi-touch Attribution in Online Advertising with Survival Theory. In: 2014 IEEE International Conference on Data Mining (ICDM). Shenzhen, China, pp. 687–696.

Appendices

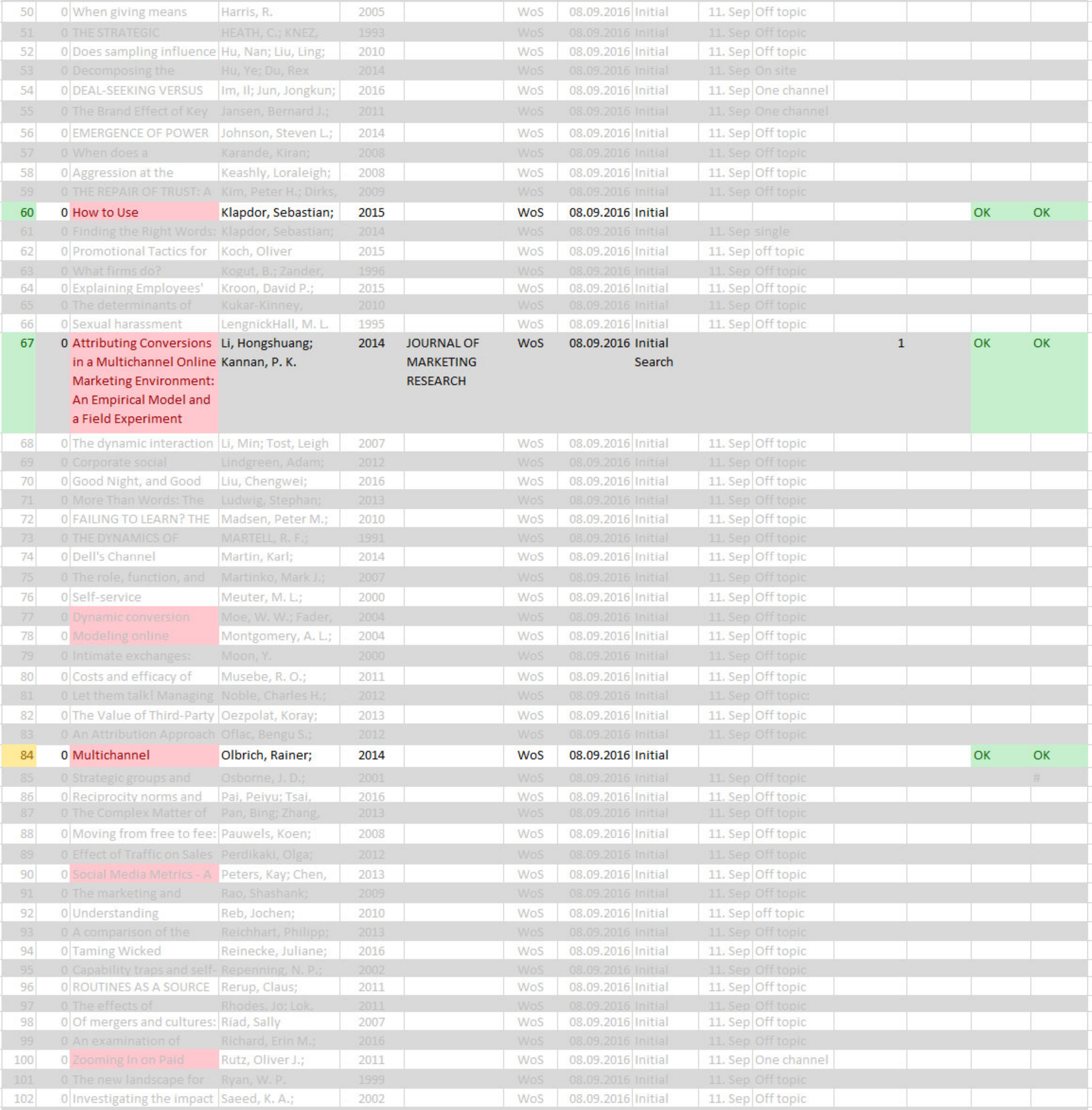
Appendix 1: Structured literature review process

The original list consists of the following additional column, which are removed for a better presentation: *DOI*, *Citations*, *Abstract*, *Keywords*, *Topic*, *Channel*, *Channel count* and *Notes*

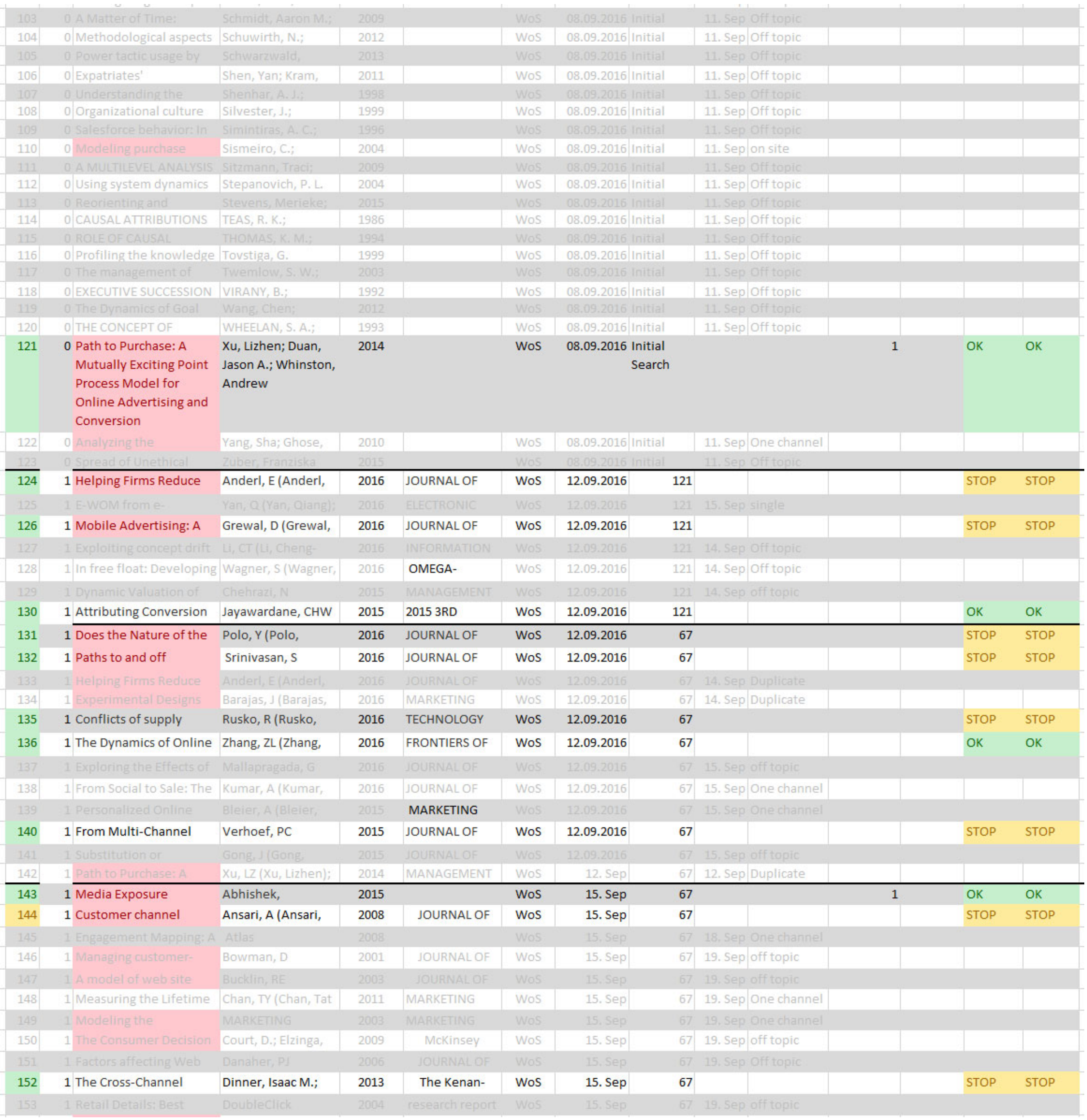
In total 632 publications (including duplicates) are analyzed. A red title indicates that the current publication is present multiple times in the list. Publications which were removed due to different reasons (see column *Removed Reason*) are grayed out. To each publication an id is assigned. The first iteration is represented by a 0 in column *Iteration*. The column *Added Reason* hold ether "initial" or a corresponding publication id. If an id is present, the current publication is added due to the publication mentioned in the column *Added Reason*.



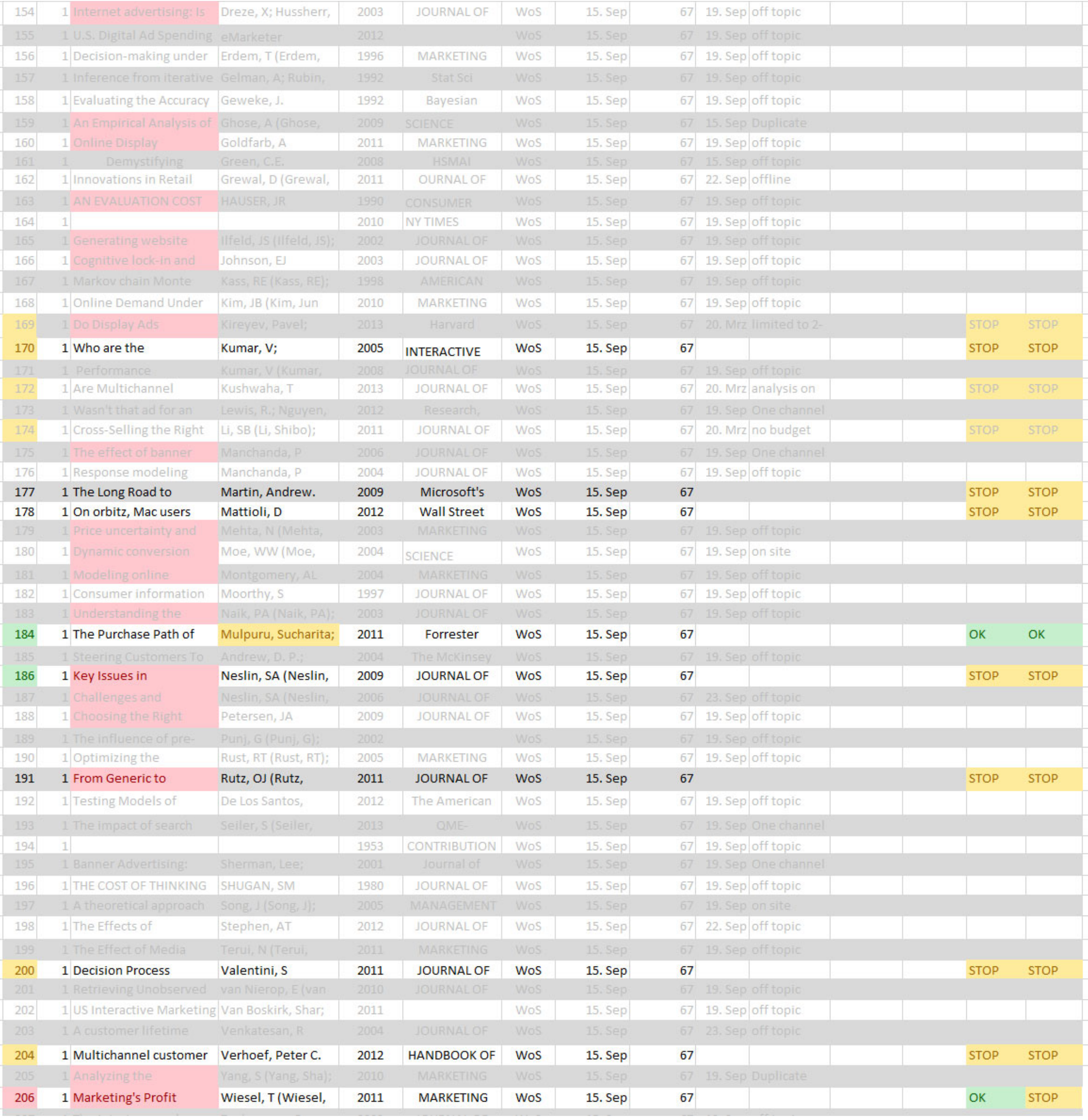
											Is Dynamic Attribution	cross-dev data	24	17
											9	2	0	0
No	Iteration	Title	Author	Year of Publicati	Journal	Sourc	Date Adde	Added Reason	Date Remov	Removed Reason	Is Dyn. Attributioi	cross-dev data	Forwarc	Backwa
1	0	Executive attention	Abebe, Michael A.	2012		WoS	08.09.2016	Initial	09. Sep	Off topic				
2	0	A novel approach for	Abou Nabout,	2015		WoS	08.09.2016	Initial	09. Sep	One channel				
3	0	Behavioral	Ackermann,	2014		WoS	08.09.2016	Initial					OK	STOP
4	0	Location, Location,	Agarwal, Ashish;	2011		WoS	08.09.2016	Initial	09. Sep	One channel				
5	0	Do Organic Results Help	Agarwal, Ashish;	2015		WoS	08.09.2016	Initial	09. Sep	One channel				
6	0	Metaphor Analysis as an	Andriessen, Daniel;	2009		WoS	08.09.2016	Initial	09. Sep	Off topic				
7	0	Efficiency Evaluation in	Ayanso, Anteneh;	2013		WoS	08.09.2016	Initial	09. Sep	One channel				
8	0	Profiling Retail Web Site	Ayanso, Anteneh;	2009		WoS	08.09.2016	Initial	09. Sep	On site				
9	0	Experimental Designs	Barajas, Joel;	2016		WoS	08.09.2016	Initial	09. Sep	One channel				
10	0	Picking winners or	Baum, J. A.C.;	2004		WoS	08.09.2016	Initial	09. Sep	Off topic				
11	0	Effects of trust beliefs	Becerra, Enrique P.;	2011		WoS	08.09.2016	Initial	09. Sep	Off topic:				
12	0	Online retailers'	Becerril-Arreola,	2013		WoS	08.09.2016	Initial	09. Sep	Off topic				
13	0	How firms learn	Bingham,	2012		WoS	08.09.2016	Initial	09. Sep	Off topic				
14	0	When plans change:	Blount, S.; Janicik,	2001		WoS	08.09.2016	Initial	09. Sep	Off topic				
15	0	Dynamic capabilities,	Blyler, M.; Coff, R.	2003		WoS	08.09.2016	Initial	09. Sep	Off topic				
16	0	Online Display	Braun, Michael;	2013		WoS	08.09.2016	Initial	09. Sep	One channel				
17	0	Am I my own worst	Brotheridge,	2012		WoS	08.09.2016	Initial	09. Sep	Off topic				
18	0	Narcissism, identity, and	Brown, A. D.	1997		WoS	08.09.2016	Initial	09. Sep	Off topic				
19	0	Status Inertia and	Bunderson, J.	2014		WoS	08.09.2016	Initial	09. Sep	Off topic				
20	0	Actionable feedback	Cannon, M. D.;	2005		WoS	08.09.2016	Initial	09. Sep	Off topic				
21	0	Analyzing conversion	Cezar, Asunur;	2016		WoS	08.09.2016	Initial	09. Sep	One channel				
22	0	When values backfire:	Cha, S. E.;	2006		WoS	08.09.2016	Initial	09. Sep	Off topic				
23	0	ABUSIVE SUPERVISION	Chan, Meowlan	2014		WoS	08.09.2016	Initial	09. Sep	Off topic				
24	0	Narrative online	Ching, Russell K. H.;	2013		WoS	08.09.2016	Initial	09. Sep	One channel				
25	0	Traditional and IS-	Choi, Jeonghye;	2012		WoS	08.09.2016	Initial	09. Sep	Off topic				
26	0	Impact of Value-Added	Chuang, Howard	2014		WoS	08.09.2016	Initial	09. Sep	Off topic:				
27	0	Beyond buying:	Close, Angeline G.;	2010		WoS	08.09.2016	Initial	09. Sep	Off topic:				
28	0	The mutual knowledge	Cramton, C. D.	2001		WoS	08.09.2016	Initial	09. Sep	Off topic				
29	0	A GOAL HIERARCHY	CROPANZANO, R.;	1992		WoS	08.09.2016	Initial	09. Sep	Off topic				
30	0	Processes underlying	Darmon, Rene Y.	2011		WoS	08.09.2016	Initial	09. Sep	Off topic				
31	0	The effects of sales	DeCarlo, T. E.	2005		WoS	08.09.2016	Initial	09. Sep	Off topic				
32	0	Learning User Real-Time	Ding, Amy	2015		WoS	08.09.2016	Initial	09. Sep	Off topic				
33	0	Labeling as a Social	Dinhopl, Anja;	2015		WoS	08.09.2016	Initial	09. Sep	Off topic				
34	0	Missing the mark:	Donovan, J. J.;	2003		WoS	08.09.2016	Initial	09. Sep	Off topic				
35	0	Puzzles in search of	Druckman, D.	2003		WoS	08.09.2016	Initial	09. Sep	Off topic				
36	0	Too hot to handle? How	Edmondson, Amy	2006		WoS	08.09.2016	Initial	09. Sep	Off topic				
37	0	The Motivating Effect of	Fagerstrom, Asle	2010		WoS	08.09.2016	Initial	09. Sep	Off topic:				
38	0	Customers behaving	Fisk, Ray; Grove,	2010		WoS	08.09.2016	Initial	09. Sep	Off topic				
39	0	Moderating unintended	Fuller, D. A.;	2004		WoS	08.09.2016	Initial	09. Sep	Off topic				
40	0	Integration of Online	Gallino, Santiago;	2014		WoS	08.09.2016	Initial					OK	STOP
41	0	Perception is truth: How	Garcia, Maria M.	2011		WoS	08.09.2016	Initial	09. Sep	Off topic				
42	0	OPPORTUNITIES AS	Gartner, William B.;	2008		WoS	08.09.2016	Initial	09. Sep	Off topic				
43	0	Leaders' charismatic	Gebert, Diether;	2016		WoS	08.09.2016	Initial	09. Sep	Off topic				
44	0	I Have Paid Less Than	Gelbrich, Katja	2011		WoS	08.09.2016	Initial	09. Sep	Off topic				
45	0	An Empirical Analysis of	Ghose, Anindya;	2009		WoS	08.09.2016	Initial	09. Sep	One channel				
46	0	Partner Reactions to	Green, Stephen G.;	2011		WoS	08.09.2016	Initial	09. Sep	Off topic				
47	0	How do different	Grueschow, Robert	2016		WoS	08.09.2016	Initial	09. Sep	Off topic:				
48	0	Search engine	Haans, Hans;	2013		WoS	08.09.2016	Initial	09. Sep	One channel				
49	0	Transformational	Harmeling, Colleen	2015		WoS	08.09.2016	Initial	11. Sep	Off topic				



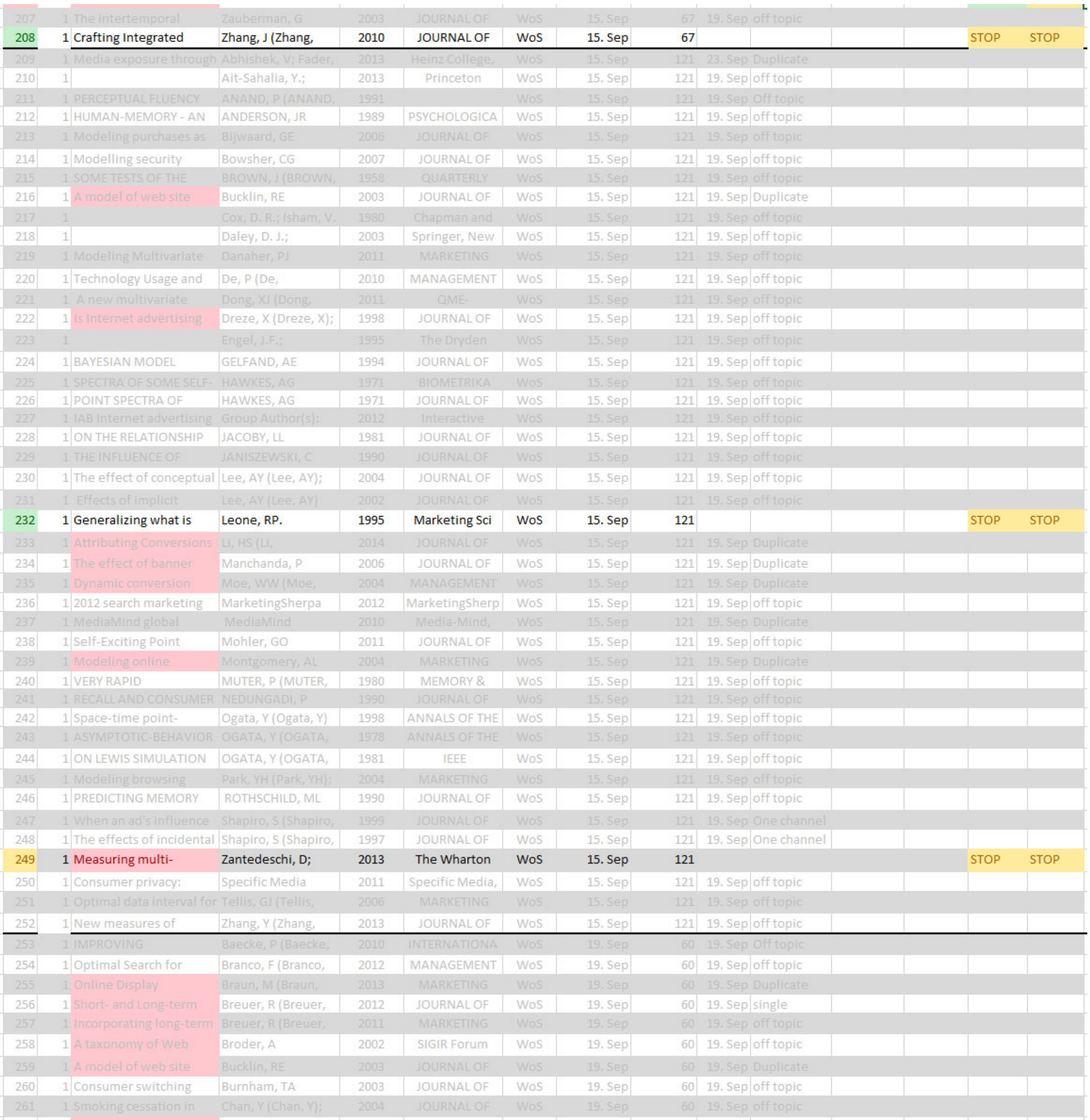
50	0	When giving means	Harris, R.	2005		WoS	08.09.2016	Initial	11. Sep	Off topic				
51	0	THE STRATEGIC	HEATH, C.; KNEZ,	1993		WoS	08.09.2016	Initial	11. Sep	Off topic				
52	0	Does sampling influence	Hu, Nan; Liu, Ling;	2010		WoS	08.09.2016	Initial	11. Sep	Off topic				
53	0	Decomposing the	Hu, Ye; Du, Rex	2014		WoS	08.09.2016	Initial	11. Sep	On site				
54	0	DEAL-SEEKING VERSUS	Im, Il; Jun, Jongkun;	2016		WoS	08.09.2016	Initial	11. Sep	One channel				
55	0	The Brand Effect of Key	Jansen, Bernard J.;	2011		WoS	08.09.2016	Initial	11. Sep	One channel				
56	0	EMERGENCE OF POWER	Johnson, Steven L.;	2014		WoS	08.09.2016	Initial	11. Sep	Off topic				
57	0	When does a	Karande, Kiran;	2008		WoS	08.09.2016	Initial	11. Sep	Off topic				
58	0	Aggression at the	Keashly, Loraleigh;	2008		WoS	08.09.2016	Initial	11. Sep	Off topic				
59	0	THE REPAIR OF TRUST: A	Kim, Peter H.; Dirks,	2009		WoS	08.09.2016	Initial	11. Sep	Off topic				
60	0	How to Use	Klapdor, Sebastian;	2015		WoS	08.09.2016	Initial					OK	OK
61	0	Finding the Right Words:	Klapdor, Sebastian;	2014		WoS	08.09.2016	Initial	11. Sep	single				
62	0	Promotional Tactics for	Koch, Oliver	2015		WoS	08.09.2016	Initial	11. Sep	off topic				
63	0	What firms do?	Kogut, B.; Zander,	1996		WoS	08.09.2016	Initial	11. Sep	Off topic				
64	0	Explaining Employees'	Kroon, David P.;	2015		WoS	08.09.2016	Initial	11. Sep	Off topic				
65	0	The determinants of	Kukar-Kinney,	2010		WoS	08.09.2016	Initial	11. Sep	Off topic				
66	0	Sexual harassment	LengnickHall, M. L.	1995		WoS	08.09.2016	Initial	11. Sep	Off topic				
67	0	Attributing Conversions in a Multichannel Online Marketing Environment: An Empirical Model and a Field Experiment	Li, Hongshuang; Kannan, P. K.	2014	JOURNAL OF MARKETING RESEARCH	WoS	08.09.2016	Initial Search			1		OK	OK
68	0	The dynamic interaction	Li, Min; Tost, Leigh	2007		WoS	08.09.2016	Initial	11. Sep	Off topic				
69	0	Corporate social	Lindgreen, Adam;	2012		WoS	08.09.2016	Initial	11. Sep	Off topic				
70	0	Good Night, and Good	Liu, Chengwei;	2016		WoS	08.09.2016	Initial	11. Sep	Off topic				
71	0	More Than Words: The	Ludwig, Stephan;	2013		WoS	08.09.2016	Initial	11. Sep	Off topic				
72	0	FAILING TO LEARN? THE	Madsen, Peter M.;	2010		WoS	08.09.2016	Initial	11. Sep	Off topic				
73	0	THE DYNAMICS OF	MARTELL, R. F.;	1991		WoS	08.09.2016	Initial	11. Sep	Off topic				
74	0	Dell's Channel	Martin, Karl;	2014		WoS	08.09.2016	Initial	11. Sep	Off topic				
75	0	The role, function, and	Martinko, Mark J.;	2007		WoS	08.09.2016	Initial	11. Sep	Off topic				
76	0	Self-service	Meuter, M. L.;	2000		WoS	08.09.2016	Initial	11. Sep	Off topic				
77	0	Dynamic conversion	Moe, W. W.; Fader,	2004		WoS	08.09.2016	Initial	11. Sep	Off topic				
78	0	Modeling online	Montgomery, A. L.;	2004		WoS	08.09.2016	Initial	11. Sep	Off topic				
79	0	Intimate exchanges:	Moon, Y.	2000		WoS	08.09.2016	Initial	11. Sep	Off topic				
80	0	Costs and efficacy of	Musebe, R. O.;	2011		WoS	08.09.2016	Initial	11. Sep	Off topic				
81	0	Let them talk! Managing	Noble, Charles H.;	2012		WoS	08.09.2016	Initial	11. Sep	Off topic:				
82	0	The Value of Third-Party	Oezpolat, Koray;	2013		WoS	08.09.2016	Initial	11. Sep	Off topic				
83	0	An Attribution Approach	Oflac, Bengu S.;	2012		WoS	08.09.2016	Initial	11. Sep	Off topic				
84	0	Multichannel	Olbrich, Rainer;	2014		WoS	08.09.2016	Initial					OK	OK
85	0	Strategic groups and	Osborne, J. D.;	2001		WoS	08.09.2016	Initial	11. Sep	Off topic				#
86	0	Reciprocity norms and	Pai, Peiyu; Tsai,	2016		WoS	08.09.2016	Initial	11. Sep	Off topic				
87	0	The Complex Matter of	Pan, Bing; Zhang,	2013		WoS	08.09.2016	Initial	11. Sep	Off topic				
88	0	Moving from free to fee:	Pauwels, Koen;	2008		WoS	08.09.2016	Initial	11. Sep	Off topic				
89	0	Effect of Traffic on Sales	Perdikaki, Olga;	2012		WoS	08.09.2016	Initial	11. Sep	Off topic				
90	0	Social Media Metrics - A	Peters, Kay; Chen,	2013		WoS	08.09.2016	Initial	11. Sep	Off topic				
91	0	The marketing and	Rao, Shashank;	2009		WoS	08.09.2016	Initial	11. Sep	Off topic				
92	0	Understanding	Reb, Jochen;	2010		WoS	08.09.2016	Initial	11. Sep	off topic				
93	0	A comparison of the	Reichhart, Philipp;	2013		WoS	08.09.2016	Initial	11. Sep	Off topic				
94	0	Taming Wicked	Reinecke, Juliane;	2016		WoS	08.09.2016	Initial	11. Sep	Off topic				
95	0	Capability traps and self-	Repenning, N. P.;	2002		WoS	08.09.2016	Initial	11. Sep	Off topic				
96	0	ROUTINES AS A SOURCE	Rerup, Claus;	2011		WoS	08.09.2016	Initial	11. Sep	Off topic				
97	0	The effects of	Rhodes, Jo; Lok,	2011		WoS	08.09.2016	Initial	11. Sep	Off topic				
98	0	Of mergers and cultures:	Riad, Sally	2007		WoS	08.09.2016	Initial	11. Sep	Off topic				
99	0	An examination of	Richard, Erin M.;	2016		WoS	08.09.2016	Initial	11. Sep	Off topic				
100	0	Zooming In on Paid	Rutz, Oliver J.;	2011		WoS	08.09.2016	Initial	11. Sep	One channel				
101	0	The new landscape for	Ryan, W. P.	1999		WoS	08.09.2016	Initial	11. Sep	Off topic				
102	0	Investigating the impact	Saeed, K. A.;	2002		WoS	08.09.2016	Initial	11. Sep	Off topic				



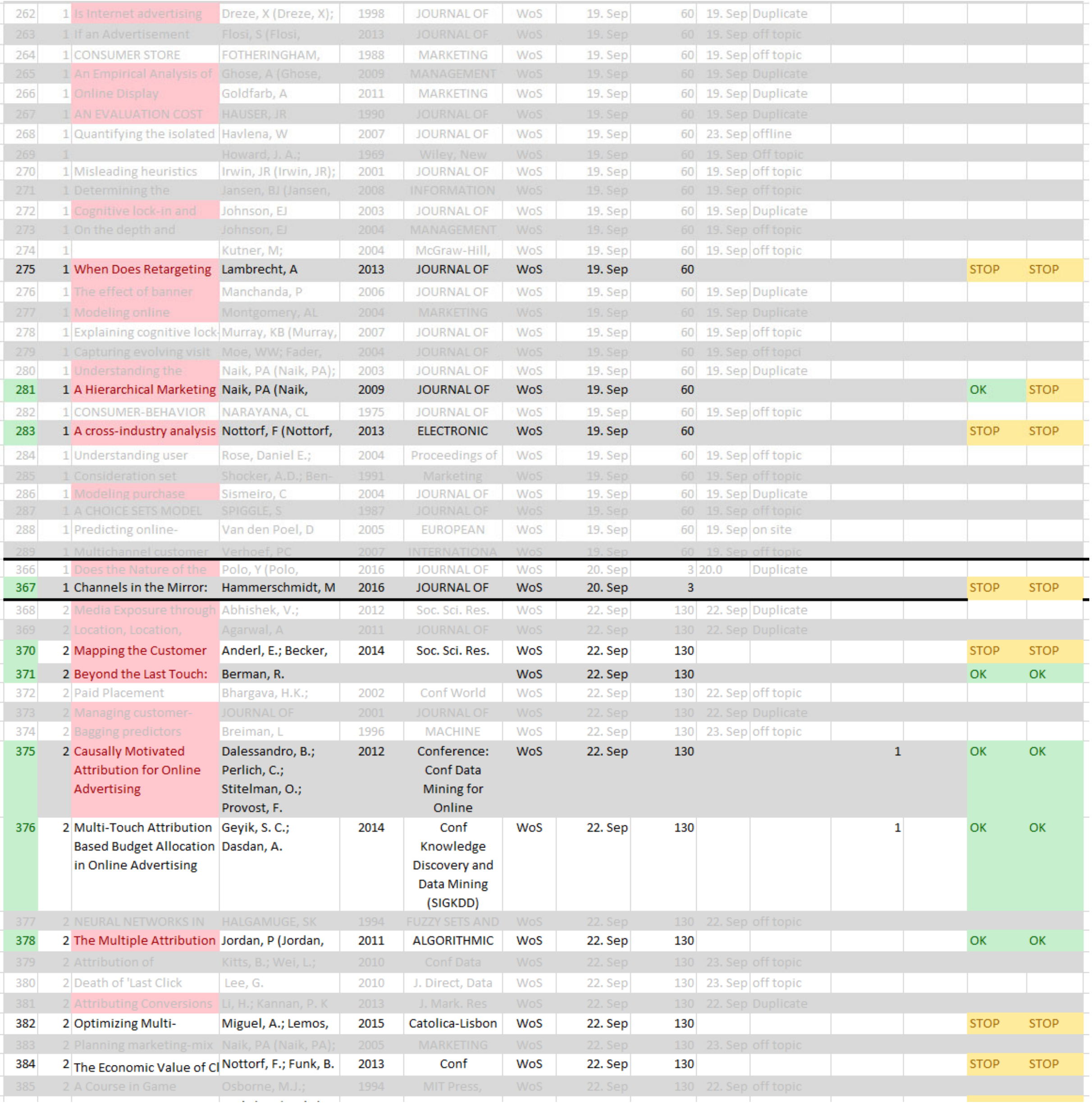
103	0	A Matter of Time:	Schmidt, Aaron M.;	2009		WoS	08.09.2016	Initial	11. Sep	Off topic				
104	0	Methodological aspects	Schuwirth, N.;	2012		WoS	08.09.2016	Initial	11. Sep	Off topic				
105	0	Power tactic usage by	Schwarzwald,	2013		WoS	08.09.2016	Initial	11. Sep	Off topic				
106	0	Expatriates'	Shen, Yan; Kram,	2011		WoS	08.09.2016	Initial	11. Sep	Off topic				
107	0	Understanding the	Shenhar, A. J.;	1998		WoS	08.09.2016	Initial	11. Sep	Off topic				
108	0	Organizational culture	Silvester, J.;	1999		WoS	08.09.2016	Initial	11. Sep	Off topic				
109	0	Salesforce behavior: In	Simintiras, A. C.;	1996		WoS	08.09.2016	Initial	11. Sep	Off topic				
110	0	Modeling purchase	Sismeiro, C.;	2004		WoS	08.09.2016	Initial	11. Sep	on site				
111	0	A MULTILEVEL ANALYSIS	Sitzmann, Traci;	2009		WoS	08.09.2016	Initial	11. Sep	Off topic				
112	0	Using system dynamics	Stepanovich, P. L.	2004		WoS	08.09.2016	Initial	11. Sep	Off topic				
113	0	Reorienting and	Stevens, Merieke;	2015		WoS	08.09.2016	Initial	11. Sep	Off topic				
114	0	CAUSAL ATTRIBUTIONS	TEAS, R. K.;	1986		WoS	08.09.2016	Initial	11. Sep	Off topic				
115	0	ROLE OF CAUSAL	THOMAS, K. M.;	1994		WoS	08.09.2016	Initial	11. Sep	Off topic				
116	0	Profiling the knowledge	Tovstiga, G.	1999		WoS	08.09.2016	Initial	11. Sep	Off topic				
117	0	The management of	Twemlow, S. W.;	2003		WoS	08.09.2016	Initial	11. Sep	Off topic				
118	0	EXECUTIVE SUCCESSION	VIRANY, B.;	1992		WoS	08.09.2016	Initial	11. Sep	Off topic				
119	0	The Dynamics of Goal	Wang, Chen;	2012		WoS	08.09.2016	Initial	11. Sep	Off topic				
120	0	THE CONCEPT OF	WHEELAN, S. A.;	1993		WoS	08.09.2016	Initial	11. Sep	Off topic				
121	0	Path to Purchase: A Mutually Exciting Point Process Model for Online Advertising and Conversion	Xu, Lizhen; Duan, Jason A.; Whinston, Andrew	2014		WoS	08.09.2016	Initial Search			1		OK	OK
122	0	Analyzing the	Yang, Sha; Ghose,	2010		WoS	08.09.2016	Initial	11. Sep	One channel				
123	0	Spread of Unethical	Zuber, Franziska	2015		WoS	08.09.2016	Initial	11. Sep	Off topic				
124	1	Helping Firms Reduce	Anderl, E (Anderl,	2016	JOURNAL OF	WoS	12.09.2016	121					STOP	STOP
125	1	E-WOM from e-	Yan, Q (Yan, Qiang);	2016	ELECTRONIC	WoS	12.09.2016	121	15. Sep	single				
126	1	Mobile Advertising: A	Grewal, D (Grewal,	2016	JOURNAL OF	WoS	12.09.2016	121					STOP	STOP
127	1	Exploiting concept drift	Li, CT (Li, Cheng-	2016	INFORMATION	WoS	12.09.2016	121	14. Sep	Off topic				
128	1	In free float: Developing	Wagner, S (Wagner,	2016	OMEGA-	WoS	12.09.2016	121	14. Sep	Off topic				
129	1	Dynamic Valuation of	Chehrazi, N	2015	MANAGEMENT	WoS	12.09.2016	121	14. Sep	off topic				
130	1	Attributing Conversion	Jayawardane, CHW	2015	2015 3RD	WoS	12.09.2016	121					OK	OK
131	1	Does the Nature of the	Polo, Y (Polo,	2016	JOURNAL OF	WoS	12.09.2016	67					STOP	STOP
132	1	Paths to and off	Srinivasan, S	2016	JOURNAL OF	WoS	12.09.2016	67					STOP	STOP
133	1	Helping Firms Reduce	Anderl, E (Anderl,	2016	JOURNAL OF	WoS	12.09.2016	67	14. Sep	Duplicate				
134	1	Experimental Designs	Barajas, J (Barajas,	2016	MARKETING	WoS	12.09.2016	67	14. Sep	Duplicate				
135	1	Conflicts of supply	Rusko, R (Rusko,	2016	TECHNOLOGY	WoS	12.09.2016	67					STOP	STOP
136	1	The Dynamics of Online	Zhang, ZL (Zhang,	2016	FRONTIERS OF	WoS	12.09.2016	67					OK	OK
137	1	Exploring the Effects of	Mallapragada, G	2016	JOURNAL OF	WoS	12.09.2016	67	15. Sep	off topic				
138	1	From Social to Sale: The	Kumar, A (Kumar,	2016	JOURNAL OF	WoS	12.09.2016	67	15. Sep	One channel				
139	1	Personalized Online	Bleier, A (Bleier,	2015	MARKETING	WoS	12.09.2016	67	15. Sep	One channel				
140	1	From Multi-Channel	Verhoef, PC	2015	JOURNAL OF	WoS	12.09.2016	67					STOP	STOP
141	1	Substitution or	Gong, J (Gong,	2015	JOURNAL OF	WoS	12.09.2016	67	15. Sep	off topic				
142	1	Path to Purchase: A	Xu, LZ (Xu, Lizhen);	2014	MANAGEMENT	WoS	12. Sep	67	12. Sep	Duplicate				
143	1	Media Exposure	Abhishek,	2015		WoS	15. Sep	67			1		OK	OK
144	1	Customer channel	Ansari, A (Ansari,	2008	JOURNAL OF	WoS	15. Sep	67					STOP	STOP
145	1	Engagement Mapping: A	Atlas	2008		WoS	15. Sep	67	18. Sep	One channel				
146	1	Managing customer-	Bowman, D	2001	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
147	1	A model of web site	Bucklin, RE	2003	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
148	1	Measuring the Lifetime	Chan, TY (Chan, Tat	2011	MARKETING	WoS	15. Sep	67	19. Sep	One channel				
149	1	Modeling the	MARKETING	2003	MARKETING	WoS	15. Sep	67	19. Sep	One channel				
150	1	The Consumer Decision	Court, D.; Elzinga,	2009	McKinsey	WoS	15. Sep	67	19. Sep	off topic				
151	1	Factors affecting Web	Danaher, PJ	2006	JOURNAL OF	WoS	15. Sep	67	19. Sep	Off topic				
152	1	The Cross-Channel	Dinner, Isaac M.;	2013	The Kenan-	WoS	15. Sep	67					STOP	STOP
153	1	Retail Details: Best	DoubleClick	2004	research report	WoS	15. Sep	67	19. Sep	off topic				



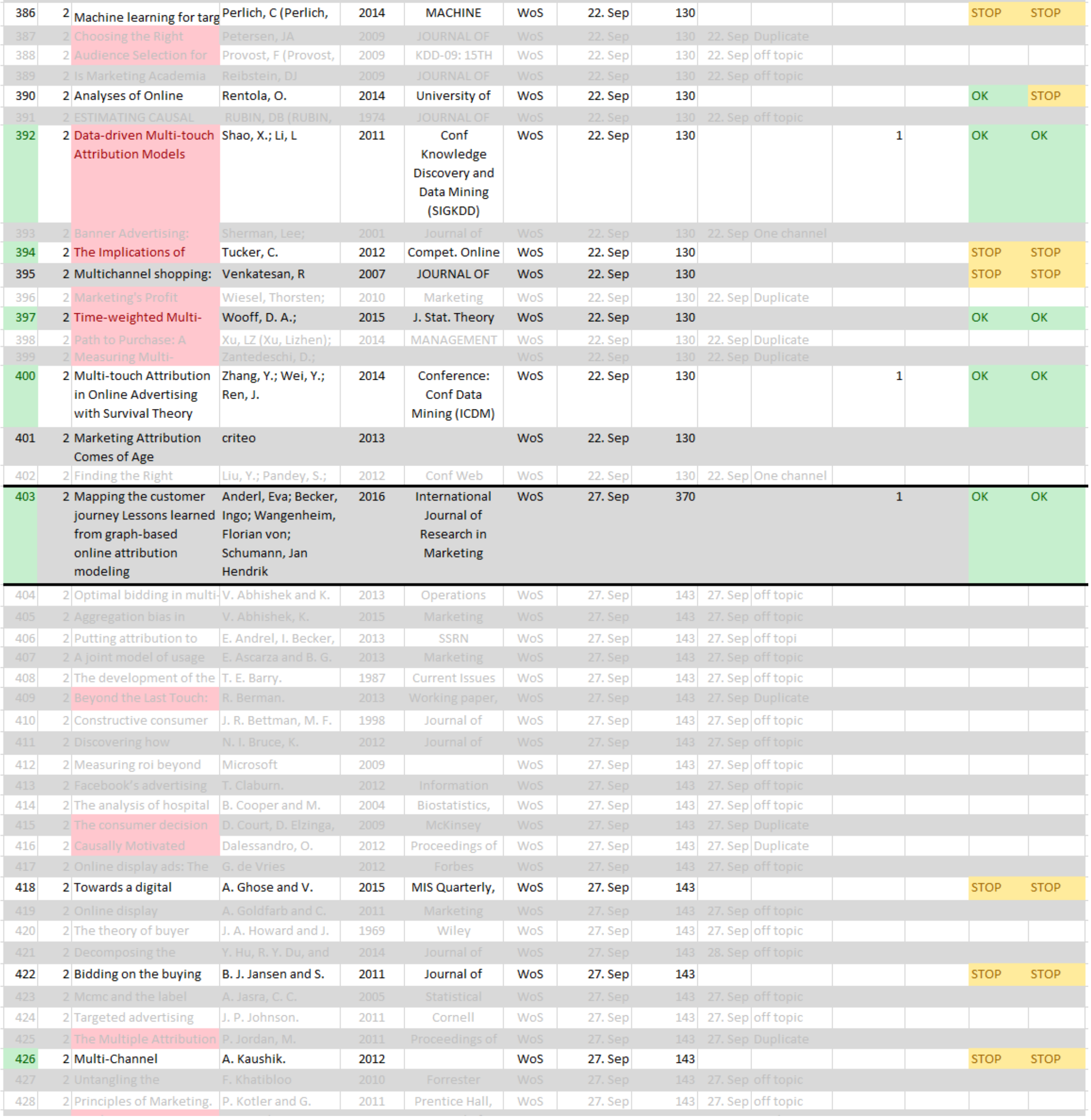
154	1	Internet advertising: Is	Dreze, X; Hussherr,	2003	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
155	1	U.S. Digital Ad Spending	eMarketer	2012		WoS	15. Sep	67	19. Sep	off topic				
156	1	Decision-making under	Erdem, T (Erdem,	1996	MARKETING	WoS	15. Sep	67	19. Sep	off topic				
157	1	Inference from iterative	Gelman, A; Rubin,	1992	Stat Sci	WoS	15. Sep	67	19. Sep	off topic				
158	1	Evaluating the Accuracy	Geweke, J.	1992	Bayesian	WoS	15. Sep	67	19. Sep	off topic				
159	1	An Empirical Analysis of	Ghose, A (Ghose,	2009	SCIENCE	WoS	15. Sep	67	15. Sep	Duplicate				
160	1	Online Display	Goldfarb, A	2011	MARKETING	WoS	15. Sep	67	19. Sep	off topic				
161	1	Demystifying	Green, C.E.	2008	HSMAI	WoS	15. Sep	67	15. Sep	off topic				
162	1	Innovations in Retail	Grewal, D (Grewal,	2011	OURNAL OF	WoS	15. Sep	67	22. Sep	offline				
163	1	AN EVALUATION COST	HAUSER, JR	1990	CONSUMER	WoS	15. Sep	67	19. Sep	off topic				
164	1			2010	NY TIMES	WoS	15. Sep	67	19. Sep	off topic				
165	1	Generating website	Ilfeld, JS (Ilfeld, JS);	2002	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
166	1	Cognitive lock-in and	Johnson, EJ	2003	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
167	1	Markov chain Monte	Kass, RE (Kass, RE);	1998	AMERICAN	WoS	15. Sep	67	19. Sep	off topic				
168	1	Online Demand Under	Kim, JB (Kim, Jun	2010	MARKETING	WoS	15. Sep	67	19. Sep	off topic				
169	1	Do Display Ads	Kireyev, Pavel;	2013	Harvard	WoS	15. Sep	67	20. Mrz	limited to 2-			STOP	STOP
170	1	Who are the	Kumar, V;	2005	INTERACTIVE	WoS	15. Sep	67					STOP	STOP
171	1	Performance	Kumar, V (Kumar,	2008	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
172	1	Are Multichannel	Kushwaha, T	2013	JOURNAL OF	WoS	15. Sep	67	20. Mrz	analysis on			STOP	STOP
173	1	Wasn't that ad for an	Lewis, R.; Nguyen,	2012	Research,	WoS	15. Sep	67	19. Sep	One channel				
174	1	Cross-Selling the Right	Li, SB (Li, Shibo);	2011	JOURNAL OF	WoS	15. Sep	67	20. Mrz	no budget			STOP	STOP
175	1	The effect of banner	Manchanda, P	2006	JOURNAL OF	WoS	15. Sep	67	19. Sep	One channel				
176	1	Response modeling	Manchanda, P	2004	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
177	1	The Long Road to	Martin, Andrew.	2009	Microsoft's	WoS	15. Sep	67					STOP	STOP
178	1	On orbitz, Mac users	Mattioli, D	2012	Wall Street	WoS	15. Sep	67					STOP	STOP
179	1	Price uncertainty and	Mehta, N (Mehta,	2003	MARKETING	WoS	15. Sep	67	19. Sep	off topic				
180	1	Dynamic conversion	Moe, WW (Moe,	2004	SCIENCE	WoS	15. Sep	67	19. Sep	on site				
181	1	Modeling online	Montgomery, AL	2004	MARKETING	WoS	15. Sep	67	19. Sep	off topic				
182	1	Consumer information	Moorthy, S	1997	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
183	1	Understanding the	Naik, PA (Naik, PA);	2003	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
184	1	The Purchase Path of	Mulpuru, Sucharita;	2011	Forrester	WoS	15. Sep	67					OK	OK
185	1	Steering Customers To	Andrew, D. P.;	2004	The McKinsey	WoS	15. Sep	67	19. Sep	off topic				
186	1	Key Issues in	Neslin, SA (Neslin,	2009	JOURNAL OF	WoS	15. Sep	67					STOP	STOP
187	1	Challenges and	Neslin, SA (Neslin,	2006	JOURNAL OF	WoS	15. Sep	67	23. Sep	off topic				
188	1	Choosing the Right	Petersen, JA	2009	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
189	1	The influence of pre-	Punj, G (Punj, G);	2002		WoS	15. Sep	67	19. Sep	off topic				
190	1	Optimizing the	Rust, RT (Rust, RT);	2005	MARKETING	WoS	15. Sep	67	19. Sep	off topic				
191	1	From Generic to	Rutz, OJ (Rutz,	2011	JOURNAL OF	WoS	15. Sep	67					STOP	STOP
192	1	Testing Models of	De Los Santos,	2012	The American	WoS	15. Sep	67	19. Sep	off topic				
193	1	The impact of search	Seiler, S (Seiler,	2013	QME-	WoS	15. Sep	67	19. Sep	One channel				
194	1			1953	CONTRIBUTION	WoS	15. Sep	67	19. Sep	off topic				
195	1	Banner Advertising:	Sherman, Lee;	2001	Journal of	WoS	15. Sep	67	19. Sep	One channel				
196	1	THE COST OF THINKING	SHUGAN, SM	1980	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
197	1	A theoretical approach	Song, J (Song, J);	2005	MANAGEMENT	WoS	15. Sep	67	19. Sep	on site				
198	1	The Effects of	Stephen, AT	2012	JOURNAL OF	WoS	15. Sep	67	22. Sep	off topic				
199	1	The Effect of Media	Terui, N (Terui,	2011	MARKETING	WoS	15. Sep	67	19. Sep	off topic				
200	1	Decision Process	Valentini, S	2011	JOURNAL OF	WoS	15. Sep	67					STOP	STOP
201	1	Retrieving Unobserved	van Nierop, E (van	2010	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
202	1	US Interactive Marketing	Van Boskirk, Shar;	2011		WoS	15. Sep	67	19. Sep	off topic				
203	1	A customer lifetime	Venkatesan, R	2004	JOURNAL OF	WoS	15. Sep	67	23. Sep	off topic				
204	1	Multichannel customer	Verhoef, Peter C.	2012	HANDBOOK OF	WoS	15. Sep	67					STOP	STOP
205	1	Analyzing the	Yang, S (Yang, Sha);	2010	MARKETING	WoS	15. Sep	67	19. Sep	Duplicate				
206	1	Marketing's Profit	Wiesel, T (Wiesel,	2011	MARKETING	WoS	15. Sep	67					OK	STOP



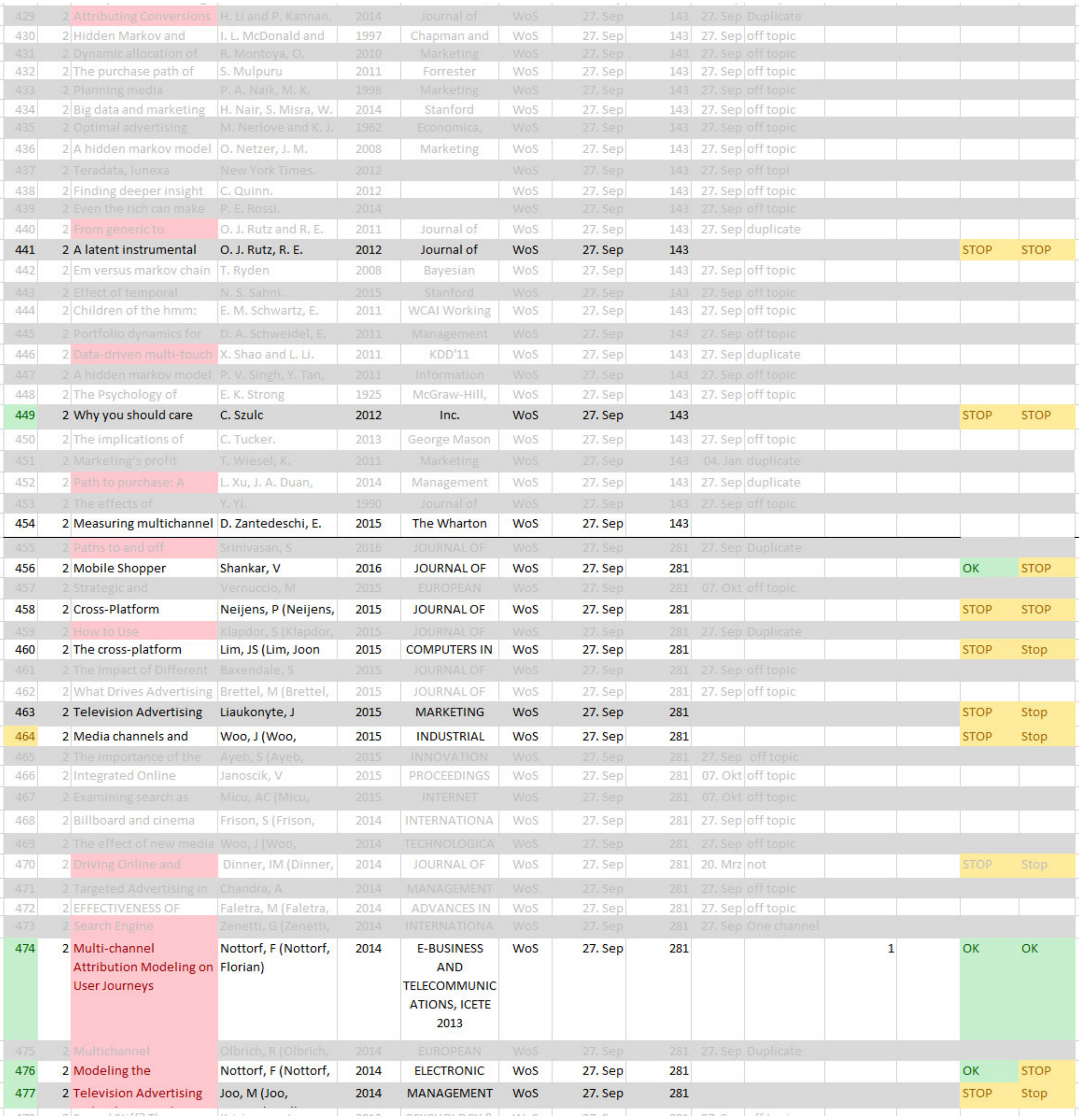
207	1	The intertemporal	Zauberman, G	2003	JOURNAL OF	WoS	15. Sep	67	19. Sep	off topic				
208	1	Crafting Integrated	Zhang, J (Zhang,	2010	JOURNAL OF	WoS	15. Sep	67					STOP	STOP
209	1	Media exposure through	Abhishek, V; Fader,	2013	Heinz College,	WoS	15. Sep	121	23. Sep	Duplicate				
210	1		Ait-Sahalia, Y.;	2013	Princeton	WoS	15. Sep	121	19. Sep	off topic				
211	1	PERCEPTUAL FLUENCY	ANAND, P (ANAND,	1991		WoS	15. Sep	121	19. Sep	Off topic				
212	1	HUMAN-MEMORY - AN	ANDERSON, JR	1989	PSYCHOLOGICA	WoS	15. Sep	121	19. Sep	off topic				
213	1	Modeling purchases as	Bijwaard, GE	2006	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
214	1	Modelling security	Bowsher, CG	2007	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
215	1	SOME TESTS OF THE	BROWN, J (BROWN,	1958	QUARTERLY	WoS	15. Sep	121	19. Sep	off topic				
216	1	A model of web site	Bucklin, RE	2003	JOURNAL OF	WoS	15. Sep	121	19. Sep	Duplicate				
217	1		Cox, D. R.; Isham, V.	1980	Chapman and	WoS	15. Sep	121	19. Sep	off topic				
218	1		Daley, D. J.;	2003	Springer, New	WoS	15. Sep	121	19. Sep	off topic				
219	1	Modeling Multivariate	Danaher, PJ	2011	MARKETING	WoS	15. Sep	121	19. Sep	off topic				
220	1	Technology Usage and	De, P (De,	2010	MANAGEMENT	WoS	15. Sep	121	19. Sep	off topic				
221	1	A new multivariate	Dong, XJ (Dong,	2011	QME-	WoS	15. Sep	121	19. Sep	off topic				
222	1	Is Internet advertising	Dreze, X (Dreze, X);	1998	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
223	1		Engel, J.F.;	1995	The Dryden	WoS	15. Sep	121	19. Sep	off topic				
224	1	BAYESIAN MODEL	GELFAND, AE	1994	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
225	1	SPECTRA OF SOME SELF-	HAWKES, AG	1971	BIOMETRIKA	WoS	15. Sep	121	19. Sep	off topic				
226	1	POINT SPECTRA OF	HAWKES, AG	1971	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
227	1	IAB Internet advertising	Group Author(s):	2012	Interactive	WoS	15. Sep	121	19. Sep	off topic				
228	1	ON THE RELATIONSHIP	JACOBY, LL	1981	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
229	1	THE INFLUENCE OF	JANISZEWSKI, C	1990	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
230	1	The effect of conceptual	Lee, AY (Lee, AY);	2004	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
231	1	Effects of implicit	Lee, AY (Lee, AY)	2002	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
232	1	Generalizing what is	Leone, RP.	1995	Marketing Sci	WoS	15. Sep	121					STOP	STOP
233	1	Attributing Conversions	Li, HS (Li,	2014	JOURNAL OF	WoS	15. Sep	121	19. Sep	Duplicate				
234	1	The effect of banner	Manchanda, P	2006	JOURNAL OF	WoS	15. Sep	121	19. Sep	Duplicate				
235	1	Dynamic conversion	Moe, WW (Moe,	2004	MANAGEMENT	WoS	15. Sep	121	19. Sep	Duplicate				
236	1	2012 search marketing	MarketingSherpa	2012	MarketingSherp	WoS	15. Sep	121	19. Sep	off topic				
237	1	MediaMind global	MediaMind	2010	Media-Mind,	WoS	15. Sep	121	19. Sep	Duplicate				
238	1	Self-Exciting Point	Mohler, GO	2011	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
239	1	Modeling online	Montgomery, AL	2004	MARKETING	WoS	15. Sep	121	19. Sep	Duplicate				
240	1	VERY RAPID	MUTER, P (MUTER,	1980	MEMORY &	WoS	15. Sep	121	19. Sep	off topic				
241	1	RECALL AND CONSUMER	NEDUNGADI, P	1990	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
242	1	Space-time point-	Ogata, Y (Ogata, Y)	1998	ANNALS OF THE	WoS	15. Sep	121	19. Sep	off topic				
243	1	ASYMPTOTIC-BEHAVIOR	OGATA, Y (OGATA,	1978	ANNALS OF THE	WoS	15. Sep	121	19. Sep	off topic				
244	1	ON LEWIS SIMULATION	OGATA, Y (OGATA,	1981	IEEE	WoS	15. Sep	121	19. Sep	off topic				
245	1	Modeling browsing	Park, YH (Park, YH);	2004	MARKETING	WoS	15. Sep	121	19. Sep	off topic				
246	1	PREDICTING MEMORY	ROTHSCHILD, ML	1990	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
247	1	When an ad's influence	Shapiro, S (Shapiro,	1999	JOURNAL OF	WoS	15. Sep	121	19. Sep	One channel				
248	1	The effects of incidental	Shapiro, S (Shapiro,	1997	JOURNAL OF	WoS	15. Sep	121	19. Sep	One channel				
249	1	Measuring multi-	Zantedeschi, D;	2013	The Wharton	WoS	15. Sep	121					STOP	STOP
250	1	Consumer privacy:	Specific Media	2011	Specific Media,	WoS	15. Sep	121	19. Sep	off topic				
251	1	Optimal data interval for	Tellis, GJ (Tellis,	2006	MARKETING	WoS	15. Sep	121	19. Sep	off topic				
252	1	New measures of	Zhang, Y (Zhang,	2013	JOURNAL OF	WoS	15. Sep	121	19. Sep	off topic				
253	1	IMPROVING	Baecke, P (Baecke,	2010	INTERNATIONA	WoS	19. Sep	60	19. Sep	Off topic				
254	1	Optimal Search for	Branco, F (Branco,	2012	MANAGEMENT	WoS	19. Sep	60	19. Sep	off topic				
255	1	Online Display	Braun, M (Braun,	2013	MARKETING	WoS	19. Sep	60	19. Sep	Duplicate				
256	1	Short- and Long-term	Breuer, R (Breuer,	2012	JOURNAL OF	WoS	19. Sep	60	19. Sep	single				
257	1	Incorporating long-term	Breuer, R (Breuer,	2011	MARKETING	WoS	19. Sep	60	19. Sep	off topic				
258	1	A taxonomy of Web	Broder, A	2002	SIGIR Forum	WoS	19. Sep	60	19. Sep	off topic				
259	1	A model of web site	Bucklin, RE	2003	JOURNAL OF	WoS	19. Sep	60	19. Sep	Duplicate				
260	1	Consumer switching	Burnham, TA	2003	JOURNAL OF	WoS	19. Sep	60	19. Sep	off topic				
261	1	Smoking cessation in	Chan, Y (Chan, Y);	2004	JOURNAL OF	WoS	19. Sep	60	19. Sep	off topic				



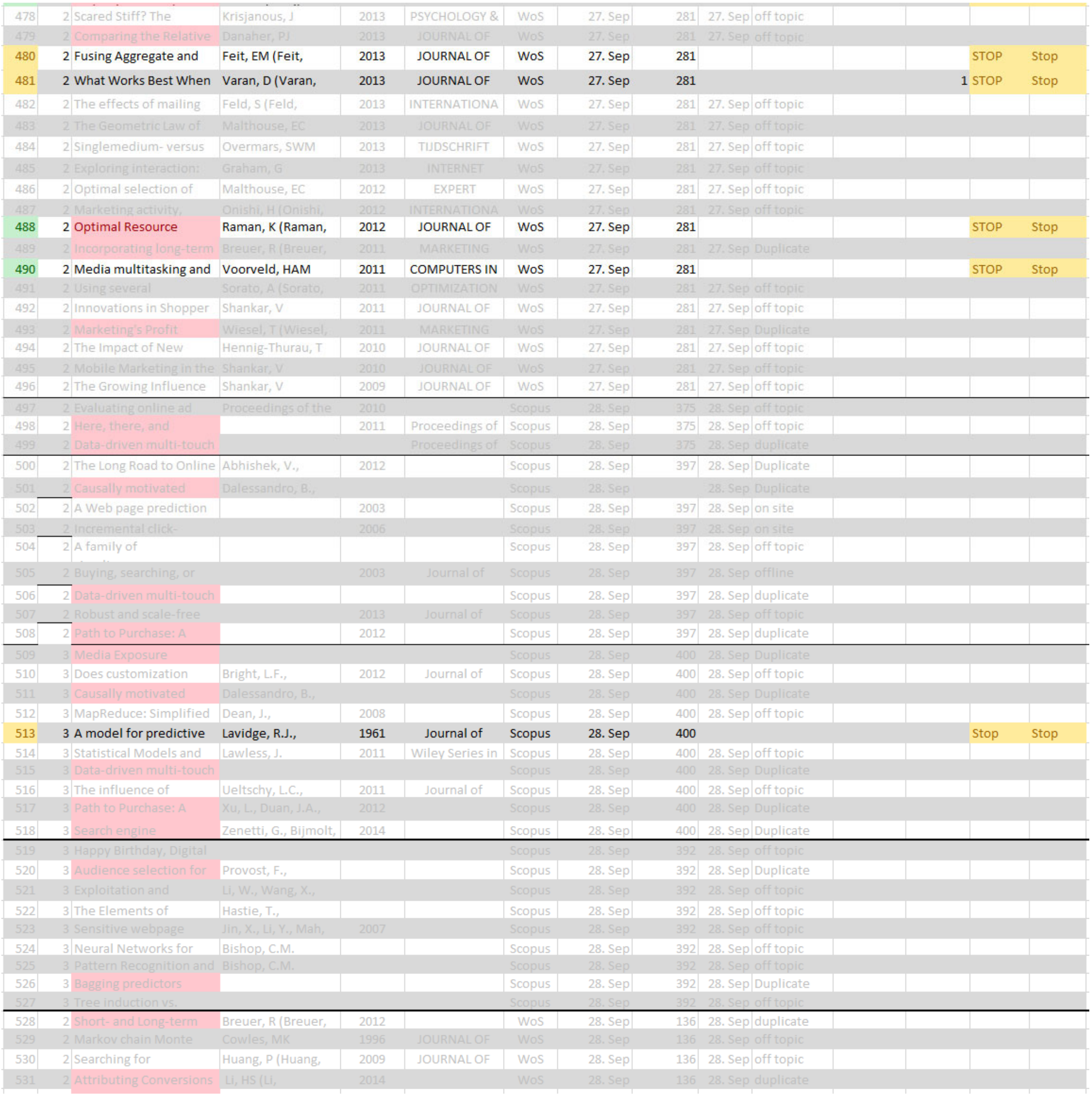
262	1	Is Internet advertising	Dreze, X (Dreze, X);	1998	JOURNAL OF	WoS	19. Sep	60	19. Sep	Duplicate				
263	1	If an Advertisement	Flosi, S (Flosi,	2013	JOURNAL OF	WoS	19. Sep	60	19. Sep	off topic				
264	1	CONSUMER STORE	FOTHERINGHAM,	1988	MARKETING	WoS	19. Sep	60	19. Sep	off topic				
265	1	An Empirical Analysis of	Ghose, A (Ghose,	2009	MANAGEMENT	WoS	19. Sep	60	19. Sep	Duplicate				
266	1	Online Display	Goldfarb, A	2011	MARKETING	WoS	19. Sep	60	19. Sep	Duplicate				
267	1	AN EVALUATION COST	HAUSER, JR	1990	JOURNAL OF	WoS	19. Sep	60	19. Sep	Duplicate				
268	1	Quantifying the isolated	Havlena, W	2007	JOURNAL OF	WoS	19. Sep	60	23. Sep	offline				
269	1		Howard, J. A.;	1969	Wiley, New	WoS	19. Sep	60	19. Sep	Off topic				
270	1	Misleading heuristics	Irwin, JR (Irwin, JR);	2001	JOURNAL OF	WoS	19. Sep	60	19. Sep	off topic				
271	1	Determining the	Jansen, BJ (Jansen,	2008	INFORMATION	WoS	19. Sep	60	19. Sep	off topic				
272	1	Cognitive lock-in and	Johnson, EJ	2003	JOURNAL OF	WoS	19. Sep	60	19. Sep	Duplicate				
273	1	On the depth and	Johnson, EJ	2004	MANAGEMENT	WoS	19. Sep	60	19. Sep	off topic				
274	1		Kutner, M;	2004	McGraw-Hill,	WoS	19. Sep	60	19. Sep	off topic				
275	1	When Does Retargeting	Lambrecht, A	2013	JOURNAL OF	WoS	19. Sep	60					STOP	STOP
276	1	The effect of banner	Manchanda, P	2006	JOURNAL OF	WoS	19. Sep	60	19. Sep	Duplicate				
277	1	Modeling online	Montgomery, AL	2004	MARKETING	WoS	19. Sep	60	19. Sep	Duplicate				
278	1	Explaining cognitive lock-	Murray, KB (Murray,	2007	JOURNAL OF	WoS	19. Sep	60	19. Sep	off topic				
279	1	Capturing evolving visit	Moe, WW; Fader,	2004	JOURNAL OF	WoS	19. Sep	60	19. Sep	off topci				
280	1	Understanding the	Naik, PA (Naik, PA);	2003	JOURNAL OF	WoS	19. Sep	60	19. Sep	Duplicate				
281	1	A Hierarchical Marketing	Naik, PA (Naik,	2009	JOURNAL OF	WoS	19. Sep	60					OK	STOP
282	1	CONSUMER-BEHAVIOR	NARAYANA, CL	1975	JOURNAL OF	WoS	19. Sep	60	19. Sep	off topic				
283	1	A cross-industry analysis	Nottorf, F (Nottorf,	2013	ELECTRONIC	WoS	19. Sep	60					STOP	STOP
284	1	Understanding user	Rose, Daniel E.;	2004	Proceedings of	WoS	19. Sep	60	19. Sep	off topic				
285	1	Consideration set	Shocker, A.D.; Ben-	1991	Marketing	WoS	19. Sep	60	19. Sep	off topic				
286	1	Modeling purchase	Sismeiro, C	2004	JOURNAL OF	WoS	19. Sep	60	19. Sep	Duplicate				
287	1	A CHOICE SETS MODEL	SPIGGLE, S	1987	JOURNAL OF	WoS	19. Sep	60	19. Sep	off topic				
288	1	Predicting online-	Van den Poel, D	2005	EUROPEAN	WoS	19. Sep	60	19. Sep	on site				
289	1	Multichannel customer	Verhoef, PC	2007	INTERNATIONA	WoS	19. Sep	60	19. Sep	off topic				
366	1	Does the Nature of the	Polo, Y (Polo,	2016	JOURNAL OF	WoS	20. Sep	3	20.0	Duplicate				
367	1	Channels in the Mirror:	Hammerschmidt, M	2016	JOURNAL OF	WoS	20. Sep	3					STOP	STOP
368	2	Media Exposure through	Abhishek, V.;	2012	Soc. Sci. Res.	WoS	22. Sep	130	22. Sep	Duplicate				
369	2	Location, Location,	Agarwal, A	2011	JOURNAL OF	WoS	22. Sep	130	22. Sep	Duplicate				
370	2	Mapping the Customer	Anderl, E.; Becker,	2014	Soc. Sci. Res.	WoS	22. Sep	130					STOP	STOP
371	2	Beyond the Last Touch:	Berman, R.			WoS	22. Sep	130					OK	OK
372	2	Paid Placement	Bhargava, H.K.;	2002	Conf World	WoS	22. Sep	130	22. Sep	off topic				
373	2	Managing customer-	JOURNAL OF	2001	JOURNAL OF	WoS	22. Sep	130	22. Sep	Duplicate				
374	2	Bagging predictors	Breiman, L	1996	MACHINE	WoS	22. Sep	130	23. Sep	off topic				
375	2	Causally Motivated Attribution for Online Advertising	Dalessandro, B.; Perlich, C.; Stitelman, O.; Provost, F.	2012	Conference: Conf Data Mining for Online	WoS	22. Sep	130			1		OK	OK
376	2	Multi-Touch Attribution Based Budget Allocation in Online Advertising	Geyik, S. C.; Dasdan, A.	2014	Conf Knowledge Discovery and Data Mining (SIGKDD)	WoS	22. Sep	130			1		OK	OK
377	2	NEURAL NETWORKS IN	HALGAMUGE, SK	1994	FUZZY SETS AND	WoS	22. Sep	130	22. Sep	off topic				
378	2	The Multiple Attribution	Jordan, P (Jordan,	2011	ALGORITHMIC	WoS	22. Sep	130					OK	OK
379	2	Attribution of	Kitts, B.; Wei, L.;	2010	Conf Data	WoS	22. Sep	130	23. Sep	off topic				
380	2	Death of 'Last Click	Lee, G.	2010	J. Direct, Data	WoS	22. Sep	130	23. Sep	off topic				
381	2	Attributing Conversions	Li, H.; Kannan, P. K	2013	J. Mark. Res	WoS	22. Sep	130	22. Sep	Duplicate				
382	2	Optimizing Multi-	Miguel, A.; Lemos,	2015	Catolica-Lisbon	WoS	22. Sep	130					STOP	STOP
383	2	Planning marketing-mix	Naik, PA (Naik, PA);	2005	MARKETING	WoS	22. Sep	130	23. Sep	off topic				
384	2	The Economic Value of Cl	Nottorf, F.; Funk, B.	2013	Conf	WoS	22. Sep	130					STOP	STOP
385	2	A Course in Game	Osborne, M.J.;	1994	MIT Press,	WoS	22. Sep	130	22. Sep	off topic				



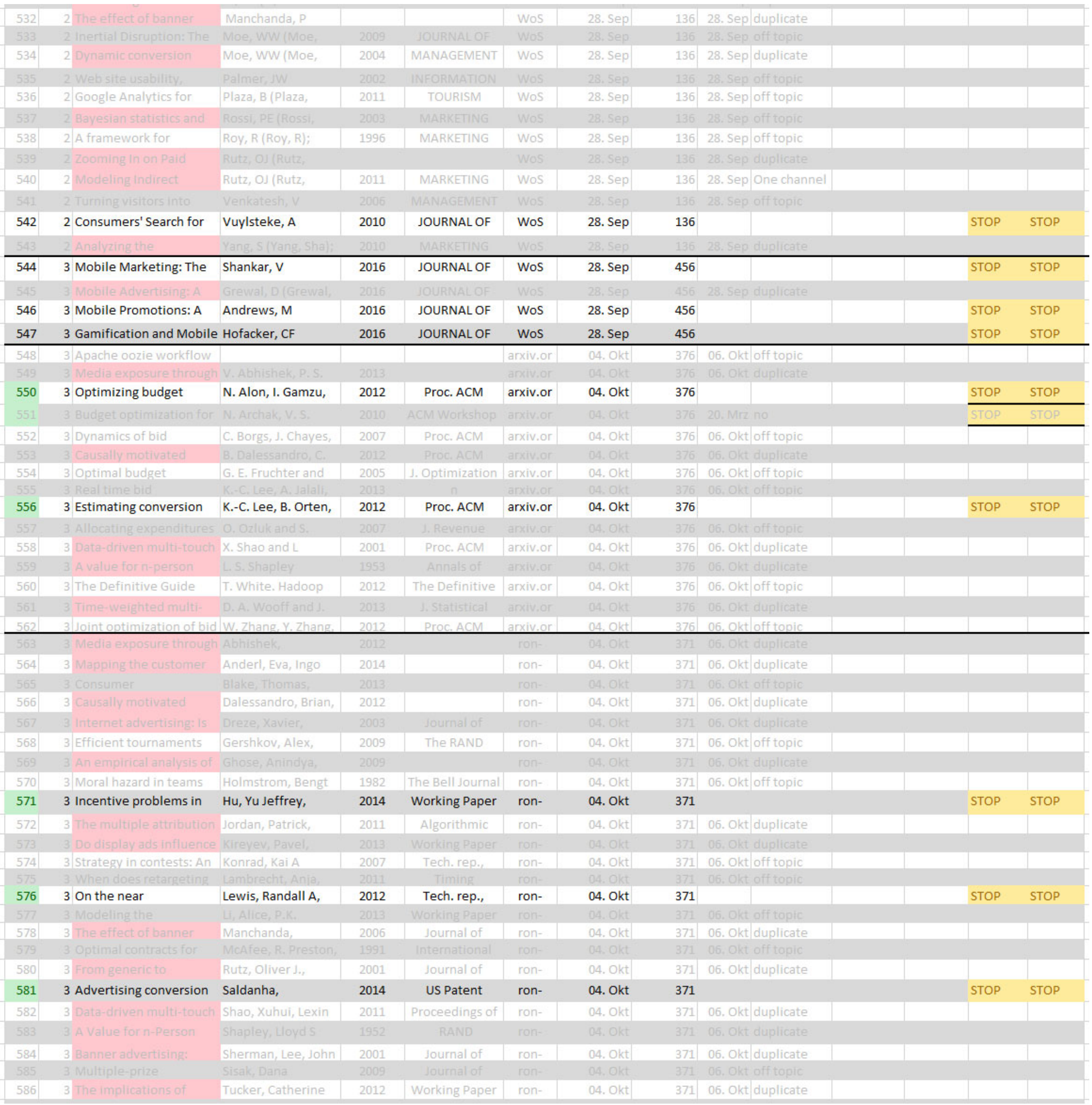
386	2	Machine learning for targ	Perlich, C (Perlich,	2014	MACHINE	WoS	22. Sep	130						STOP	STOP
387	2	Choosing the Right	Petersen, JA	2009	JOURNAL OF	WoS	22. Sep	130	22. Sep	Duplicate					
388	2	Audience Selection for	Provost, F (Provost,	2009	KDD-09: 15TH	WoS	22. Sep	130	22. Sep	off topic					
389	2	Is Marketing Academia	Reibstein, DJ	2009	JOURNAL OF	WoS	22. Sep	130	22. Sep	off topic					
390	2	Analyses of Online	Rentola, O.	2014	University of	WoS	22. Sep	130						OK	STOP
391	2	ESTIMATING CAUSAL	RUBIN, DB (RUBIN,	1974	JOURNAL OF	WoS	22. Sep	130	22. Sep	off topic					
392	2	Data-driven Multi-touch Attribution Models	Shao, X.; Li, L	2011	Conf Knowledge Discovery and Data Mining (SIGKDD)	WoS	22. Sep	130				1		OK	OK
393	2	Banner Advertising:	Sherman, Lee;	2001	Journal of	WoS	22. Sep	130	22. Sep	One channel					
394	2	The Implications of	Tucker, C.	2012	Compet. Online	WoS	22. Sep	130						STOP	STOP
395	2	Multichannel shopping:	Venkatesan, R	2007	JOURNAL OF	WoS	22. Sep	130						STOP	STOP
396	2	Marketing's Profit	Wiesel, Thorsten;	2010	Marketing	WoS	22. Sep	130	22. Sep	Duplicate					
397	2	Time-weighted Multi-	Wooff, D. A.;	2015	J. Stat. Theory	WoS	22. Sep	130						OK	OK
398	2	Path to Purchase: A	Xu, LZ (Xu, Lizhen);	2014	MANAGEMENT	WoS	22. Sep	130	22. Sep	Duplicate					
399	2	Measuring Multi-	Zantedeschi, D.;			WoS	22. Sep	130	22. Sep	Duplicate					
400	2	Multi-touch Attribution in Online Advertising with Survival Theory	Zhang, Y.; Wei, Y.; Ren, J.	2014	Conference: Conf Data Mining (ICDM)	WoS	22. Sep	130				1		OK	OK
401	2	Marketing Attribution Comes of Age	criteo	2013		WoS	22. Sep	130							
402	2	Finding the Right	Liu, Y.; Pandey, S.;	2012	Conf Web	WoS	22. Sep	130	22. Sep	One channel					
403	2	Mapping the customer journey Lessons learned from graph-based online attribution modeling	Anderl, Eva; Becker, Ingo; Wangenheim, Florian von; Schumann, Jan Hendrik	2016	International Journal of Research in Marketing	WoS	27. Sep	370				1		OK	OK
404	2	Optimal bidding in multi-	V. Abhishek and K.	2013	Operations	WoS	27. Sep	143	27. Sep	off topic					
405	2	Aggregation bias in	V. Abhishek, K.	2015	Marketing	WoS	27. Sep	143	27. Sep	off topic					
406	2	Putting attribution to	E. Andrel, I. Becker,	2013	SSRN	WoS	27. Sep	143	27. Sep	off topi					
407	2	A joint model of usage	E. Ascarza and B. G.	2013	Marketing	WoS	27. Sep	143	27. Sep	off topic					
408	2	The development of the	T. E. Barry.	1987	Current Issues	WoS	27. Sep	143	27. Sep	off topic					
409	2	Beyond the Last Touch:	R. Berman.	2013	Working paper,	WoS	27. Sep	143	27. Sep	Duplicate					
410	2	Constructive consumer	J. R. Bettman, M. F.	1998	Journal of	WoS	27. Sep	143	27. Sep	off topic					
411	2	Discovering how	N. I. Bruce, K.	2012	Journal of	WoS	27. Sep	143	27. Sep	off topic					
412	2	Measuring roi beyond	Microsoft	2009		WoS	27. Sep	143	27. Sep	off topic					
413	2	Facebook's advertising	T. Claburn.	2012	Information	WoS	27. Sep	143	27. Sep	off topic					
414	2	The analysis of hospital	B. Cooper and M.	2004	Biostatistics,	WoS	27. Sep	143	27. Sep	off topic					
415	2	The consumer decision	D. Court, D. Elzinga,	2009	McKinsey	WoS	27. Sep	143	27. Sep	Duplicate					
416	2	Causally Motivated	Dalessandro, O.	2012	Proceedings of	WoS	27. Sep	143	27. Sep	Duplicate					
417	2	Online display ads: The	G. de Vries	2012	Forbes	WoS	27. Sep	143	27. Sep	off topic					
418	2	Towards a digital	A. Ghose and V.	2015	MIS Quarterly,	WoS	27. Sep	143						STOP	STOP
419	2	Online display	A. Goldfarb and C.	2011	Marketing	WoS	27. Sep	143	27. Sep	off topic					
420	2	The theory of buyer	J. A. Howard and J.	1969	Wiley	WoS	27. Sep	143	27. Sep	off topic					
421	2	Decomposing the	Y. Hu, R. Y. Du, and	2014	Journal of	WoS	27. Sep	143	28. Sep	off topic					
422	2	Bidding on the buying	B. J. Jansen and S.	2011	Journal of	WoS	27. Sep	143						STOP	STOP
423	2	Mcmc and the label	A. Jasra, C. C.	2005	Statistical	WoS	27. Sep	143	27. Sep	off topic					
424	2	Targeted advertising	J. P. Johnson.	2011	Cornell	WoS	27. Sep	143	27. Sep	off topic					
425	2	The Multiple Attribution	P. Jordan, M.	2011	Proceedings of	WoS	27. Sep	143	27. Sep	Duplicate					
426	2	Multi-Channel	A. Kaushik.	2012		WoS	27. Sep	143						STOP	STOP
427	2	Untangling the	F. Khatibloo	2010	Forrester	WoS	27. Sep	143	27. Sep	off topic					
428	2	Principles of Marketing.	P. Kotler and G.	2011	Prentice Hall,	WoS	27. Sep	143	27. Sep	off topic					



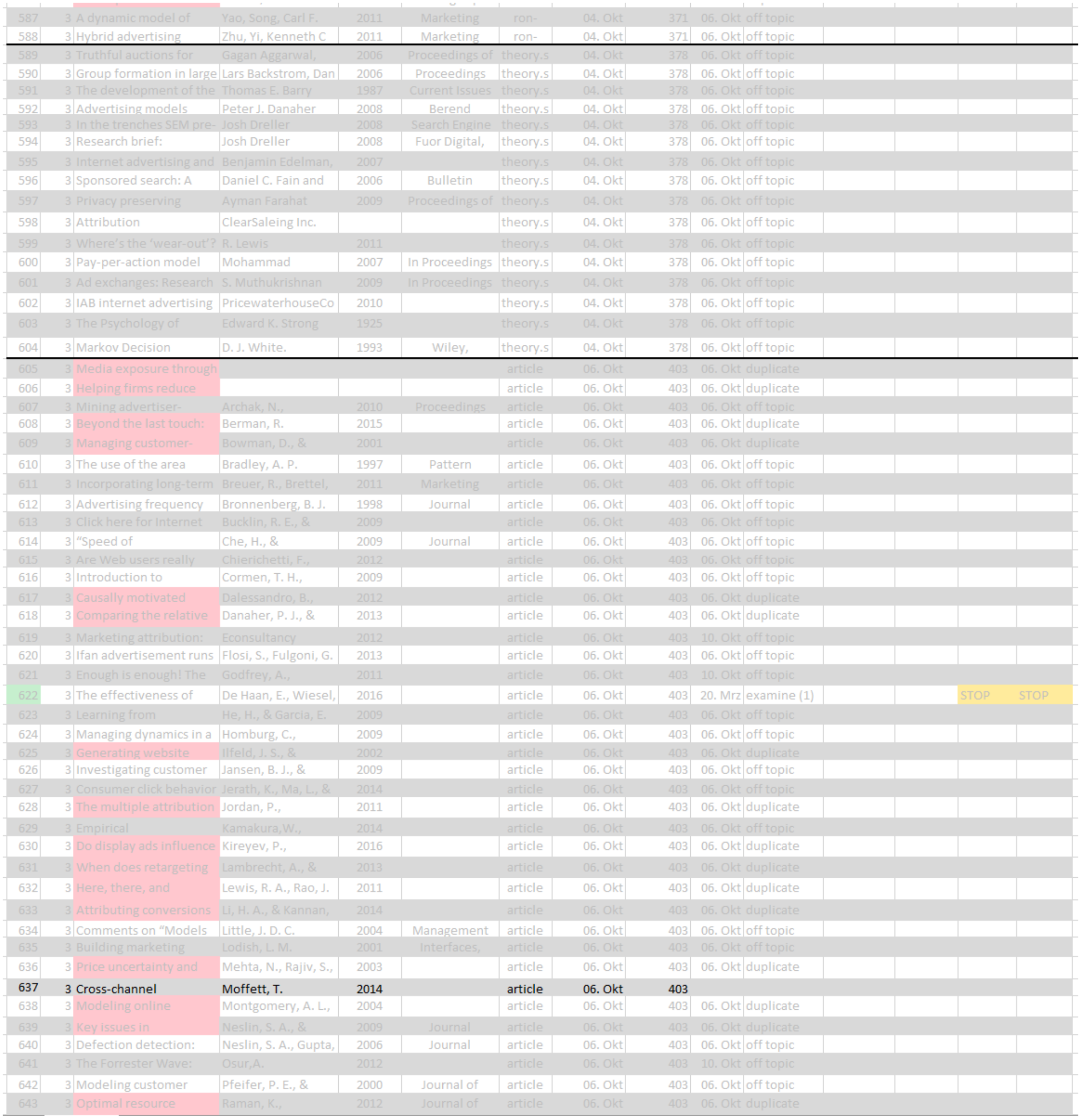
429	2	Attributing Conversions	H. Li and P. Kannan.	2014	Journal of	WoS	27. Sep	143	27. Sep	Duplicate				
430	2	Hidden Markov and	I. L. McDonald and	1997	Chapman and	WoS	27. Sep	143	27. Sep	off topic				
431	2	Dynamic allocation of	R. Montoya, O.	2010	Marketing	WoS	27. Sep	143	27. Sep	off topic				
432	2	The purchase path of	S. Mulpuru	2011	Forrester	WoS	27. Sep	143	27. Sep	off topic				
433	2	Planning media	P. A. Naik, M. K.	1998	Marketing	WoS	27. Sep	143	27. Sep	off topic				
434	2	Big data and marketing	H. Nair, S. Misra, W.	2014	Stanford	WoS	27. Sep	143	27. Sep	off topic				
435	2	Optimal advertising	M. Nerlove and K. J.	1962	Economica,	WoS	27. Sep	143	27. Sep	off topic				
436	2	A hidden markov model	O. Netzer, J. M.	2008	Marketing	WoS	27. Sep	143	27. Sep	off topic				
437	2	Teradata, lunexa	New York Times.	2012		WoS	27. Sep	143	27. Sep	off topi				
438	2	Finding deeper insight	C. Quinn.	2012		WoS	27. Sep	143	27. Sep	off topic				
439	2	Even the rich can make	P. E. Rossi.	2014		WoS	27. Sep	143	27. Sep	off topic				
440	2	From generic to	O. J. Rutz and R. E.	2011	Journal of	WoS	27. Sep	143	27. Sep	duplicate				
441	2	A latent instrumental	O. J. Rutz, R. E.	2012	Journal of	WoS	27. Sep	143					STOP	STOP
442	2	Em versus markov chain	T. Ryden	2008	Bayesian	WoS	27. Sep	143	27. Sep	off topic				
443	2	Effect of temporal	N. S. Sahni.	2015	Stanford	WoS	27. Sep	143	27. Sep	off topic				
444	2	Children of the hmm:	E. M. Schwartz, E.	2011	WCAI Working	WoS	27. Sep	143	27. Sep	off topic				
445	2	Portfolio dynamics for	D. A. Schweidel, E.	2011	Management	WoS	27. Sep	143	27. Sep	off topic				
446	2	Data-driven multi-touch	X. Shao and L. Li.	2011	KDD'11	WoS	27. Sep	143	27. Sep	duplicate				
447	2	A hidden markov model	P. V. Singh, Y. Tan,	2011	Information	WoS	27. Sep	143	27. Sep	off topic				
448	2	The Psychology of	E. K. Strong	1925	McGraw-Hill,	WoS	27. Sep	143	27. Sep	off topic				
449	2	Why you should care	C. Szulc	2012	Inc.	WoS	27. Sep	143					STOP	STOP
450	2	The implications of	C. Tucker.	2013	George Mason	WoS	27. Sep	143	27. Sep	off topic				
451	2	Marketing's profit	T. Wiesel, K.	2011	Marketing	WoS	27. Sep	143	04. Jan	duplicate				
452	2	Path to purchase: A	L. Xu, J. A. Duan,	2014	Management	WoS	27. Sep	143	27. Sep	duplicate				
453	2	The effects of	Y. Yi.	1990	Journal of	WoS	27. Sep	143	27. Sep	off topic				
454	2	Measuring multichannel	D. Zantedeschi, E.	2015	The Wharton	WoS	27. Sep	143						
455	2	Paths to and off	Srinivasan, S	2016	JOURNAL OF	WoS	27. Sep	281	27. Sep	Duplicate				
456	2	Mobile Shopper	Shankar, V	2016	JOURNAL OF	WoS	27. Sep	281					OK	STOP
457	2	Strategic and	Vernuccio, M	2015	EUROPEAN	WoS	27. Sep	281	07. Okt	off topic				
458	2	Cross-Platform	Neijens, P (Neijens,	2015	JOURNAL OF	WoS	27. Sep	281					STOP	STOP
459	2	How to Use	Klapdor, S (Klapdor,	2015	JOURNAL OF	WoS	27. Sep	281	27. Sep	Duplicate				
460	2	The cross-platform	Lim, JS (Lim, Joon	2015	COMPUTERS IN	WoS	27. Sep	281					STOP	Stop
461	2	The Impact of Different	Baxendale, S	2015	JOURNAL OF	WoS	27. Sep	281	27. Sep	off topic				
462	2	What Drives Advertising	Brettel, M (Brettel,	2015	JOURNAL OF	WoS	27. Sep	281	27. Sep	off topic				
463	2	Television Advertising	Liaukonyte, J	2015	MARKETING	WoS	27. Sep	281					STOP	Stop
464	2	Media channels and	Woo, J (Woo,	2015	INDUSTRIAL	WoS	27. Sep	281					STOP	Stop
465	2	The importance of the	Ayeb, S (Ayeb,	2015	INNOVATION	WoS	27. Sep	281	27. Sep	off topic				
466	2	Integrated Online	Janoscik, V	2015	PROCEEDINGS	WoS	27. Sep	281	07. Okt	off topic				
467	2	Examining search as	Micu, AC (Micu,	2015	INTERNET	WoS	27. Sep	281	07. Okt	off topic				
468	2	Billboard and cinema	Frison, S (Frison,	2014	INTERNATIONA	WoS	27. Sep	281	27. Sep	off topic				
469	2	The effect of new media	Woo, J (Woo,	2014	TECHNOLOGICA	WoS	27. Sep	281	27. Sep	off topic				
470	2	Driving Online and	Dinner, IM (Dinner,	2014	JOURNAL OF	WoS	27. Sep	281	20. Mrz	not			STOP	Stop
471	2	Targeted Advertising in	Chandra, A	2014	MANAGEMENT	WoS	27. Sep	281	27. Sep	off topic				
472	2	EFFECTIVENESS OF	Faletra, M (Faletra,	2014	ADVANCES IN	WoS	27. Sep	281	27. Sep	off topic				
473	2	Search Engine	Zenetti, G (Zenetti,	2014	INTERNATIONA	WoS	27. Sep	281	27. Sep	One channel				
474	2	Multi-channel Attribution Modeling on User Journeys	Nottorf, F (Nottorf, Florian)	2014	E-BUSINESS AND TELECOMMUNICATIONS, ICETE 2013	WoS	27. Sep	281			1		OK	OK
475	2	Multichannel	Olbrich, R (Olbrich,	2014	EUROPEAN	WoS	27. Sep	281	27. Sep	Duplicate				
476	2	Modeling the	Nottorf, F (Nottorf,	2014	ELECTRONIC	WoS	27. Sep	281					OK	STOP
477	2	Television Advertising	Joo, M (Joo,	2014	MANAGEMENT	WoS	27. Sep	281					STOP	Stop



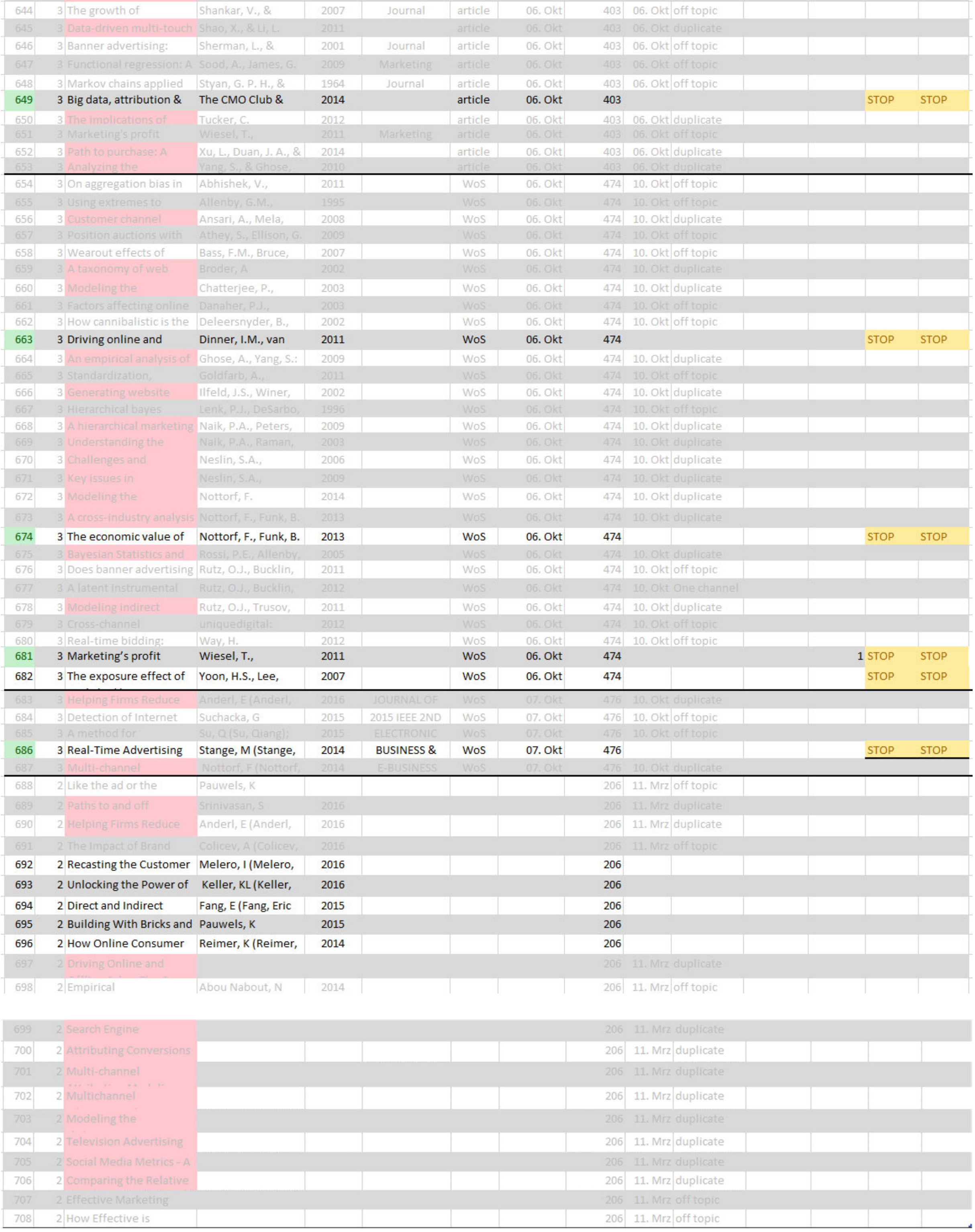
478	2	Scared Stiff? The	Krisjanous, J	2013	PSYCHOLOGY &	WoS	27. Sep	281	27. Sep	off topic				
479	2	Comparing the Relative	Danaher, PJ	2013	JOURNAL OF	WoS	27. Sep	281	27. Sep	off topic				
480	2	Fusing Aggregate and	Feit, EM (Feit,	2013	JOURNAL OF	WoS	27. Sep	281					STOP	Stop
481	2	What Works Best When	Varan, D (Varan,	2013	JOURNAL OF	WoS	27. Sep	281				1	STOP	Stop
482	2	The effects of mailing	Feld, S (Feld,	2013	INTERNATIONA	WoS	27. Sep	281	27. Sep	off topic				
483	2	The Geometric Law of	Malthouse, EC	2013	JOURNAL OF	WoS	27. Sep	281	27. Sep	off topic				
484	2	Singlemedium- versus	Overmars, SWM	2013	TIJDSCHRIFT	WoS	27. Sep	281	27. Sep	off topic				
485	2	Exploring interaction:	Graham, G	2013	INTERNET	WoS	27. Sep	281	27. Sep	off topic				
486	2	Optimal selection of	Malthouse, EC	2012	EXPERT	WoS	27. Sep	281	27. Sep	off topic				
487	2	Marketing activity,	Onishi, H (Onishi,	2012	INTERNATIONA	WoS	27. Sep	281	27. Sep	off topic				
488	2	Optimal Resource	Raman, K (Raman,	2012	JOURNAL OF	WoS	27. Sep	281					STOP	Stop
489	2	Incorporating long-term	Breuer, R (Breuer,	2011	MARKETING	WoS	27. Sep	281	27. Sep	Duplicate				
490	2	Media multitasking and	Voorveld, HAM	2011	COMPUTERS IN	WoS	27. Sep	281					STOP	Stop
491	2	Using several	Sorato, A (Sorato,	2011	OPTIMIZATION	WoS	27. Sep	281	27. Sep	off topic				
492	2	Innovations in Shopper	Shankar, V	2011	JOURNAL OF	WoS	27. Sep	281	27. Sep	off topic				
493	2	Marketing's Profit	Wiesel, T (Wiesel,	2011	MARKETING	WoS	27. Sep	281	27. Sep	Duplicate				
494	2	The Impact of New	Hennig-Thurau, T	2010	JOURNAL OF	WoS	27. Sep	281	27. Sep	off topic				
495	2	Mobile Marketing in the	Shankar, V	2010	JOURNAL OF	WoS	27. Sep	281	27. Sep	off topic				
496	2	The Growing Influence	Shankar, V	2009	JOURNAL OF	WoS	27. Sep	281	27. Sep	off topic				
497	2	Evaluating online ad	Proceedings of the	2010		Scopus	28. Sep	375	28. Sep	off topic				
498	2	Here, there, and		2011	Proceedings of	Scopus	28. Sep	375	28. Sep	off topic				
499	2	Data-driven multi-touch			Proceedings of	Scopus	28. Sep	375	28. Sep	duplicate				
500	2	The Long Road to Online	Abhishek, V.,	2012		Scopus	28. Sep	397	28. Sep	Duplicate				
501	2	Causally motivated	Dalessandro, B.,			Scopus	28. Sep		28. Sep	Duplicate				
502	2	A Web page prediction		2003		Scopus	28. Sep	397	28. Sep	on site				
503	2	Incremental click-		2006		Scopus	28. Sep	397	28. Sep	on site				
504	2	A family of				Scopus	28. Sep	397	28. Sep	off topic				
505	2	Buying, searching, or		2003	Journal of	Scopus	28. Sep	397	28. Sep	offline				
506	2	Data-driven multi-touch				Scopus	28. Sep	397	28. Sep	duplicate				
507	2	Robust and scale-free		2013	Journal of	Scopus	28. Sep	397	28. Sep	off topic				
508	2	Path to Purchase: A		2012		Scopus	28. Sep	397	28. Sep	duplicate				
509	3	Media Exposure				Scopus	28. Sep	400	28. Sep	Duplicate				
510	3	Does customization	Bright, L.F.,	2012	Journal of	Scopus	28. Sep	400	28. Sep	off topic				
511	3	Causally motivated	Dalessandro, B.,			Scopus	28. Sep	400	28. Sep	Duplicate				
512	3	MapReduce: Simplified	Dean, J.,	2008		Scopus	28. Sep	400	28. Sep	off topic				
513	3	A model for predictive	Lavidge, R.J.,	1961	Journal of	Scopus	28. Sep	400					Stop	Stop
514	3	Statistical Models and	Lawless, J.	2011	Wiley Series in	Scopus	28. Sep	400	28. Sep	off topic				
515	3	Data-driven multi-touch				Scopus	28. Sep	400	28. Sep	Duplicate				
516	3	The influence of	Ueltschy, L.C.,	2011	Journal of	Scopus	28. Sep	400	28. Sep	off topic				
517	3	Path to Purchase: A	Xu, L., Duan, J.A.,	2012		Scopus	28. Sep	400	28. Sep	Duplicate				
518	3	Search engine	Zenetti, G., Bijmolt,	2014		Scopus	28. Sep	400	28. Sep	Duplicate				
519	3	Happy Birthday, Digital				Scopus	28. Sep	392	28. Sep	off topic				
520	3	Audience selection for	Provost, F.,			Scopus	28. Sep	392	28. Sep	Duplicate				
521	3	Exploitation and	Li, W., Wang, X.,			Scopus	28. Sep	392	28. Sep	off topic				
522	3	The Elements of	Hastie, T.,			Scopus	28. Sep	392	28. Sep	off topic				
523	3	Sensitive webpage	Jin, X., Li, Y., Mah,	2007		Scopus	28. Sep	392	28. Sep	off topic				
524	3	Neural Networks for	Bishop, C.M.			Scopus	28. Sep	392	28. Sep	off topic				
525	3	Pattern Recognition and	Bishop, C.M.			Scopus	28. Sep	392	28. Sep	off topic				
526	3	Bagging predictors				Scopus	28. Sep	392	28. Sep	Duplicate				
527	3	Tree induction vs.				Scopus	28. Sep	392	28. Sep	off topic				
528	2	Short- and Long-term	Breuer, R (Breuer,	2012		WoS	28. Sep	136	28. Sep	duplicate				
529	2	Markov chain Monte	Cowles, MK	1996	JOURNAL OF	WoS	28. Sep	136	28. Sep	off topic				
530	2	Searching for	Huang, P (Huang,	2009	JOURNAL OF	WoS	28. Sep	136	28. Sep	off topic				
531	2	Attributing Conversions	Li, HS (Li,	2014		WoS	28. Sep	136	28. Sep	duplicate				



532	2	The effect of banner	Manchanda, P			WoS	28. Sep	136	28. Sep	duplicate				
533	2	Inertial Disruption: The	Moe, WW (Moe,	2009	JOURNAL OF	WoS	28. Sep	136	28. Sep	off topic				
534	2	Dynamic conversion	Moe, WW (Moe,	2004	MANAGEMENT	WoS	28. Sep	136	28. Sep	duplicate				
535	2	Web site usability,	Palmer, JW	2002	INFORMATION	WoS	28. Sep	136	28. Sep	off topic				
536	2	Google Analytics for	Plaza, B (Plaza,	2011	TOURISM	WoS	28. Sep	136	28. Sep	off topic				
537	2	Bayesian statistics and	Rossi, PE (Rossi,	2003	MARKETING	WoS	28. Sep	136	28. Sep	off topic				
538	2	A framework for	Roy, R (Roy, R);	1996	MARKETING	WoS	28. Sep	136	28. Sep	off topic				
539	2	Zooming In on Paid	Rutz, OJ (Rutz,			WoS	28. Sep	136	28. Sep	duplicate				
540	2	Modeling Indirect	Rutz, OJ (Rutz,	2011	MARKETING	WoS	28. Sep	136	28. Sep	One channel				
541	2	Turning visitors into	Venkatesh, V	2006	MANAGEMENT	WoS	28. Sep	136	28. Sep	off topic				
542	2	Consumers' Search for	Vuylsteke, A	2010	JOURNAL OF	WoS	28. Sep	136					STOP	STOP
543	2	Analyzing the	Yang, S (Yang, Sha);	2010	MARKETING	WoS	28. Sep	136	28. Sep	duplicate				
544	3	Mobile Marketing: The	Shankar, V	2016	JOURNAL OF	WoS	28. Sep	456					STOP	STOP
545	3	Mobile Advertising: A	Grewal, D (Grewal,	2016	JOURNAL OF	WoS	28. Sep	456	28. Sep	duplicate				
546	3	Mobile Promotions: A	Andrews, M	2016	JOURNAL OF	WoS	28. Sep	456					STOP	STOP
547	3	Gamification and Mobile	Hofacker, CF	2016	JOURNAL OF	WoS	28. Sep	456					STOP	STOP
548	3	Apache oozie workflow				arxiv.or	04. Okt	376	06. Okt	off topic				
549	3	Media exposure through	V. Abhishek, P. S.	2013		arxiv.or	04. Okt	376	06. Okt	duplicate				
550	3	Optimizing budget	N. Alon, I. Gamzu,	2012	Proc. ACM	arxiv.or	04. Okt	376					STOP	STOP
551	3	Budget optimization for	N. Archak, V. S.	2010	ACM Workshop	arxiv.or	04. Okt	376	20. Mrz	no			STOP	STOP
552	3	Dynamics of bid	C. Borgs, J. Chayes,	2007	Proc. ACM	arxiv.or	04. Okt	376	06. Okt	off topic				
553	3	Causally motivated	B. Dalessandro, C.	2012	Proc. ACM	arxiv.or	04. Okt	376	06. Okt	duplicate				
554	3	Optimal budget	G. E. Fruchter and	2005	J. Optimization	arxiv.or	04. Okt	376	06. Okt	off topic				
555	3	Real time bid	K.-C. Lee, A. Jalali,	2013	n	arxiv.or	04. Okt	376	06. Okt	off topic				
556	3	Estimating conversion	K.-C. Lee, B. Orten,	2012	Proc. ACM	arxiv.or	04. Okt	376					STOP	STOP
557	3	Allocating expenditures	O. Ozluk and S.	2007	J. Revenue	arxiv.or	04. Okt	376	06. Okt	off topic				
558	3	Data-driven multi-touch	X. Shao and L	2001	Proc. ACM	arxiv.or	04. Okt	376	06. Okt	duplicate				
559	3	A value for n-person	L. S. Shapley	1953	Annals of	arxiv.or	04. Okt	376	06. Okt	duplicate				
560	3	The Definitive Guide	T. White. Hadoop	2012	The Definitive	arxiv.or	04. Okt	376	06. Okt	off topic				
561	3	Time-weighted multi-	D. A. Wooff and J.	2013	J. Statistical	arxiv.or	04. Okt	376	06. Okt	duplicate				
562	3	Joint optimization of bid	W. Zhang, Y. Zhang,	2012	Proc. ACM	arxiv.or	04. Okt	376	06. Okt	off topic				
563	3	Media exposure through	Abhishek,	2012		ron-	04. Okt	371	06. Okt	duplicate				
564	3	Mapping the customer	Anderl, Eva, Ingo	2014		ron-	04. Okt	371	06. Okt	duplicate				
565	3	Consumer	Blake, Thomas,	2013		ron-	04. Okt	371	06. Okt	off topic				
566	3	Causally motivated	Dalessandro, Brian,	2012		ron-	04. Okt	371	06. Okt	duplicate				
567	3	Internet advertising: Is	Dreze, Xavier,	2003	Journal of	ron-	04. Okt	371	06. Okt	duplicate				
568	3	Efficient tournaments	Gershkov, Alex,	2009	The RAND	ron-	04. Okt	371	06. Okt	off topic				
569	3	An empirical analysis of	Ghose, Anindya,	2009		ron-	04. Okt	371	06. Okt	duplicate				
570	3	Moral hazard in teams	Holmstrom, Bengt	1982	The Bell Journal	ron-	04. Okt	371	06. Okt	off topic				
571	3	Incentive problems in	Hu, Yu Jeffrey,	2014	Working Paper	ron-	04. Okt	371					STOP	STOP
572	3	The multiple attribution	Jordan, Patrick,	2011	Algorithmic	ron-	04. Okt	371	06. Okt	duplicate				
573	3	Do display ads influence	Kireyev, Pavel,	2013	Working Paper	ron-	04. Okt	371	06. Okt	duplicate				
574	3	Strategy in contests: An	Konrad, Kai A	2007	Tech. rep.,	ron-	04. Okt	371	06. Okt	off topic				
575	3	When does retargeting	Lambrecht, Anja,	2011	Timing	ron-	04. Okt	371	06. Okt	off topic				
576	3	On the near	Lewis, Randall A,	2012	Tech. rep.,	ron-	04. Okt	371					STOP	STOP
577	3	Modeling the	Li, Alice, P.K.	2013	Working Paper	ron-	04. Okt	371	06. Okt	off topic				
578	3	The effect of banner	Manchanda,	2006	Journal of	ron-	04. Okt	371	06. Okt	duplicate				
579	3	Optimal contracts for	McAfee, R. Preston,	1991	International	ron-	04. Okt	371	06. Okt	off topic				
580	3	From generic to	Rutz, Oliver J.,	2001	Journal of	ron-	04. Okt	371	06. Okt	duplicate				
581	3	Advertising conversion	Saldanha,	2014	US Patent	ron-	04. Okt	371					STOP	STOP
582	3	Data-driven multi-touch	Shao, Xuhui, Lexin	2011	Proceedings of	ron-	04. Okt	371	06. Okt	duplicate				
583	3	A Value for n-Person	Shapley, Lloyd S	1952	RAND	ron-	04. Okt	371	06. Okt	duplicate				
584	3	Banner advertising:	Sherman, Lee, John	2001	Journal of	ron-	04. Okt	371	06. Okt	duplicate				
585	3	Multiple-prize	Sisak, Dana	2009	Journal of	ron-	04. Okt	371	06. Okt	off topic				
586	3	The implications of	Tucker, Catherine	2012	Working Paper	ron-	04. Okt	371	06. Okt	duplicate				



587	3	A dynamic model of	Yao, Song, Carl F.	2011	Marketing	ron-	04. Okt	371	06. Okt	off topic			
588	3	Hybrid advertising	Zhu, Yi, Kenneth C	2011	Marketing	ron-	04. Okt	371	06. Okt	off topic			
589	3	Truthful auctions for	Gagan Aggarwal,	2006	Proceedings of	theory.s	04. Okt	378	06. Okt	off topic			
590	3	Group formation in large	Lars Backstrom, Dan	2006	Proceedings	theory.s	04. Okt	378	06. Okt	off topic			
591	3	The development of the	Thomas E. Barry	1987	Current Issues	theory.s	04. Okt	378	06. Okt	off topic			
592	3	Advertising models	Peter J. Danaher	2008	Berend	theory.s	04. Okt	378	06. Okt	off topic			
593	3	In the trenches SEM pre-	Josh Dreller	2008	Search Engine	theory.s	04. Okt	378	06. Okt	off topic			
594	3	Research brief:	Josh Dreller	2008	Fuor Digital,	theory.s	04. Okt	378	06. Okt	off topic			
595	3	Internet advertising and	Benjamin Edelman,	2007		theory.s	04. Okt	378	06. Okt	off topic			
596	3	Sponsored search: A	Daniel C. Fain and	2006	Bulletin	theory.s	04. Okt	378	06. Okt	off topic			
597	3	Privacy preserving	Ayman Farahat	2009	Proceedings of	theory.s	04. Okt	378	06. Okt	off topic			
598	3	Attribution	ClearSaleing Inc.			theory.s	04. Okt	378	06. Okt	off topic			
599	3	Where's the 'wear-out'?	R. Lewis	2011		theory.s	04. Okt	378	06. Okt	off topic			
600	3	Pay-per-action model	Mohammad	2007	In Proceedings	theory.s	04. Okt	378	06. Okt	off topic			
601	3	Ad exchanges: Research	S. Muthukrishnan	2009	In Proceedings	theory.s	04. Okt	378	06. Okt	off topic			
602	3	IAB internet advertising	PricewaterhouseCo	2010		theory.s	04. Okt	378	06. Okt	off topic			
603	3	The Psychology of	Edward K. Strong	1925		theory.s	04. Okt	378	06. Okt	off topic			
604	3	Markov Decision	D. J. White.	1993	Wiley,	theory.s	04. Okt	378	06. Okt	off topic			
605	3	Media exposure through				article	06. Okt	403	06. Okt	duplicate			
606	3	Helping firms reduce				article	06. Okt	403	06. Okt	duplicate			
607	3	Mining advertiser-	Archak, N.,	2010	Proceedings	article	06. Okt	403	06. Okt	off topic			
608	3	Beyond the last touch:	Berman, R.	2015		article	06. Okt	403	06. Okt	duplicate			
609	3	Managing customer-	Bowman, D., &	2001		article	06. Okt	403	06. Okt	duplicate			
610	3	The use of the area	Bradley, A. P.	1997	Pattern	article	06. Okt	403	06. Okt	off topic			
611	3	Incorporating long-term	Breuer, R., Brettel,	2011	Marketing	article	06. Okt	403	06. Okt	off topic			
612	3	Advertising frequency	Bronnenberg, B. J.	1998	Journal	article	06. Okt	403	06. Okt	off topic			
613	3	Click here for Internet	Bucklin, R. E., &	2009		article	06. Okt	403	06. Okt	off topic			
614	3	"Speed of	Che, H., &	2009	Journal	article	06. Okt	403	06. Okt	off topic			
615	3	Are Web users really	Chierichetti, F.,	2012		article	06. Okt	403	06. Okt	off topic			
616	3	Introduction to	Cormen, T. H.,	2009		article	06. Okt	403	06. Okt	off topic			
617	3	Causally motivated	Dalessandro, B.,	2012		article	06. Okt	403	06. Okt	duplicate			
618	3	Comparing the relative	Danaher, P. J., &	2013		article	06. Okt	403	06. Okt	duplicate			
619	3	Marketing attribution:	Econsultancy	2012		article	06. Okt	403	10. Okt	off topic			
620	3	Ifan advertisement runs	Flosi, S., Fulgoni, G.	2013		article	06. Okt	403	06. Okt	off topic			
621	3	Enough is enough! The	Godfrey, A.,	2011		article	06. Okt	403	10. Okt	off topic			
622	3	The effectiveness of	De Haan, E., Wiesel,	2016		article	06. Okt	403	20. Mrz	examine (1)		STOP	STOP
623	3	Learning from	He, H., & Garcia, E.	2009		article	06. Okt	403	06. Okt	off topic			
624	3	Managing dynamics in a	Homburg, C.,	2009		article	06. Okt	403	06. Okt	off topic			
625	3	Generating website	Ilfeld, J. S., &	2002		article	06. Okt	403	06. Okt	duplicate			
626	3	Investigating customer	Jansen, B. J., &	2009		article	06. Okt	403	06. Okt	off topic			
627	3	Consumer click behavior	Jerath, K., Ma, L., &	2014		article	06. Okt	403	06. Okt	off topic			
628	3	The multiple attribution	Jordan, P.,	2011		article	06. Okt	403	06. Okt	duplicate			
629	3	Empirical	Kamakura,W.,	2014		article	06. Okt	403	06. Okt	off topic			
630	3	Do display ads influence	Kireyev, P.,	2016		article	06. Okt	403	06. Okt	duplicate			
631	3	When does retargeting	Lambrecht, A., &	2013		article	06. Okt	403	06. Okt	duplicate			
632	3	Here, there, and	Lewis, R. A., Rao, J.	2011		article	06. Okt	403	06. Okt	duplicate			
633	3	Attributing conversions	Li, H. A., & Kannan,	2014		article	06. Okt	403	06. Okt	duplicate			
634	3	Comments on "Models	Little, J. D. C.	2004	Management	article	06. Okt	403	06. Okt	off topic			
635	3	Building marketing	Lodish, L. M.	2001	Interfaces,	article	06. Okt	403	06. Okt	off topic			
636	3	Price uncertainty and	Mehta, N., Rajiv, S.,	2003		article	06. Okt	403	06. Okt	duplicate			
637	3	Cross-channel	Moffett, T.	2014		article	06. Okt	403					
638	3	Modeling online	Montgomery, A. L.,	2004		article	06. Okt	403	06. Okt	duplicate			
639	3	Key issues in	Neslin, S. A., &	2009	Journal	article	06. Okt	403	06. Okt	duplicate			
640	3	Defection detection:	Neslin, S. A., Gupta,	2006	Journal	article	06. Okt	403	06. Okt	off topic			
641	3	The Forrester Wave:	Osur,A.	2012		article	06. Okt	403	10. Okt	off topic			
642	3	Modeling customer	Pfeifer, P. E., &	2000	Journal of	article	06. Okt	403	06. Okt	off topic			
643	3	Optimal resource	Raman, K.,	2012	Journal of	article	06. Okt	403	06. Okt	duplicate			



644	3	The growth of	Shankar, V., &	2007	Journal	article	06. Okt	403	06. Okt	off topic				
645	3	Data-driven multi-touch	Shao, X., & Li, L.	2011		article	06. Okt	403	06. Okt	duplicate				
646	3	Banner advertising:	Sherman, L., &	2001	Journal	article	06. Okt	403	06. Okt	off topic				
647	3	Functional regression: A	Sood, A., James, G.	2009	Marketing	article	06. Okt	403	06. Okt	off topic				
648	3	Markov chains applied	Styan, G. P. H., &	1964	Journal	article	06. Okt	403	06. Okt	off topic				
649	3	Big data, attribution &	The CMO Club &	2014		article	06. Okt	403					STOP	STOP
650	3	The implications of	Tucker, C.	2012		article	06. Okt	403	06. Okt	duplicate				
651	3	Marketing's profit	Wiesel, T.,	2011	Marketing	article	06. Okt	403	06. Okt	off topic				
652	3	Path to purchase: A	Xu, L., Duan, J. A., &	2014		article	06. Okt	403	06. Okt	duplicate				
653	3	Analyzing the	Yang, S., & Ghose,	2010		article	06. Okt	403	06. Okt	duplicate				
654	3	On aggregation bias in	Abhishek, V.,	2011		WoS	06. Okt	474	10. Okt	off topic				
655	3	Using extremes to	Allenby, G.M.,	1995		WoS	06. Okt	474	10. Okt	off topic				
656	3	Customer channel	Ansari, A., Mela,	2008		WoS	06. Okt	474	10. Okt	duplicate				
657	3	Position auctions with	Athey, S., Ellison, G.	2009		WoS	06. Okt	474	10. Okt	off topic				
658	3	Wearout effects of	Bass, F.M., Bruce,	2007		WoS	06. Okt	474	10. Okt	off topic				
659	3	A taxonomy of web	Broder, A	2002		WoS	06. Okt	474	10. Okt	duplicate				
660	3	Modeling the	Chatterjee, P.,	2003		WoS	06. Okt	474	10. Okt	duplicate				
661	3	Factors affecting online	Danaher, P.J.,	2003		WoS	06. Okt	474	10. Okt	off topic				
662	3	How cannibalistic is the	Deleersnyder, B.,	2002		WoS	06. Okt	474	10. Okt	off topic				
663	3	Driving online and	Dinner, I.M., van	2011		WoS	06. Okt	474					STOP	STOP
664	3	An empirical analysis of	Ghose, A., Yang, S.:	2009		WoS	06. Okt	474	10. Okt	duplicate				
665	3	Standardization,	Goldfarb, A.,	2011		WoS	06. Okt	474	10. Okt	off topic				
666	3	Generating website	Ilfeld, J.S., Winer,	2002		WoS	06. Okt	474	10. Okt	duplicate				
667	3	Hierarchical bayes	Lenk, P.J., DeSarbo,	1996		WoS	06. Okt	474	10. Okt	off topic				
668	3	A hierarchical marketing	Naik, P.A., Peters,	2009		WoS	06. Okt	474	10. Okt	duplicate				
669	3	Understanding the	Naik, P.A., Raman,	2003		WoS	06. Okt	474	10. Okt	duplicate				
670	3	Challenges and	Neslin, S.A.,	2006		WoS	06. Okt	474	10. Okt	duplicate				
671	3	Key issues in	Neslin, S.A.,	2009		WoS	06. Okt	474	10. Okt	duplicate				
672	3	Modeling the	Nottorf, F.	2014		WoS	06. Okt	474	10. Okt	duplicate				
673	3	A cross-industry analysis	Nottorf, F., Funk, B.	2013		WoS	06. Okt	474	10. Okt	duplicate				
674	3	The economic value of	Nottorf, F., Funk, B.	2013		WoS	06. Okt	474					STOP	STOP
675	3	Bayesian Statistics and	Rossi, P.E., Allenby,	2005		WoS	06. Okt	474	10. Okt	duplicate				
676	3	Does banner advertising	Rutz, O.J., Bucklin,	2011		WoS	06. Okt	474	10. Okt	off topic				
677	3	A latent instrumental	Rutz, O.J., Bucklin,	2012		WoS	06. Okt	474	10. Okt	One channel				
678	3	Modeling indirect	Rutz, O.J., Trusov,	2011		WoS	06. Okt	474	10. Okt	duplicate				
679	3	Cross-channel	uniquedigital:	2012		WoS	06. Okt	474	10. Okt	off topic				
680	3	Real-time bidding:	Way, H.	2012		WoS	06. Okt	474	10. Okt	off topic				
681	3	Marketing's profit	Wiesel, T.,	2011		WoS	06. Okt	474				1	STOP	STOP
682	3	The exposure effect of	Yoon, H.S., Lee,	2007		WoS	06. Okt	474					STOP	STOP
683	3	Helping Firms Reduce	Anderl, E (Anderl,	2016	JOURNAL OF	WoS	07. Okt	476	10. Okt	duplicate				
684	3	Detection of Internet	Suchacka, G	2015	2015 IEEE 2ND	WoS	07. Okt	476	10. Okt	off topic				
685	3	A method for	Su, Q (Su, Qiang);	2015	ELECTRONIC	WoS	07. Okt	476	10. Okt	off topic				
686	3	Real-Time Advertising	Stange, M (Stange,	2014	BUSINESS &	WoS	07. Okt	476					STOP	STOP
687	3	Multi-channel	Nottorf, F (Nottorf,	2014	E-BUSINESS	WoS	07. Okt	476	10. Okt	duplicate				
688	2	Like the ad or the	Pauwels, K					206	11. Mrz	off topic				
689	2	Paths to and off	Srinivasan, S	2016				206	11. Mrz	duplicate				
690	2	Helping Firms Reduce	Anderl, E (Anderl,	2016				206	11. Mrz	duplicate				
691	2	The Impact of Brand	Colicev, A (Colicev,	2016				206	11. Mrz	off topic				
692	2	Recasting the Customer	Melero, I (Melero,	2016				206						
693	2	Unlocking the Power of	Keller, KL (Keller,	2016				206						
694	2	Direct and Indirect	Fang, E (Fang, Eric	2015				206						
695	2	Building With Bricks and	Pauwels, K	2015				206						
696	2	How Online Consumer	Reimer, K (Reimer,	2014				206						
697	2	Driving Online and						206	11. Mrz	duplicate				
698	2	Empirical	Abou Nabout, N	2014				206	11. Mrz	off topic				
699	2	Search Engine						206	11. Mrz	duplicate				
700	2	Attributing Conversions						206	11. Mrz	duplicate				
701	2	Multi-channel						206	11. Mrz	duplicate				
702	2	Multichannel						206	11. Mrz	duplicate				
703	2	Modeling the						206	11. Mrz	duplicate				
704	2	Television Advertising						206	11. Mrz	duplicate				
705	2	Social Media Metrics - A						206	11. Mrz	duplicate				
706	2	Comparing the Relative						206	11. Mrz	duplicate				
707	2	Effective Marketing						206	11. Mrz	off topic				
708	2	How Effective is						206	11. Mrz	off topic				

Appendix 2: Interview Coding Categories

Data Flow IPO

(1) Input data

Criteria	Statements
Ability to handle input sources containing hard facts Notes: • Tracking-Daten www • Tracking-Daten App • Company internal data sources • External data sources • Channel data • Online and offline data	➔ alle Datenquellen, die der Firma in diesem Fall zur Verfügung stehen [09] ➔ Je mehr desto besser [09] ➔ und dann kommen noch die dazu, die man kauft, je nachdem, was gerade braucht, ja äh, die man extern am Markt zu den eigenen Daten braucht um sie dann zu verwenden [09] ➔ IP-Adressen sein, das können GEO-locations sein, das können Branchen sein [09] ➔ Nein, ausschließen sollte man am Anfang möglichst wenig an sich, natürlich ähm, weil wenn man sich zu sehr einengt, dann kann natürlich auch ähm schnell ein falscher Eindruck des Users irgendwo entstehen, dementsprechend wie auch zielgruppenspezifische Daten mit einzubeziehen weiche Faktoren mit einzubeziehen ist äh natürlich ganze vorne mit dabei was die Prio angeht [04] ➔ sehr viele Datentöpfe zur Verfügung stehen, aber grundlegend um mal ein grundlegendes Modell zu skizzieren ist natürlich ähm so was wie die ABC-Analyse bzw. die ABC-Variable natürlich sehr sehr wichtig. [04] ➔ Das ganze muss man natürlich um zielgruppenspezifische Daten ergänzen, beispielsweise eben halt äh durch demografische Daten – wie z.B. Geschlecht oder eben halt Alter natürlich ähm, spezifische Kampagneninformationen und natürlich spezifische Informationen, die das komplette ähm Produkt noch einmal beschreiben an sich [04] ➔ , tracken den User über die komplette Seite an sich, erfassen natürlich die Kampagnendaten dann mit nun, wie gesagt schon, wie verhält er sich auf dieser Seite oder der App und was kommt hinten letztendlich bei raus. Natürlich haben wir noch andere Tools, sowas wie AdWords natürlich oder eben halt ein mobil campaign tracking tracking was das Ganze natürlich ergänzt bzw. komplementär dazu erweist bzw. sich komplettiert [04] ➔ Event-Daten, Wetter, GEO-Daten, Saison [03] ➔ Kombination, ähm , ganz zentral, eine Kombination aus den Tracking-Daten und, ähm, den den Daten. Du hast sie jetzt hier externe Daten genannt, ja, also, ähm, den dem dem Nutzungsverhalten bspw. aus ähm den verschieden Sourcen wie den verschiedenen social media Kanälen und so weiter, ne [03] ➔ in Zukunft möglich oder nötig, dass man ähm auch Quellen wie offline z.B. mitberücksichtigt, dass man, wenn man einen point of sale hat quasi die Daten auch mit den online-Daten verknüpfen kann [05] ➔ alles, was man so erfassen kann um die Interaktion des Nutzers herum ähm, das kann primär sein seine Interaktion auf der Seite [07] ➔ , die Quelle, wo er gerade herkommt und ähm ähm, welche Produkte ein Nutzer sich ansieht, seine [07] ➔ bevorzugt würde ich gerne interne anwenden und die Daten intern modellieren, aber ähm externe Daten sollten auf jeden Fall auch anwendbar sein. . Ein ähm bestes Beispiel wenn ich wenn ich online auf `ner Seite meine Nutzer sehe, ansprechen möchte ähm wird`s auch wichtig, ob die jetzt offline irgendwo ein Abo abgeschlossen haben, ähm dann könnte ich noch eigene Daten aus aus ner offline Quelle noch mit anbinden [07] ➔ nach Möglichkeit aber auch von anderen Quellen [07] ➔ je mehr ich über den Nutzer weiß, desto mehr habe ich die Chance ihn richtig anzusprechen und ähm etwas zu erstellen, was was ihn halt wirklich anspricht [07] ➔ bevorzugt würde ich gerne interne anwenden und die Daten intern modellieren, aber ähm externe Daten sollten auf jeden Fall auch anwendbar sein. Ein ähm bestes Beispiel wenn ich wenn ich online auf `ner Seite meine Nutzer sehe, ansprechen möchte ähm wird`s auch wichtig, ob die jetzt offline irgendwo ein Abo abgeschlossen haben, ähm dann könnte ich noch eigene Daten aus aus ner offline Quelle noch mit anbinden [07] ➔ Wo er die äh, was er wo guckt, was er möchte, von wo er kommt, dazu könnte man natürlich unsere Objektdaten nehmen, vielleicht als Hintergrundinformation in was für einem Preissegment sucht er, nach was für einer Region, wenn das relevant ist [08] ➔ aber natürlich auch noch externe Daten hinzunehmen, mm, genau vielleicht auch Marktforschungsdaten, aber vielleicht auch weiche Faktoren, die man nicht direkt messen kann [08] ➔ Auch die historischen Daten [08]

Ability to handle input sources containing soft facts Notes: • Interests • Feelings • Attitudes	➔ Saison (z.B. Weihnachten bzw. Indikation, dass es bald passiert) [03] ➔ seine Interessen [04] ➔ Ich denke, das wird immer wichtiger ähm genauso als mit mit ähm mit, zu erfassen, modellieren und ähm man sieht immer mehr das Nutzer äh auch mehr nach Gefühlslage entscheiden und nicht einfach nur nach nachhaken [07] ➔ WEICHE FAKTOREN: für was sich der Nutzer interessiert, für was er vielleicht sensibel ist, wie man ihn ansprechen könnte, was man nicht direkt messen kann [05] ➔ Predictive-Analysen müssen zusätzlich weitere Signale berücksichtigen: Saisonalität, Auktionsdynamik (siehe Input Variables) [08] ➔ Weiche Faktoren so etwas wie Interessen [09] ➔ je mehr ich über den Nutzer weiß, desto mehr habe ich die Chance ihn richtig anzusprechen und ähm etwas zu erstellen, was was ihn halt wirklich anspricht [07]
Ability to add/remove data sources	➔ Einfaches hinzufügen wäre gut fürs testen [03] ➔ Wenn wir einen Anbieter wechseln, ähm muss dieses umgesetzt werden können [07] ➔ Wünschenswert ist natürlich beides: auf der einen Seite möchte ich System haben, wo ich ähm Tools sehr sehr schnell ja anschließen, aber irgendwann auch wieder wegnehmen kann, auf der anderen Seite möchte ich natürlich auch, dass es fehlerfrei läuft, was meistens damit einhergeht, dass man ein bestehendes System nimmt und das immer wieder nachjustiert letztendlich [04] ➔ Cross-Device-Daten: Realisiert mit Cookies und Login, Fingerprinting ist eher ein "erraten" [02] ➔ Neue Datenquellen müssen integrierbar sein. [08] ➔ man schon auf seine individuellen KPIs bezogen den richtigen Mix wählen [09]

(2) Data quality

Criteria	Statements
Highest possible data granularity of input sources	➔ Granularität erweitern [01] ➔ Ähm darauf müssen wir natürlich achten, dass wir den user nicht nicht über zwei Geräte zweimal abholen, sondern ihn wirklich relevant einmal wie wir ihn sehen ins Visier nehmen quasi und gerade beim crossdevice tracking und auch beim crossdomain-tracking ist es immer wichtig die Daten miteinander zu verknüpfen, um auch hier eben halt effizient arbeiten zu können [04] ➔ Ach so, da fällt immer das Wort Stitching, da würde ich die einzelnen Profile, die der User über verschiedene Domains über verschiedene Devices, dass man die zusammen führen kann [04] ➔ also die Datenqualität muss auf jeden Fall immer geprüft werden, das ist einer der wichtigsten Schritte, bevor man überhaupt irgendetwas analysiert mmmm die tracking-Daten natürlich hat man da erst einmal eine riesige Menge an Daten im Vergleich jetzt zu unseren Objektdaten, die sind ja deutlich geringer [06] ➔ Und da weiß man natürlich nie, was der Makler wirklich eingibt, z.B. da können natürlich deutlich mehr Fehler drinnen sein, als wenn die automatisiert erhoben werden, aber die muss man natürlich auch prüfen, ob da alles so richtig einläuft [06] ➔ Ich denk intern wird schon viel getrackt und da wird auch viel auf die Qualität geachtet, gerade weil es die eigenen Daten sind. Bei äh externen Daten bin ich immer erst mal sehr skeptisch, sofern sie nicht aus einer anderen offline-Quelle von mir z.B. kommen [07] ➔ Daten müssen so granular vorliegen wie möglich (Keywords, Placement, Timestamp) [08] ➔ alle Datenquellen, die der Firma in diesem Fall zur Verfügung stehen ähm, dass können CAM-Daten sein, das sind auf jeden Fall Bewegungsdaten, das sind Bewegungsdaten im Shop, Kaufinformationen, Warenkorbartikel, jedenfalls alles was man auswerten kann eigentlich. Je mehr desto besser [09]
Stich ability of a single user cross-devices	➔ Ähm darauf müssen wir natürlich achten, dass wir den user nicht nicht über zwei Geräte zweimal abholen, sondern ihn wirklich relevant einmal wie wir ihn sehen ins Visier nehmen quasi und gerade beim crossdevice tracking und auch beim crossdomain-tracking ist es immer wichtig die Daten miteinander zu verknüpfen, um auch hier eben halt effizient arbeiten zu können [04] ➔ Ach so, da fällt immer das Wort stitching, da würde ich die einzelnen Profile, die der user über verschiedene domains über verschiedene devices, dass man die zusammenführen kann [04] ➔ User Profil erstellen können (*impliziert geräteübergreifende Aktivitäten*) [02] ➔ Profil über verschiedene Devices hinweg [02] ➔ also die Datenqualität muss auf jeden Fall immer geprüft werden, das ist einer der wichtigsten Schritte, bevor man überhaupt irgendetwas analysiert mmmm die tracking-Daten natürlich hat man da erst einmal eine riesige Menge an Daten im Vergleich jetzt zu unseren Objektdaten, die sind ja deutlich geringer [06] ➔ ich denke es ist am besten, wenn die ganze Qualitätskette unter einer Hand zu haben [07] ➔ Also so so schnell wie es in der Situation irgendwie geht, damit der Nutzer in der Situation auf allen Geräten abgeholt wird, wo er gerade ist [07] ➔ CrossDeviceDaten: Realisiert mit Cookies und Login, Fingerprinting ist eher ein "erraten" [08] ➔ Wie bisher immer aus einem bestimmten Grund das Handy in die Hand nimmt, meist ist es immer i want to know, i want to buy, i want to get. Das sind immer klare Definitionen, warum man das Handy in die Hand nimmt, in dem Moment muss man halt aktiv handeln [09]
Linkable data sources	➔ wo der user sehr sehr viele Angebote natürlich hat und diese über sehr sehr viele verschiedene Wege auch wahrnehmen kann, dementsprechend müssen wir natürlich auch beispielsweise wie Remarktetingkampagnen irgendwie aussteuern und das am besten natürlich kosteneffizient, so dass wir beispielsweise die crossdevice tracking, was natürlich in diesem Fall sehr sehr wichtig ist [04] ➔ dass man über ähm gleiche tracking-Lösungen oder sehr sehr ähnliche tracking-Lösungen ohne große Abweichungen, die einzelnen Kanäle messen kann, damit da auch eine Vergleichbarkeit entsteht [06] ➔ also die Datenqualität muss auf jeden Fall immer geprüft werden, das ist einer der wichtigsten Schritte, bevor man überhaupt irgendetwas analysiert mmmm die tracking-Daten natürlich hat man da erst einmal eine riesige Menge an Daten im Vergleich jetzt zu unseren Objektdaten, die sind ja deutlich geringer [06] ➔ und das man verschiedenste Datenquellen auch verbinden kann [06] ➔ Sehr Schnittstellen-Intensiv (*Daten müssen verknüpfbar sein*) [02] ➔ Also so so schnell wie es in der Situation irgendwie geht, damit der Nutzer in der Situation auf allen Geräten abgeholt wird, wo er gerade ist [07] ➔ da wird auch viel auf die Qualität geachtet, gerade weil es die eigenen Daten sind. Bei äh externen Daten bin ich immer erst mal sehr skeptisch, sofern sie nicht aus einer anderen offline-Quelle von mir z.B. kommen [07] ➔ Hauptproblem: Unterschiede in der Datenqualität der Input-Sourcen *(Eine Verknüpfbarkeit wird vorausgesetzt)* [08] ➔ alle Datenquellen, die der Firma in diesem Fall zur Verfügung stehen [09]

(3) Calculation

Combination of *mathematical/ statistical approach* and *calculation*

Criteria	Statements
Ability to calculate in real-time	➔ akuten Interessen und Bedürfnisse eines Nutzers, die es schaffen, dass er die Aufmerksamkeit auf das Werbemittel lenkt, das er gerne haben möchte und das alles in einer real-time-attribution [09] ➔ und dann das Momentum definitiv verspielt ist [09] ➔ Wie bisher immer aus einem bestimmten Grund das Handy in die Hand nimmt, meist ist es immer i want to know, i want to buy, i want to get. Das sind immer klare Definitionen, warum man das Handy in die Hand nimmt, in dem Moment muss man halt aktiv handeln [09] ➔ Ja, also bei der Berechnung ist es immer so, dass man sich tatsächlich verschiedene Zeithorizonte anschauen muss. Also wir haben auf der einen Seite natürlich kurzfristige Ergebnisse in hier oder real time, die wir natürlich sehr sehr schnell brauchen und verarbeiten beispielsweise wenn in bestimmten Kampagnen, wo man davon ausgeht, dass die sofort performen sollen immer da brauchen wir schnellstmöglich die Daten dazu um evtl. korrigierend eingreifen zu können. Natürlich möchten wir aber auch, um überhaupt solche Kampagnen starten zu können oder um eben halt neue Produkte einstellen zu können, brauchen wir natürlich predictive analytics, wo man evtl. schon mal versucht zu erahnen was der user brauchen könnte, was er bisher noch nicht gebraucht hat. [04] ➔ Die Berechnung muss in Echtzeit erfolgen. Ein Modell mit Latenzen ist für die Zukunft ungeeignet der Attribution ungeeignet [03] ➔ Daten sollten auch, wenn es geht, automatisiert und in Echtzeit angepasst werden, sprich ein Kunde, der mal ein Kunde war, wieder aktiv wird, sollte möglichst auch über zielgerichtete Werbung, die vielleicht auch dynamisch und automatisch ausgespielt werden kann, wieder angesprochen werden [05] ➔ Also so so schnell wie es in der Situation irgendwie geht, damit der Nutzer in der Situation auf allen Geräten abgeholt wird, wo er gerade ist. [02] ➔ um dann nachzugucken, was man damit machen kann und dann kann man es natürlich immer weiter entwickeln mit selbstlernenden [06] ➔ also Ziel in der Zukunft sollte natürlich sein alles in Echtzeit hinzukriegen [05] ➔ eine nicht-zeitversetzte bzw. Echtzeitanalyse dessen, was kanalspezifisch Kunden tun [02] ➔ Die Berechnung muss in Echtzeit erfolgen. Ein Modell mit Latenzen ist für die Zukunft ungeeignet der Attribution ungeeignet [08]
Incremental learning process	➔ Konzept der Inkrementalität (Frage: Ist ein Touch oder eine andere Aktion relevant für den Outcome [08] ➔ Budget wird auf Userebene immer weiter verfeinert (inkrementelles Vorgehen und testen) [08] ➔ und vor allem auch zu sehen, welche Nutzer sind uninteressant, also das ist auch ein ganz wichtiger Punkt. Weil meines Erachtens viel Budget verschwendet wird, auf Nutzer, die für das jeweilige Geschäftsmodell nicht so interessant sind. (*Lernen aus Fehlern*) [05] ➔ Auto-Pilot-Modus (*beste Option wählen und lernen*) [02] ➔ selbstlernendes Verfahren anwenden kann, wie neuronales Netz z.B [06] ➔ Also aber vor allem auf den Benutzer, dass man eher stärker sich auf den Nutzer konzentriert ihn individual, individuell ihn anspricht und auch abholt [06] ➔ Der Einsatz von selbstlernenden Algorithmen ist ebenfalls denkbar *und* Regelmäßige Berechnung wie oft ist Industrie abhängig (*Lernprozess*). [03] ➔ akuten Interessen und Bedürfnisse eines Nutzers, die es schaffen, dass er die Aufmerksamkeit auf das Werbemittel lenkt, das er gerne haben möchte und das alles in einer real-time-attribution, das heißt, dass derjenige, der sich eine Digitalkamera anguckt, möchte vielleicht auch gerade eine Digitalkamera kaufen und da bringt es doch nichts, wenn man die Information übermorgen hat wenn er sich doch vorgestern eine Digitalkamera angeguckt hat *(Aus Nutzerverhalten lernen)* [09]

Ability to predict future actions	➔ Weil einfach nur Daten zu haben und Daten auszuwerten ohne sie hochzurechnen wird dir im Marketing nicht weiterhelfen [09] ➔ würde ich sagen predictive analytics [10] ➔ [Predictive] Ja, sind sehr interessant , sind werden ja auch teilweise jetzt schon eingesetzt, wenn man meinetwegen Wetterdaten etc. und unterschiedliche Datenquellen, die noch herangezogen werden, sind interessant, ähm die Quellen, die dafür herangezogen werden für die Berechnung müssten wenn es geht natürlich ähm ähm nachvollziehbar sein und so genau wie möglich [05] ➔ Also man hat so einen großen Datenstock an historischen Daten und auf Basis diesem dieser Historie kann man natürlich selbstlernende Modelle entwickeln bzw. predictive, was in Zukunft wahrscheinliche passieren wird auf Basis der historischen Daten, dass man sich z.B. einen Zeitraum nimmt und guckt, ob es für einen anderen Zeitraum auch gültig wäre dieses Modell [06] ➔ Predictive ist ein muss. Für omni-channel marketing Strategie. [02] ➔ Neben Insights aus historischen Daten müssen Vorhersagen getroffen werden können [03] ➔ Predictive Ansätze müssen auf User-Ebene (nicht Kanal-Ebene) arbeiten. D.h. es wird ein Kunden-Wert berechnet (CustomerValue) [08] ➔ Predictive Analysen müssen zusätzlich weitere Signale berücksichtigen: Sesonalität, Auktionsdynamic [08] ➔ das heißt, dass derjenige, der sich eine Digitalkamera anguckt, möchte vielleicht auch gerade eine Digitalkamera kaufen und da bringt es doch nichts, wenn man die Information übermorgen hat wenn er sich doch vorgestern eine Digitalkamera angeguckt hat [09]
Value calculation on user level	➔ immer ein Zielgruppenansatz ist und das man innerhalb dieser Zielgruppen dann auf jeden Fall runterbricht und versucht so individuell versucht arbeiten zu können [09] ➔ Das heißt, die Wertigkeit würde eher so ein bisschen beim Nutzer liegen, d.h. nicht mehr quasi an dem Verhalten direkt, sondern ähm eine Wertigkeit des Nutzers – würdest du das auch so sehen oder habe das missverstanden? [04] ➔ A.B. Nein, das ist absolut richtig, aber es geht mehr in die Richtung auf Basisinhalte, neue Tools und neue Prozesse , die wir haben, das wir eine Individualansprache machen und nicht nur wirklich die großen Segmente uns anschauen, sondern wir stellen den customer a über den customer b wenn er quasi pro visit oder pro sagen wir mal im seinem Lebenszyklus ➔ Customer Value berechnen [04] ➔ bin ich der Meinung, dass es immer mehr in die Richtung Individualisierung geht, wo ich die user wirklich einzeln ansprechen muss und ich nachgeschaltete Konzepte finden muss, dass kann natürlich nur passieren, wenn sich auch die Technik weiter entwickelt [04] ➔ – ähm im Idealfall geht das in Zukunft runter auf äh Personenbasis [08] ➔ Ziel: Budgetierung auf Kundenebene (User) [08] ➔ also ich denke, es wird immer immer wichtiger auf einzelne Personen zu gucken um möglichst den Interessen der Einzelnen gerecht zu werden und auch weil gerade viele Kanäle von immer mehr Personen benutzt werden, so dass die Wahl eines Kanals an sich u.U. gar nicht mehr so viel ausmacht. [04] ➔ Also aber vor allem auf den Benutzer, dass man eher stärker sich auf den Nutzer konzentriert ihn individual, individuell ihn anspricht und auch abholt und natürlich guckt, welcher Nutzer wie viel Kosten sollte oder wie viel man ihn in ihn in ihn investiert [06] ➔ Also Ausgangslage, äh, oder oder, ich glaube es wird mehr, oder wird Richtung einzelnen Kunden gehen [10] ➔ Predictive-Ansätze müssen auf User-Ebene (nicht Kanal-Ebene) arbeiten. D.h. es wird ein Kunden-Wert berechnet (Customer-Value) [08] ➔ Es muss personalisiert werden und dieses muss bei der Attribution berücksichtigt werden ➔ "Endziel" ist nicht Kanalbudgetberechnung oder User-Budgetberechnung sondern ein Gewinn im Unternehmen zu erzielen [08] ➔ Änderung von Signalen werden berücksichtigt [08] ➔ kanalübergreifend, ähm, diese Information, inwiefern das Gut bereits gekauft, ja, ähm, das sind glaube ich alles Dinge, die dann bei einer Ansprache in einem anderen Kanal, ähm, über z.B. Online-Werbung wesentlich berücksichtigt werden müssten. Also, äh, ich bin mir da sicher, dass das, äh, weiter zunehmen wird, also diese Individualisierung, ähm, in der Online-Werbung [10] ➔ Nicht umsetzbar: Pro Nutzer zu attribuieren (*Einschätzung des Experten, aber gewünscht*) [01] ➔ User-Profile erstellen können (gezielte Ansprache). (*Impliziert gezielte Berechnung*) [02] ➔ Ziel: Profilabhängig aussteuern und berechnen (= Budget auf Profilebene) [03] ➔ Daten sollten auch, wenn es geht, automatisiert und in Echtzeit angepasst werden, sprich ein Kunde, der mal ein Kunde war, wieder aktiv wird, sollte möglichst auch über zielgerichtete Werbung, die vielleicht auch dynamisch und automatisch ausgespielt werden kann, wieder angesprochen werden. (Arbeiten auf Kundenebene) [05]

Value calculation on audience basis	➔ Ach so, was jetzt gerade schon passiert ist, ist, das die meisten Budgets nicht mehr auf Kanal, sondern auf Audience geschiftet werden [09] ➔ Nutzer sind für mich zusammengefasst in einer Audience [09] ➔ um das Ganze nicht zu komplex zu machen, gewisse (husten) gewisse Cluster nach z.B. Merkmalen, Nutzungsverhalten zu bilden und vielleicht so in einer ich sag jetzt mal Vorstufe zu der kompletten Individualisierung, ähm, ist vielleicht so zu managen [10] ➔ es Sinn machen kann in der nächsten Stufe vielleicht gewisse Cluster zu bilden, gewisse Gruppierungen zu bilden oder sagen wir mal in einer vorgelagerten Stufe [10] ➔ diese zielgerichteten Botschaften für sagen wir jetzt ein Kundencluster oder einen einzelnen Kunden [05] ➔ das was ich zuvor angesprochen hatte steht der Nutzer im Mittelpunkt und meine ganze Betrachtung, die ich jetzt wiederum auch dieses System ähm überstülpe ist, immer wieder diese Betrachtung, dass der Nutzer oder ein Cluster von interessanten Nutzern quasi der Vordergrund ist [05] ➔ also nicht mehr auf Kanalebene sondern auf Segmentebende -> genau. aber innerhalb der Segmente gibt`s ja wiederrum die Kanäle die aber verschwimmen [05] ➔ nutzerzentriert zu arbeiten, dass man weiß, wer der Nutzer ist ähm, wieviel auch der einzelne Nutzer wert ist [05] ➔ und vor allem auch zu sehen, welche Nutzer sind uninteressant, also das ist auch ein ganz wichtiger Punkt [05] ➔ Nicht umsetzbar: Pro Nutzer zu attribuieren (*Einschätzung des Experten, eher auf Audience Ebene*) [01] ➔ User-Profile erstellen können (neue Touchpoints) (*ggf. auch aggregiert zu Gruppen*) [02] ➔ Das ganze muss man natürlich um zielgruppenspezifische Daten ergänzen [04] ➔ Ziel: Budgetierung auf Kundenebene (User) (*Vorstufe: Gruppenebene*) [08]
Machine learning / Artificial Intelligence approach Notes: because the user behavior is dynamic	➔ Machine-Learning-Technik und noch ein bisschen mein ähm mein Datentopf, in dem ich weiß, wer jetzt ein konkretes Interesse am Autokauf hat, weil es preislich möglich ist [09] ➔ Den Vorcast wird es natürlich meiner Meinung nach immer geben an sich, es kommt natürlich immer sag ich mal einen gewissen einen gewissen Blick in die Zukunft geben an sich. Ähm ich glaube eher, dass die die Berechnungsgrundlagen für die Erhebungsarbeit sich in den nächsten Jahren nicht nur etwas sondern sich grundlegend ändern wird wenn immer mehr sag ich mal das Eingreifen der künstlichen Intelligenz miterleben an sich [03] ➔ werden sehr sehr viel Automatisierung, sehr viel machine learning einfach erleben. Wenn man sich beispielsweise so was wie IBM Watson anschaut wo man jetzt einfach Daten per csv-Datei jetzt schon reinschmeisst und bekommt man bekommt die Korrelation in sehr sehr kurzer Zeit ausgespuckt und auch die Empfehlungen an sich, ähm, dann wird sich das auf jeden Fall vom sag ich mal händischen Ansatz eher zum Automatisierungsansatz hin entwickeln. Genauso ist es wichtig, dass mit Hilfe von mathematischen und statistischen Ansätzen natürlich irgendwann auch seine auch seine Remarketingkampagnen aufsetzt und nicht nur händisch, sondern dass man auch dafür tools letztendlich hat. [03] ➔ Genau richtig und auch Tools, die das Ganze auch dann verarbeiten können, dass man im Hintergrund die künstliche Intelligenz hat, die viele Sachen berechnet, die sie weitergibt per keine Ahnung per Connector [03] ➔ Also die Möglichkeit, eben auf der, sagen wir mal, Basis vorhandener Daten eben und entsprechend multivariater Verfahren eben, in die Zukunft zu schauen, ähm, was, äh, wir jetzt hier bei vielen Firmenprojekten, Firmenkooperationen auch, ähm, nutzen, ist sicherlich das Thema künstliche Intelligenz. Ja also, dass ich Systeme habe, die, ähm, ja, Lernen können, ja [10] ➔ selbstlernendes Verfahren anwenden kann, wie neuronales Netz z.B. [06] ➔ Regelbasierte Systeme sind nicht zukunftsfähig, kurzfristig aber noch relevant. Zukunft: Data Driven Ansätze (*Machine-Learning*) [08] ➔ Auto-Pilot-Modus (*Machine-Learning*) [02] ➔ Dynamisch -> KI. Nutzerverhalten ist auch dynamisch. [01]
Data driven calculation – not rule based	➔ Regelbasierte Systeme sind nicht zukunftsfähig, kurzfristig aber noch relevant. Zukunft: Data-Driven-Ansätze [08] ➔ Ich würde auf jeden Fall datengetriebene Modelle empfehlen die dynamisch, nah an Echtzeit auf Änderungen des Angebot und Nachfrage Volumens und auf externe Faktoren wie Werbung oder große Events reagieren können und den wirklichen Mehrwert einer Kampagne (Inkrementalität oder "was wäre passiert wenn der Benutzer die Werbung nicht gesehen hätte, hätte sie trotzdem gekauft") [08] ➔ akuten Interessen und Bedürfnisse eines Nutzers, die es schaffen, dass er die Aufmerksamkeit auf das Werbemittel lenkt, das er gerne haben möchte und das alles in einer real-time-attribution, das heißt, dass derjenige, der sich eine Digitalkamera anguckt, möchte vielleicht auch gerade eine Digitalkamera kaufen und da bringt es doch nichts, wenn man die Information übermorgen hat wenn er sich doch vorgestern eine Digitalkamera angeguckt hat (*Für alles eine Regel ist zu komplex*) [09]

(4) Output

Criteria	Statements
High-quality output	➔ Auto-Pilot-Modus (*Gute Qualität*) [02] ➔ also die Datenqualität muss auf jeden Fall immer geprüft werden, das ist einer der wichtigsten Schritte (*Schlechte Qualität rein -> schlechte Qualität raus*) [06] ➔ Sehr großes Problem: die Ergebnisse müssen gut sein [08] ➔ hat man prinzipiell immer das Problem zwischen Qualität und Reichweite und dann muss man schon auf seine individuellen KPIs bezogen den richtigen Mix wählen. [09] ➔ wir wollen natürlich schon äh viel Automatisierung eben halt haben, wir wollen natürlich ähm uns weiterentwickeln, wir wollen natürlich eine eine eine hohe Qualität der Daten, dementsprechend dann auch natürlich auch bei der Auswertung eine hohe Qualität haben [04] ➔ ich muss wirklich darauf achten, dass technische Hindernisse beseitigt sind, ähm, bzw. eben halt, dass man darauf achtet, dass ein tracking tool nicht unbedingt von einem app blocker geblockt wird letztendlich oder die user sich sofort outloggen können natürlich muss die Möglichkeit da sein für den user an sich [04] ➔ Also gerade bei den Daten von facebook wäre ich eher, was heißt vorsichtiger, aber. Ähm, da hätte ich zu mindestens, ähm, daran, kommt auf die Teilbereiche drauf an, ne, die auch einen interessieren. Ähm, dort wäre ich eher etwas vorsichtiger, ne? (*Schlechte Qualität rein -> schlechte Qualität raus*) [10] ➔ Also denen müsste man sicherlich schon ein bisschen kommt auf das Thema drauf an, sicherlich mit einer gewissen Skepsis, äh, begegnen, wohingegen ich die Daten von google, ähm, vom Gefühl her eher, jetzt eher, als qualitativ hochwertiger einschätzen würde (*Schlechte Qualität rein -> schlechte Qualität raus*) [10] ➔ wird schon viel getrackt und da wird auch viel auf die Qualität geachtet, gerade weil es die eigenen Daten sind. Bei äh externen Daten bin ich immer erst mal sehr skeptisch, sofern sie nicht aus einer anderen offline-Quelle von mir z.B. kommen (*Schlechte Qualität rein -> schlechte Qualität raus*) [01] ➔ also die Datenqualität muss auf jeden Fall immer geprüft werden, das ist einer der wichtigsten Schritte, bevor man überhaupt irgendetwas analysiert mmmm die tracking-Daten natürlich hat man da erst einmal eine riesige Menge an Daten im Vergleich jetzt zu unseren Objektdaten, die sind ja deutlich geringer (*Schlechte Qualität rein -> schlechte Qualität raus*) [06] ➔ Und da weiß man natürlich nie, was der Makler wirklich eingibt, z.B. da können natürlich deutlich mehr Fehler drinnen sein, als wenn die automatisiert erhoben werden, aber die muss man natürlich auch prüfen, ob da alles so richtig einläuft (*Schlechte Qualität rein -> schlechte Qualität raus*) [06]
Ability to connect (third party) vendors directly (automated connection)	➔ Genau richtig und auch Tools, die das Ganze auch dann verarbeiten können, dass man im Hintergrund die künstliche Intelligenz hat, die viele Sachen berechnet, die sie weitergibt per keine Ahnung per Connector [02] ➔ wir wollen natürlich schon äh viel Automatisierung eben halt haben, wir wollen natürlich ähm uns weiterentwickeln, wir wollen natürlich eine eine eine hohe Qualität der Daten, dementsprechend dann auch natürlich auch bei der Auswertung eine hohe Qualität haben [04] ➔ wenn ich mich für ein Tool entscheide und der Meinung bin, dass der Algorithmus auch wirklich funktioniert, gehe ich auch davon aus, dass das Tool eigenständig entscheiden kann, wenn Echtzeit sein soll ähm, wie Budget geshiftet wird (*komplett automatisiert, da zu komplex*) [05] ➔ Schnittstellgetrieben [08] ➔ also es sollte es sollte einen Punkt haben, in dem ich alle Informationen über die Nutzer dann umschlage (*…um diese dann weiter zu geben)* [07] ➔ gleichzeitig gibt sollte es ein Schnittstellenmangement geben [09]
Performance test of the model outcome/data validation	➔ AB-Funktion [09] ➔ glaube ich würde versuchen es zu validieren immer, wo an welcher Stelle es auch immer geht [07] ➔ würden auch Daten beim Einkauf, die ich wirklich halt schon hab oder nochmal verifizier gegen also oder auf jeden Fall da wo es geht verifizieren gegen gegen harte Daten [07] ➔ (Datenquellen) ersetzen Proof des Modells, funktioniert es richtig? [08] ➔ Kernfrage: Entsteht ein Mehrwert? (Tests: UserSplits über gewisse Segemnte / A/B Tests auf Basis von Modellen) [08] ➔ Testen können [03] ➔ Kennzahl: Qualität des Modells [02] ➔ Budgetallokation so in Echtzeit und ähm intelligent vornimmt, dass quasi ähm der Kanal äh, dass in den Kanal das Bugdet geshiftet wird, der der quasi nach unseren KPIs auch der performance test ist [05]
intuitive interface	➔ das user-interfaces, dass jemand bedienen kann der keine Info Informatik studiert hat, das ein ganz normaler Projektmanager auch bedienen kann [05] ➔ also auf jeden Fall eine gute Nutzeroberfläche ähm, dass man sich schnell schnell einfindet, schnell damit was machen kann und auf der anderen Seite trotzdem Gestaltungsmöglichkeiten bis in die Detailtiefe, wenn man das System dann mal schon gut versteht. [07] ➔ Komplexität muss händelbar sein. -> Doku [02]

Integration

(Unterscheiden: Know-How, Implementierung und Herausforderungen im Unternehmen)

Criteria	Statements
Interface driven design Notes: • Hook up input sources • Integration in existing system environment	➔ es wird so viele verschiedene Datenquellen geben, dass man immer eine gewisse Art von Schnittstellenmangement betreiben muss, weil ein System, was in irgendeiner Art und Weise geschlossen ist, wird nie ins komplette Bild für die Attribution in irgendeiner Art und Weise widerspiegeln können [09] ➔ gleichzeitig gibt sollte es ein Schnittstellenmangement geben, an die ähn an die bestehenden Systeme CRM etwa wie Ad-Server, GSP, SSP, des Kunden, also ich glaube, das ist eine richtige Infrastruktur, die man dort braucht. [09] ➔ ganz eher muss so ,n System ne gewisse, äh, muss dynamisch sein, so n System muss sicherlich Schnittstellen oder Möglichkeiten für, ich sag mal, ähm, Anbindung an bereits auch bestehende Systeme haben [10] ➔ aber du würdest eher was schnittstellengetriebenes in der Zukunft sehen, als quasi so`n Monoprogramm, dass alles -> JA [06] ➔ Struktur: Schnittstellengetrieben, das Modell soll keine Systeme (Datenquellen) ersetzen [08] ➔ einer Datamanagementplattform, die letztendlich dann auf Basis des Segmente, die dann berechnet wurden oder vom user eingestellt wurden bzw. vom Marketingmanager das ganze automatisch abgerufen wird [04] ➔ nutzerfreudig ist und auch intuitiv bedienbar ist [04] ➔ Reportingfunktionen wären natürlich top. [02]
Interface definitions / standards	➔ d.h. für mich ist es schon so, dass man einen großen zentralen Pott hat, in man alle Daten rein gibt, in dem sich historische Daten befinden, genauso wie auch Echtzeitdaten, die man, je nachdem, für was man sie gerade benötigt individuell hochrechnen kann [09] ➔ Ich glaube, alles was standardisiert werden kann, soll standardisiert werden ähm, es wird dann wo vierzig fünfzig Grundsegmente geben, die auch auf Internet- und Rohdaten beruhen und die auch immer pauschal genutzt werden können und dann wird es kundenindividuelle Lösungen geben [09] ➔ Struktur: Schnittstellengetrieben, das Modell soll keine Systeme (Datenquellen) ersetzen [08] ➔ es muss natürliche irgendwie anwendbar sein, also es sollte nicht zu speziell vielleicht sein [06] ➔ Plug and Play mit ETL (muss in bestehende Infrastruktur eingepasst werden) [02] ➔ und an den vorherigen natürlich anschließt irgendwo und man sollte natürlich auch darauf abzielen, dass es im gewissen Maße auch äh nutzerfreudig ist und auch intuitiv bedienbar ist, was immer dahingeht, dass man die Leute schult und natürlich auch äh, dass man den Leuten, wie sie damit umgehen können. [04]
Plug and play	➔ es kann nicht sein, dass man dafür coden können muss um ein solches Instrument zu bedienen, dann wird es nie in irgendeiner Art und Weise die Relevanz bekommen, die es braucht im Marketing, weil diese Leute haben die da größtenteils nicht in den Köpfen [09] ➔ Plug and Play mit ETL (muss in bestehende Infrastruktur eingepasst werden) [02] ➔ und an den vorherigen natürlich anschließt irgendwo und man sollte natürlich auch darauf abzielen, dass es im gewissen Maße auch äh nutzerfreudig ist und auch intuitiv bedienbar ist, was immer dahingeht, dass man die Leute schult und natürlich auch äh, dass man den Leuten, wie sie damit umgehen können. [04]

Personalisierung

Removed

BI Knowledge

Added

Criteria	Statements
Basic skills from the Busniess Inetelligence (BI)	➔ Also ich glaube, dass es schon immer technischer wird und auch immer technischer werden sollte und das jeder vernünftige KPI- und auch jeder BI-Kenntnisse mitbringen sollte, der in solchen Bereichen arbeitet [09] ➔ Also, ich brauch keinen, also ich brauche keinen Akademiker dafür, der mir das hochwissenschaftlich berechnet, da haben wir auch nie die Zeit zu (= Zeitkritisch) [09] ➔ Try & Error: Kaum die Möglichkeit Channel-übergreifend zu messen (*es wird Fachwissen benötigt*) [01] ➔ und äh ja know-how muss im Unternehmen auf jeden Fall vorhanden sein, es muss auch, sag ich mal ausgeweitet werden, d.h. derjenige, der initial dafür zuständig ist muss natürlich dafür Sorge tragen, dass know-how über über das Tool über das System letztendlich ähm ja weitflächig gestreut wird und auch irgendwann den Fortbestand von so einer mathematischen Weiterentwicklung so gewährleisten ähm genauso muss natürlich weitergedacht werden, man darf nicht verharren, das ist äh wie bei allen Ansätzen, die es in unserem Bereich gibt [04] ➔ bei uns ist es jetzt auch einfach so, dass die tools, die wir neu nutzen, auch wenn das quasi self-service-tools sind ähm, braucht man doch noch jemanden, der einen quasi customer-service-mäßig an die Hand nimmt und zumindest erklären kann, ähm, wie der Algorithmus arbeitet [03] ➔ Änderung von Signalen werden berücksichtigt (Diese müssen erkannt werden -> Fachwissen) [08]
Basic understanding of technical aspects Notes: Identify promising measures and strategies	➔ und äh dies in einem Unternehmen zu streuen, dass dies wirklich wichtig ist, ist glaub ich schon eine sehr sehr große Hürde an sich, d.h.er muss sehr sehr viel geschult werden [04] ➔ dass das System eine entsprechende Bedeutung hat und das auch vor allem abteilungsübergreifend, um da so ein bisschen auf den organisatorischen Aspekt mit einzugehen [10] ➔ Änderung von Signalen werden berücksichtigt (Diese müssen erkannt werden -> Fachwissen) [08] ➔ also ich denke, dass auch wichtig ist, das irgendwie auch beim Personal da Verständnis für vorhanden ist oder sie ja geschult werden, damit sie damit umgehen können, und man muss auch die Systeme auch immer kritisch hinterfragen und es nicht einfach so hinnehmen was da rauskommt, sondern, dass man sich wirklich damit auseinandersetzt [09] ➔ Try & Error: Kaum die Möglichkeit Channel-übergreifend zu messen (*es wird technisches Grundverständnis benötigt*) [01] ➔ muss muss äh irgendwo der Datenschutz gewährleistet sein natürlich der user muss sich sag ich mal, auf uns als Unternehmen verlassen können, das wir keinen Schmu machen mit den Daten [05]

Appendix 3: Questionnaire Guideline

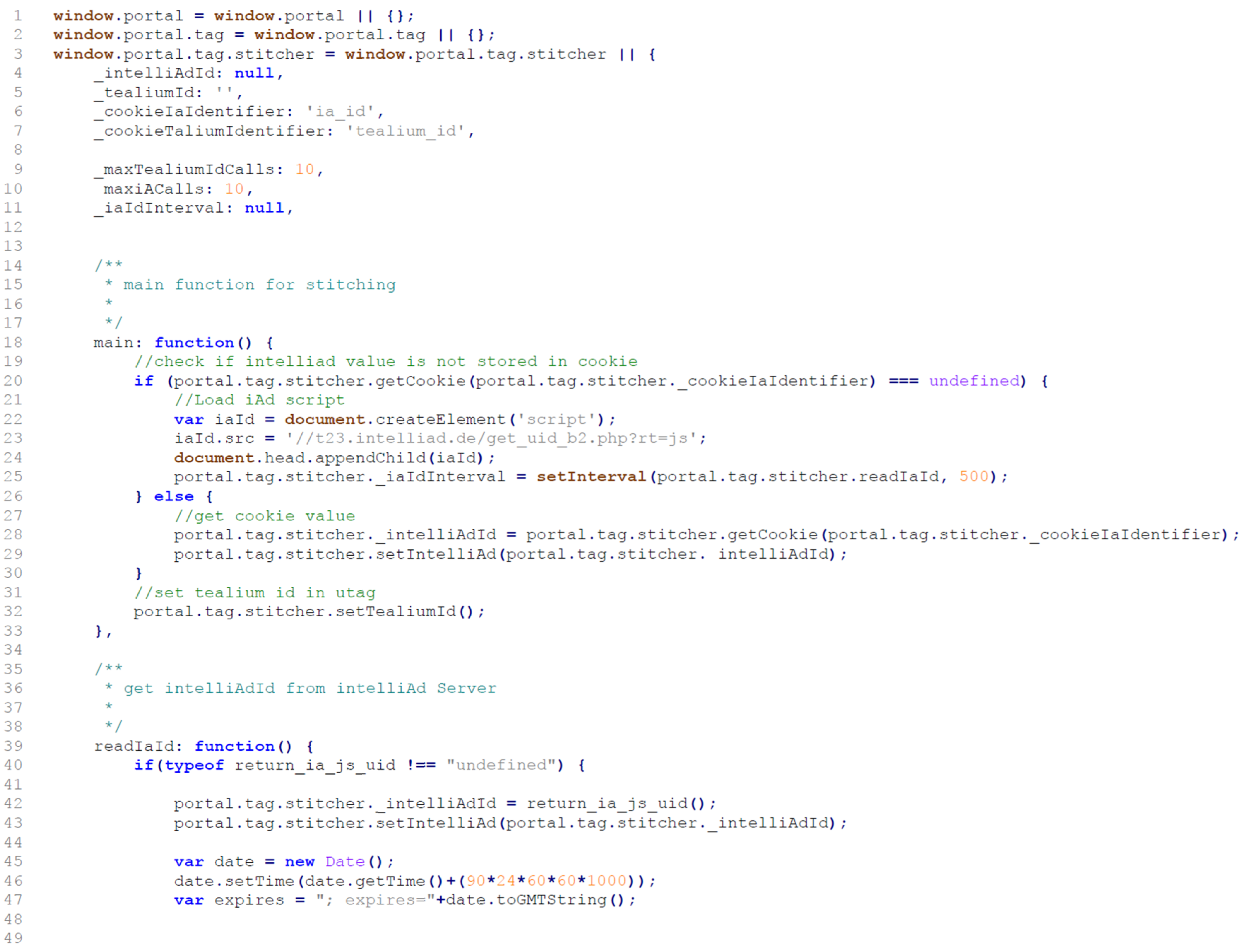
Questionnaire Guideline for the semi-structured interviews. (All interviews were conducted in German)

Questionnaire Guideline

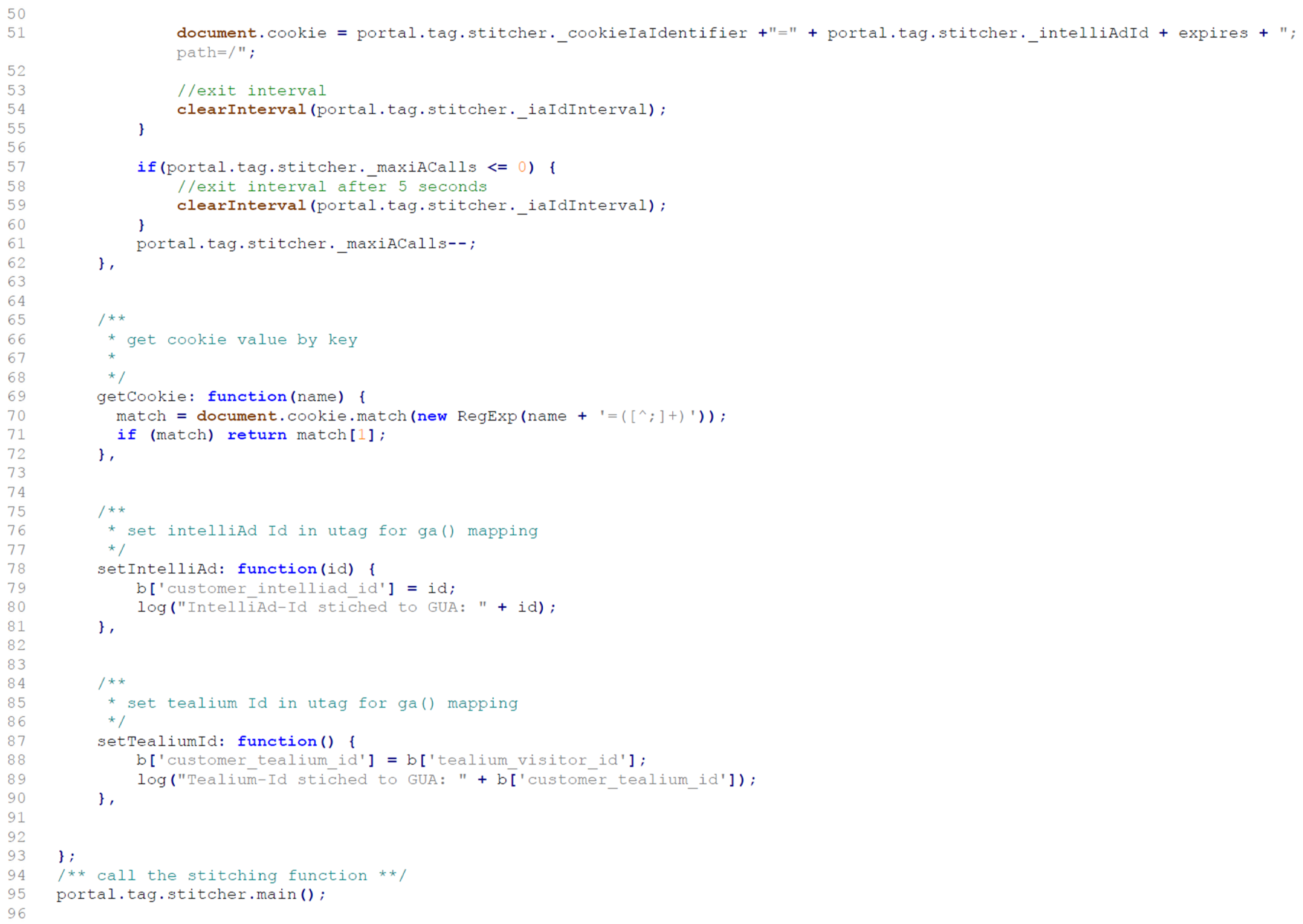
The interviewee needs to be informed about / asked for the following aspects:

- The collected data will only be used for the dissertation project of Ole Nass.
- Data will be anonymized.
- Is a voice recording acceptable?
- Explain the following terms and answer arising questions before the interview begins:
    - Dynamic attribution: An attribution model/approach utilizing generated data (e.g. user interaction data) to perform the budget allocation
    - Omni-Channel: Cross-channel marketing activities with a seamless switching options between channels. A central data hub is present to support marketing decisions.
    - IPO-Model: Input, process, output.
- Discuss / Talk about features, requirements, and/or personal opinions in terms of the following predefined categories. Each category is explained shortly by the interviewer. The interviewee should explain requirements towards importance and personal opinions.
    - INPUT
        - Variablen (variables)
        - Datengundlage / Datenqualität (data foundation / data quality)
    - PROCESS
        - Mathematischer / statistischer Ansatz (mathematical / statistical approach)
        - Berechnung (calculation)
    - OUPUT
        - Anbindung (connection (to other systems), third party vendors)
    - INDIVIDULAZATION
        - Personalisierung (personalization)
    - IMPLEMENTATION
        - Productiver Einsatz (productive use)
        - Implementierung (implementation)
    - OTHERS
        - Allgemeine Informationen (general information)
        - Sonstiges (other)

Appendix 4: Java-Script: ID-Matching

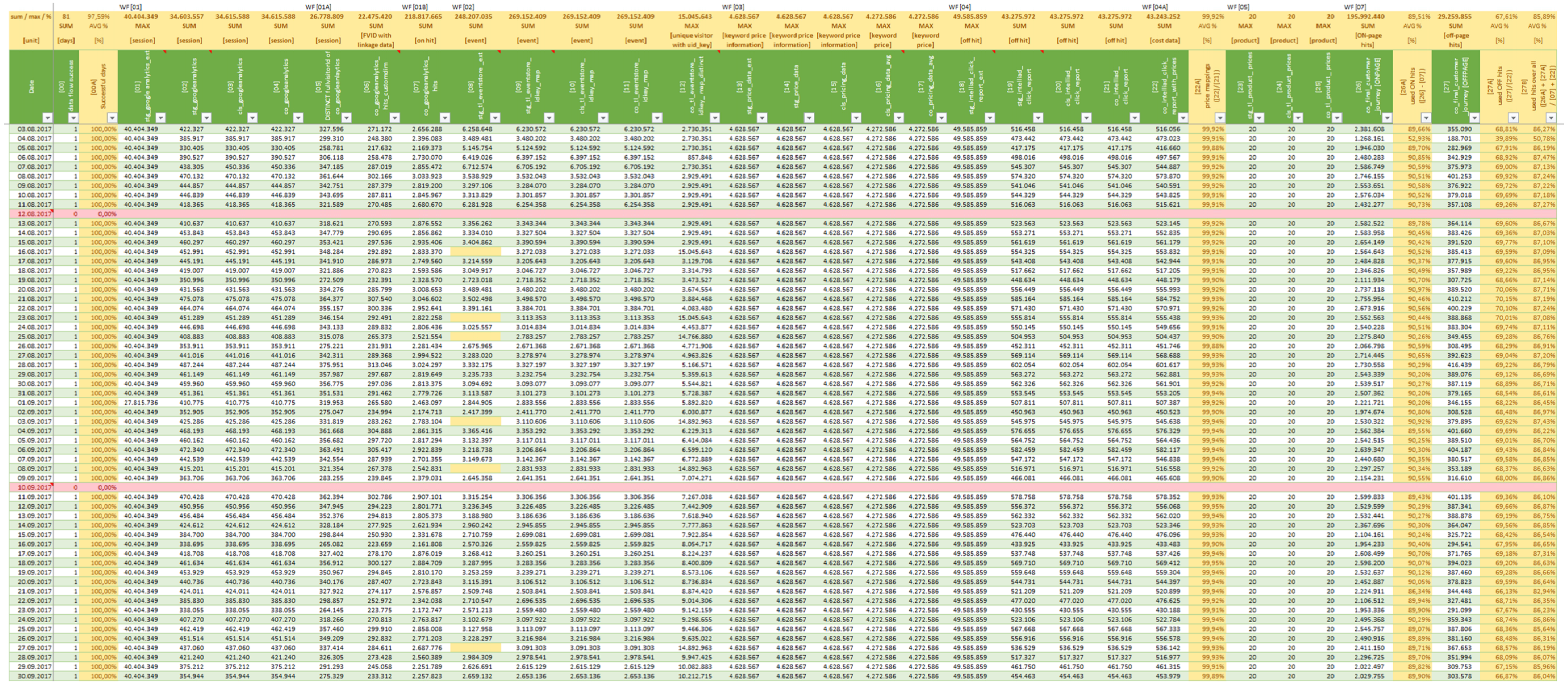


```
window.portal = window.portal || {};
window.portal.tag = window.portal.tag || {};
window.portal.tag.stitcher = window.portal.tag.stitcher || {
    _intelliAdId: null,
    _tealiumId: '',
    _cookieIaIdentifier: 'ia_id',
    _cookieTaliumIdentifier: 'tealium_id',

    _maxTealiumIdCalls: 10,
     maxiACalls: 10,
    _iaIdInterval: null,


    /**
     * main function for stitching
     *
     */
    main: function() {
        //check if intelliad value is not stored in cookie
        if (portal.tag.stitcher.getCookie(portal.tag.stitcher._cookieIaIdentifier) === undefined) {
            //Load iAd script
            var iaId = document.createElement('script');
            iaId.src = '//t23.intelliad.de/get_uid_b2.php?rt=js';
            document.head.appendChild(iaId);
            portal.tag.stitcher._iaIdInterval = setInterval(portal.tag.stitcher.readIaId, 500);
        } else {
            //get cookie value
            portal.tag.stitcher._intelliAdId = portal.tag.stitcher.getCookie(portal.tag.stitcher._cookieIaIdentifier);
            portal.tag.stitcher.setIntelliAd(portal.tag.stitcher. intelliAdId);
        }
        //set tealium id in utag
        portal.tag.stitcher.setTealiumId();
    },

    /**
     * get intelliAdId from intelliAd Server
     *
     */
    readIaId: function() {
        if(typeof return_ia_js_uid !== "undefined") {

            portal.tag.stitcher._intelliAdId = return_ia_js_uid();
            portal.tag.stitcher.setIntelliAd(portal.tag.stitcher._intelliAdId);

            var date = new Date();
            date.setTime(date.getTime()+(90*24*60*60*1000));
            var expires = "; expires="+date.toGMTString();

```

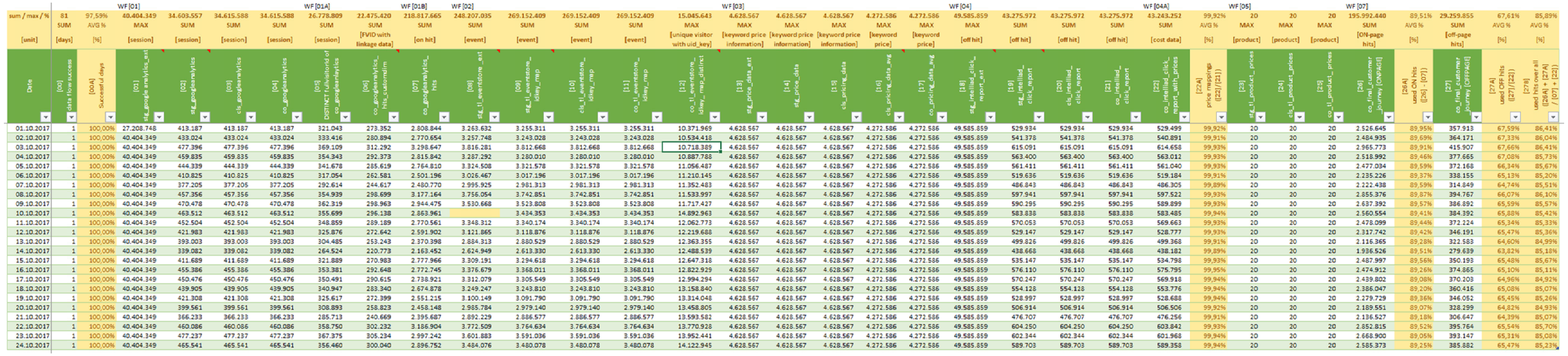


```
            document.cookie = portal.tag.stitcher._cookieIaIdentifier +"=" + portal.tag.stitcher._intelliAdId + expires + ";
            path=/";

            //exit interval
            clearInterval(portal.tag.stitcher._iaIdInterval);
        }

        if(portal.tag.stitcher._maxiACalls <= 0) {
            //exit interval after 5 seconds
            clearInterval(portal.tag.stitcher._iaIdInterval);
        }
        portal.tag.stitcher._maxiACalls--;
    },


    /**
     * get cookie value by key
     *
     */
    getCookie: function(name) {
       match = document.cookie.match(new RegExp(name + '=([^;]+)'));
       if (match) return match[1];
    },


    /**
     * set intelliAd Id in utag for ga() mapping
     */
    setIntelliAd: function(id) {
        b['customer_intelliad_id'] = id;
        log("IntelliAd-Id stiched to GUA: " + id);
    },


    /**
     * set tealium Id in utag for ga() mapping
     */
    setTealiumId: function() {
        b['customer_tealium_id'] = b['tealium_visitor_id'];
        log("Tealium-Id stiched to GUA: " + b['customer_tealium_id']);
    },


};
/** call the stitching function **/
portal.tag.stitcher.main();

```

Appendix 5: Transformation performance



			WF [01]				WF [01A]		WF [01B]	WF [02]					WF [03]					WF [04]				WF [04A]		WF [05]			WF [07]				
sum / max / %	81	97,59%	40.404.349	34.603.557	34.615.588	34.615.588	26.778.809	22.475.420	218.817.665	248.207.035	269.152.409	269.152.409	269.152.409	15.045.643	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	43.275.972	43.275.972	43.275.972	43.243.252	99,92%	20	20	20	195.992.440	89,51%	29.259.855	67,61%	85,89%
	SUM	AVG %	MAX	SUM	SUM	SUM	SUM	SUM	SUM	SUM	SUM	SUM	SUM	MAX	MAX	MAX	MAX	MAX	MAX	MAX	SUM	SUM	SUM	SUM	AVG %	MAX	MAX	MAX	SUM	AVG %	SUM	AVG %	AVG %
[unit]	[days]	[%]	[session]	[session]	[session]	[session]	[session]	[FVID with linkage data]	[on hit]	[event]	[event]	[event]	[event]	[unique visitor with uid_key]	[keyword price information]	[keyword price information]	[keyword price information]	[keyword price]	[keyword price]	[off hit]	[off hit]	[off hit]	[off hit]	[cost data]	[%]	[product]	[product]	[product]	[ON-page hits]	[%]	[off-page hits]	[%]	[%]
Date	[00] data flow success	[00A] Successful days	[01] stg_google analytics_ext	[02] stg_googleanalytics	[03] cls_googleanalytics	[04] co_googleanalytics	[05] DISTINCT fullvisitorid of co_googleanlaytics	[06] co_googleanalytics_ hits_customdim	[07] co_googleanalytics_ hits	[08] stg_tl_eventstore _ext	[09] stg_tl_eventstore_ idkey_map	[10] cls_tl_eventstore_ idkey_map	[11] co_tl_eventstore_ idkey_map	[12] co_tl_eventstore_ idkey_ map_distinct	[13] stg_price_data_ext	[14] stg_price_data	[15] cls_pricing_data	[16] cls_pricing_data_avg	[17] co_pricing_data_avg	[18] stg_intelliad_click_ report_ext	[19] stg_intelliad_ click_report	[20] cls_intelliad_ click_report	[21] co_intelliad_ click_report	[22] co_intelliad_click_ report_with_prices	[22A] price mappings ([22]/[21])	[23] stg_tl_product_ prices	[24] cls_tl_product_prices	[25] co_tl_product_ prices	[26] co_final_customer _journey [ONPAGE]	[26A] used ON hits ([26] - [07])	[27] co_final_customer _journey [OFFPAGE]	[27A] used OFF hits ([27]/[22])	[27B] used hits over all ([26A] + [27A] / [07] + [22])
03.08.2017	1	100,00%	40.404.349	422.327	422.327	422.327	327.596	271.172	2.656.288	6.258.648	6.230.572	6.230.572	6.230.572	2.730.351	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	516.458	516.458	516.458	516.056	99,92%	20	20	20	2.381.608	89,66%	355.090	68,81%	86,27%
04.08.2017	1	100,00%	40.404.349	385.917	385.917	385.917	299.310	248.380	2.396.083	3.489.481	3.480.202	3.480.202	3.480.202	2.730.351	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	473.442	473.442	473.442	473.023	99,91%	20	20	20	1.268.161	52,93%	188.701	39,89%	50,78%
05.08.2017	1	100,00%	40.404.349	330.405	330.405	330.405	258.781	217.632	2.169.373	5.145.754	5.124.592	5.124.592	5.124.592	2.730.351	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	417.175	417.175	417.175	416.660	99,88%	20	20	20	1.946.030	89,70%	282.969	67,91%	86,19%
06.08.2017	1	100,00%	40.404.349	390.527	390.527	390.527	306.118	258.478	2.730.070	6.419.026	6.397.152	6.397.152	6.397.152	857.848	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	498.016	498.016	498.016	497.567	99,91%	20	20	20	2.480.283	90,85%	342.929	68,92%	87,47%
07.08.2017	1	100,00%	40.404.349	438.305	450.336	450.336	347.185	287.019	2.855.472	6.712.574	6.705.192	6.705.192	6.705.192	2.730.351	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	545.307	545.307	545.307	544.887	99,92%	20	20	20	2.586.749	90,59%	375.973	69,00%	87,13%
08.08.2017	1	100,00%	40.404.349	470.132	470.132	470.132	361.644	302.166	3.033.923	3.538.929	3.532.043	3.532.043	3.532.043	2.929.491	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	574.320	574.320	574.320	573.870	99,92%	20	20	20	2.746.155	90,51%	401.253	69,92%	87,24%
09.08.2017	1	100,00%	40.404.349	444.857	444.857	444.857	342.751	287.379	2.819.200	3.297.106	3.284.070	3.284.070	3.284.070	2.929.491	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	541.046	541.046	541.046	540.591	99,92%	20	20	20	2.553.651	90,58%	376.922	69,72%	87,22%
10.08.2017	1	100,00%	40.404.349	446.839	446.839	446.839	343.695	287.811	2.845.967	3.313.829	3.301.857	3.301.857	3.301.857	2.929.491	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	544.329	544.329	544.329	543.825	99,91%	20	20	20	2.576.034	90,52%	379.018	69,69%	87,18%
11.08.2017	1	100,00%	40.404.349	418.365	418.365	418.365	321.589	270.485	2.680.670	6.281.928	6.254.358	6.254.358	6.254.358	2.929.491	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	516.063	516.063	516.063	515.621	99,91%	20	20	20	2.432.277	90,73%	357.108	69,26%	87,27%
12.08.2017	0	0,00%																															
13.08.2017	1	100,00%	40.404.349	410.637	410.637	410.637	318.621	270.593	2.876.552	3.356.262	3.343.344	3.343.344	3.343.344	2.929.491	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	523.563	523.563	523.563	523.145	99,92%	20	20	20	2.582.522	89,78%	364.114	69,60%	86,67%
14.08.2017	1	100,00%	40.404.349	453.843	453.843	453.843	347.779	290.695	2.856.862	3.334.010	3.327.504	3.327.504	3.327.504	2.929.491	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	553.271	553.271	553.271	552.835	99,92%	20	20	20	2.583.958	90,45%	383.426	69,36%	87,03%
15.08.2017	1	100,00%	40.404.349	460.297	460.297	460.297	353.421	297.536	2.935.406	3.404.862	3.390.594	3.390.594	3.390.594	2.929.491	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	561.619	561.619	561.619	561.179	99,92%	20	20	20	2.654.149	90,42%	391.520	69,77%	87,10%
16.08.2017	1	100,00%	40.404.349	452.991	452.991	452.991	348.284	292.892	2.833.370		3.272.033	3.272.033	3.272.033	15.045.643	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	554.325	554.325	554.325	553.832	99,91%	20	20	20	2.564.643	90,52%	385.413	69,59%	87,09%
17.08.2017	1	100,00%	40.404.349	445.191	445.191	445.191	341.910	286.973	2.749.560	3.214.559	3.205.643	3.205.643	3.205.643	3.129.708	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	543.408	543.408	543.408	542.944	99,91%	20	20	20	2.484.828	90,37%	377.915	69,60%	86,95%
18.08.2017	1	100,00%	40.404.349	419.007	419.007	419.007	321.886	270.823	2.593.586	3.049.917	3.046.727	3.046.727	3.046.727	3.314.793	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	517.662	517.662	517.662	517.205	99,91%	20	20	20	2.346.826	90,49%	357.989	69,22%	86,95%
19.08.2017	1	100,00%	40.404.349	350.996	350.996	350.996	272.509	232.391	2.328.570	2.723.018	2.718.352	2.718.352	2.718.352	3.473.527	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	448.634	448.634	448.634	448.179	99,90%	20	20	20	2.111.934	90,70%	307.725	68,66%	87,14%
20.08.2017	1	100,00%	40.404.349	431.563	431.563	431.563	334.276	285.799	3.008.653	3.489.481	3.480.202	3.480.202	3.480.202	3.674.554	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	556.449	556.449	556.449	555.993	99,92%	20	20	20	2.737.118	90,97%	389.520	70,06%	87,71%
21.08.2017	1	100,00%	40.404.349	475.078	475.078	475.078	364.377	307.540	3.046.602	3.502.498	3.498.570	3.498.570	3.498.570	3.884.468	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	585.164	585.164	585.164	584.752	99,93%	20	20	20	2.755.954	90,46%	410.212	70,15%	87,19%
22.08.2017	1	100,00%	40.404.349	464.074	464.074	464.074	355.157	300.336	2.952.641	3.391.161	3.384.701	3.384.701	3.384.701	4.083.480	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	571.430	571.430	571.430	570.971	99,92%	20	20	20	2.673.916	90,56%	400.229	70,10%	87,24%
23.08.2017	1	100,00%	40.404.349	451.289	451.289	451.289	346.154	292.491	2.822.258		3.113.353	3.113.353	3.113.353	15.045.643	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	555.814	555.814	555.814	555.438	99,93%	20	20	20	2.552.563	90,44%	388.868	70,01%	87,08%
24.08.2017	1	100,00%	40.404.349	446.698	446.698	446.698	343.133	289.832	2.806.436	3.025.557	3.014.834	3.014.834	3.014.834	4.453.877	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	550.145	550.145	550.145	549.656	99,91%	20	20	20	2.540.228	90,51%	383.304	69,74%	87,11%
25.08.2017	1	100,00%	40.404.349	408.883	408.883	408.883	315.078	265.373	2.521.554		2.783.257	2.783.257	2.783.257	14.766.880	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	504.953	504.953	504.953	504.437	99,90%	20	20	20	2.275.840	90,26%	349.455	69,28%	86,76%
26.08.2017	1	100,00%	40.404.349	353.911	353.911	353.911	275.221	231.931	2.281.434	2.675.965	2.671.368	2.671.368	2.671.368	4.771.908	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	452.311	452.311	452.311	451.746	99,88%	20	20	20	2.066.798	90,59%	308.495	68,29%	86,91%
27.08.2017	1	100,00%	40.404.349	441.016	441.016	441.016	342.311	289.368	2.994.522	3.283.020	3.278.974	3.278.974	3.278.974	4.963.826	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	569.114	569.114	569.114	568.688	99,93%	20	20	20	2.714.445	90,65%	392.623	69,04%	87,20%
28.08.2017	1	100,00%	40.404.349	487.244	487.244	487.244	375.951	313.046	3.024.297	3.332.175	3.327.197	3.327.197	3.327.197	5.166.571	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	602.054	602.054	602.054	601.617	99,93%	20	20	20	2.730.558	90,29%	416.439	69,22%	86,79%
29.08.2017	1	100,00%	40.404.349	461.149	461.149	461.149	357.987	297.687	2.819.649	3.235.733	3.232.754	3.232.754	3.232.754	5.359.613	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	563.272	563.272	563.272	562.881	99,93%	20	20	20	2.543.339	90,20%	389.076	69,12%	86,69%
30.08.2017	1	100,00%	40.404.349	459.960	459.960	459.960	356.775	297.036	2.813.375	3.094.692	3.093.077	3.093.077	3.093.077	5.544.821	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	562.326	562.326	562.326	561.901	99,92%	20	20	20	2.539.517	90,27%	387.119	68,89%	86,71%
31.08.2017	1	100,00%	40.404.349	451.361	451.361	451.361	351.531	291.462	2.779.726	3.113.587	3.101.273	3.101.273	3.101.273	5.728.387	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	553.545	553.545	553.545	553.205	99,94%	20	20	20	2.507.362	90,20%	379.165	68,54%	86,61%
01.09.2017	1	100,00%	27.815.736	410.775	410.775	410.775	319.953	265.580	2.463.097	2.844.905	2.833.556	2.833.556	2.833.556	5.892.820	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	507.811	507.811	507.811	507.387	99,92%	20	20	20	2.221.721	90,20%	346.155	68,22%	86,45%
02.09.2017	1	100,00%	40.404.349	352.905	352.905	352.905	275.047	234.994	2.174.713	2.417.399	2.411.770	2.411.770	2.411.770	6.030.877	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	450.963	450.963	450.963	450.523	99,90%	20	20	20	1.974.674	90,80%	308.528	68,48%	86,97%
03.09.2017	1	100,00%	40.404.349	425.286	425.286	425.286	331.819	283.262	2.783.104		3.110.606	3.110.606	3.110.606	14.892.963	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	545.975	545.975	545.975	545.638	99,94%	20	20	20	2.530.322	90,92%	379.895	69,62%	87,43%
04.09.2017	1	100,00%	40.404.349	468.193	468.193	468.193	361.668	304.888	2.861.315	3.365.416	3.353.292	3.353.292	3.353.292	6.229.313	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	576.655	576.655	576.655	576.329	99,94%	20	20	20	2.562.384	89,55%	401.660	69,69%	86,22%
05.09.2017	1	100,00%	40.404.349	460.162	460.162	460.162	356.682	297.720	2.817.294	3.132.397	3.117.011	3.117.011	3.117.011	6.414.084	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	564.752	564.752	564.752	564.436	99,94%	20	20	20	2.542.515	90,25%	389.510	69,01%	86,70%
06.09.2017	1	100,00%	40.404.349	472.340	472.340	472.340	363.491	305.417	2.922.839	3.218.738	3.206.864	3.206.864	3.206.864	6.599.120	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	582.459	582.459	582.459	582.117	99,94%	20	20	20	2.639.347	90,30%	404.187	69,43%	86,84%
07.09.2017	1	100,00%	40.404.349	442.539	442.539	442.539	342.554	287.939	2.701.355	3.149.673	3.142.367	3.142.367	3.142.367	6.772.889	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	547.172	547.172	547.172	546.838	99,94%	20	20	20	2.440.680	90,35%	380.517	69,58%	86,85%
08.09.2017	1	100,00%	40.404.349	415.201	415.201	415.201	321.354	267.378	2.542.831		2.831.933	2.831.933	2.831.933	14.892.963	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	516.971	516.971	516.971	516.558	99,92%	20	20	20	2.297.257	90,34%	353.189	68,37%	86,63%
09.09.2017	1	100,00%	40.404.349	363.706	363.706	363.706	283.255	239.845	2.379.031	2.645.358	2.641.351	2.641.351	2.641.351	7.074.271	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	466.081	466.081	466.081	465.608	99,90%	20	20	20	2.154.231	90,55%	316.610	68,00%	86,86%
10.09.2017	0	0,00%																															
11.09.2017	1	100,00%	40.404.349	470.428	470.428	470.428	362.394	302.786	2.907.101	3.315.254	3.306.356	3.306.356	3.306.356	7.267.038	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	578.758	578.758	578.758	578.352	99,93%	20	20	20	2.599.833	89,43%	401.135	69,36%	86,10%
12.09.2017	1	100,00%	40.404.349	450.956	450.956	450.956	347.945	294.223	2.801.771	3.236.345	3.226.485	3.226.485	3.226.485	7.442.909	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	556.372	556.372	556.372	556.068	99,95%	20	20	20	2.529.599	90,29%	387.341	69,66%	86,87%
13.09.2017	1	100,00%	40.404.349	456.484	456.484	456.484	352.376	294.813	2.805.373	3.188.980	3.186.636	3.186.636	3.186.636	7.618.940	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	562.332	562.332	562.332	562.020	99,94%	20	20	20	2.532.441	90,27%	388.878	69,19%	86,75%
14.09.2017	1	100,00%	40.404.349	424.612	424.612	424.612	328.184	277.925	2.621.934	2.960.242	2.945.855	2.945.855	2.945.855	7.777.863	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	523.703	523.703	523.703	523.346	99,93%	20	20	20	2.367.696	90,30%	364.047	69,56%	86,85%
15.09.2017	1	100,00%	40.404.349	384.700	384.700	384.700	298.844	250.930	2.331.678	2.710.759	2.699.081	2.699.081	2.699.081	7.922.854	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	476.440	476.440	476.440	476.096	99,93%	20	20	20	2.104.161	90,24%	325.722	68,42%	86,54%
16.09.2017	1	100,00%	40.404.349	338.695	338.695	338.695	265.082	223.659	2.161.808	2.570.326	2.559.825	2.559.825	2.559.825	8.054.717	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	433.925	433.925	433.925	433.483	99,90%	20	20	20	1.954.233	90,40%	294.541	67,95%	86,65%
17.09.2017	1	100,00%	40.404.349	418.708	418.708	418.708	327.402	278.170	2.876.019	3.268.412	3.260.251	3.260.251	3.260.251	8.224.237	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	537.748	537.748	537.748	537.426	99,94%	20	20	20	2.608.499	90,70%	371.765	69,18%	87,31%
18.09.2017	1	100,00%	40.404.349	461.634	461.634	461.634	356.912	300.127	2.884.709	3.287.995	3.283.356	3.283.356	3.283.356	8.400.809	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	569.710	569.710	569.710	569.412	99,95%	20	20	20	2.598.200	90,07%	394.023	69,20%	86,63%
19.09.2017	1	100,00%	40.404.349	453.929	453.929	453.929	350.967	294.845	2.810.170	3.253.259	3.239.271	3.239.271	3.239.271	8.573.106	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	559.648	559.648	559.648	559.304	99,94%	20	20	20	2.532.637	90,12%	387.460	69,28%	86,66%
20.09.2017	1	100,00%	40.404.349	440.736	440.736	440.736	340.176	287.407	2.723.843	3.115.391	3.106.512	3.106.512	3.106.512	8.736.834	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	544.731	544.731	544.731	544.397	99,94%	20	20	20	2.452.887	90,05%	378.823	69,59%	86,64%
21.09.2017	1	100,00%	40.404.349	424.011	424.011	424.011	327.922	274.117	2.576.857	2.509.748	2.503.841	2.503.841	2.503.841	8.874.420	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	521.209	521.209	521.209	520.899	99,94%	20	20	20	2.224.911	86,34%	344.448	66,13%	82,94%
22.09.2017	1	100,00%	40.404.349	385.830	385.830	385.830	298.857	252.972	2.342.038	2.710.547	2.696.535	2.696.535	2.696.535	9.014.306	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	477.020	477.020	477.020	476.625	99,92%	20	20	20	2.106.512	89,94%	327.481	68,71%	86,35%
23.09.2017	1	100,00%	40.404.349	338.055	338.055	338.055	264.145	223.775	2.172.747	2.571.213	2.559.480	2.559.480	2.559.480	9.142.159	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	430.555	430.555	430.555	430.188	99,91%	20	20	20	1.953.336	89,90%	291.099	67,67%	86,23%
24.09.2017	1	100,00%	40.404.349	407.270	407.270	407.270	318.266	270.813	2.763.817	3.102.679	3.097.922	3.097.922	3.097.922	9.298.655	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	523.106	523.106	523.106	522.784	99,94%	20	20	20	2.495.368	90,29%	359.343	68,74%	86,86%
25.09.2017	1	100,00%	40.404.349	462.419	462.419	462.419	357.460	299.910	2.858.008	3.127.958	3.113.097	3.113.097	3.113.097	9.466.306	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	567.668	567.668	567.668	567.333	99,94%	20	20	20	2.545.757	89,07%	387.806	68,36%	85,64%
26.09.2017	1	100,00%	40.404.349	451.514	451.514	451.514	349.209	292.832	2.771.203	3.228.297	3.216.984	3.216.984	3.216.984	9.635.022	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	556.916	556.916	556.916	556.578	99,94%	20	20	20	2.490.916	89,89%	381.160	68,48%	86,31%
27.09.2017	1	100,00%	40.404.349	437.060	437.060	437.060	337.414	284.611	2.687.776		3.091.303	3.091.303	3.091.303	14.892.963	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	536.529	536.529	536.529	536.142	99,93%	20	20	20	2.411.150	89,71%	367.653	68,57%	86,19%
28.09.2017	1	100,00%	40.404.349	421.240	421.240	421.240	326.305	273.428	2.560.389	2.984.309	2.978.541	2.978.541	2.978.541	9.947.425	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	517.327	517.327	517.327	516.977	99,93%	20	20	20	2.296.725	89,70%	351.994	68,09%	86,07%
29.09.2017	1	100,00%	40.404.349	375.212	375.212	375.212	291.293	245.058	2.251.789	2.626.691	2.615.129	2.615.129	2.615.129	10.082.883	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	461.750	461.750	461.750	461.315	99,91%	20	20	20	2.022.497	89,82%	309.753	67,15%	85,96%
30.09.2017	1	100,00%	40.404.349	354.944	354.944	354.944	275.329	233.312	2.257.823	2.659.132	2.653.136	2.653.136	2.653.136	10.212.715	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586	49.585.859	454.463	454.463	454.463	453.979	99,89%	20	20	20	2.029.755	89,90%	303.578	66,87%	86,04%



			WF [01]				WF [01A]		WF [01B]
sum / max / %	81	97,59%	40.404.349	34.603.557	34.615.588	34.615.588	26.778.809	22.475.420	218.817.665
	SUM	AVG %	MAX	SUM	SUM	SUM	SUM	SUM	SUM
[unit]	[days]	[%]	[session]	[session]	[session]	[session]	[session]	[FVID with linkage data]	[on hit]
Date	[00] data flow success	[00A] Successful days	[01] stg_google analytics_ext	[02] stg_googleanalytics	[03] cls_googleanalytics	[04] co_googleanalytics	[05] DISTINCT fullvisitorId of co_googleanalytics	[06] co_googleanalytics_ hits_customdim	[07] co_googleanalytics_ hits
01.10.2017	1	100,00%	27.208.748	413.187	413.187	413.187	321.043	273.352	2.808.844
02.10.2017	1	100,00%	40.404.349	433.024	433.024	433.024	333.416	280.894	2.770.654
03.10.2017	1	100,00%	40.404.349	477.396	477.396	477.396	369.109	312.292	3.298.647
04.10.2017	1	100,00%	40.404.349	459.835	459.835	459.835	354.343	292.373	2.815.842
05.10.2017	1	100,00%	40.404.349	444.339	444.339	444.339	341.678	285.619	2.764.810
06.10.2017	1	100,00%	40.404.349	410.825	410.825	410.825	317.054	262.581	2.501.196
07.10.2017	1	100,00%	40.404.349	377.205	377.205	377.205	292.614	244.617	2.480.770
08.10.2017	1	100,00%	40.404.349	457.356	457.356	457.356	354.939	298.699	3.177.164
09.10.2017	1	100,00%	40.404.349	470.478	470.478	470.478	362.319	298.963	2.944.475
10.10.2017	1	100,00%	40.404.349	463.512	463.512	463.512	355.699	296.138	2.863.961
11.10.2017	1	100,00%	40.404.349	452.504	452.504	452.504	348.859	289.189	2.770.561
12.10.2017	1	100,00%	40.404.349	421.983	421.983	421.983	325.876	272.642	2.591.902
13.10.2017	1	100,00%	40.404.349	393.003	393.003	393.003	304.485	253.243	2.370.398
14.10.2017	1	100,00%	40.404.349	339.082	339.082	339.082	264.524	220.773	2.163.452
15.10.2017	1	100,00%	40.404.349	411.689	411.689	411.689	321.889	270.983	2.777.966
16.10.2017	1	100,00%	40.404.349	455.386	455.386	455.386	353.381	292.648	2.772.745
17.10.2017	1	100,00%	40.404.349	450.476	450.476	450.476	350.491	290.615	2.738.921
18.10.2017	1	100,00%	40.404.349	439.905	439.905	439.905	340.947	283.340	2.674.878
19.10.2017	1	100,00%	40.404.349	421.308	421.308	421.308	325.617	272.399	2.551.215
20.10.2017	1	100,00%	40.404.349	399.561	399.561	399.561	308.893	258.823	2.458.148
21.10.2017	1	100,00%	40.404.349	366.233	366.233	366.233	285.713	240.669	2.395.687
22.10.2017	1	100,00%	40.404.349	460.086	460.086	460.086	358.750	302.232	3.186.904
23.10.2017	1	100,00%	40.404.349	477.237	477.237	477.237	367.375	305.234	2.997.242
24.10.2017	1	100,00%	40.404.349	465.541	465.541	465.541	356.460	300.040	2.896.752

	WF [02]					WF [03]				
sum / max / %	248.207.035	269.152.409	269.152.409	269.152.409	15.045.643	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
	SUM	SUM	SUM	SUM	MAX	MAX	MAX	MAX	MAX	MAX
[unit]	[event]	[event]	[event]	[event]	[unique visitor with uid_key]	[keyword price information]	[keyword price information]	[keyword price information]	[keyword price]	[keyword price]
Date	[08] stg_tl_eventstore _ext	[09] stg_tl_eventstore_ idkey_map	[10] cls_tl_eventstore_ idkey_map	[11] co_tl_eventstore_ idkey_map	[12] co_tl_eventstore_ idkey_ map_distinct	[13] stg_price_data_ext	[14] stg_price_data	[15] cls_pricing_data	[16] cls_pricing_data_avg	[17] co_pricing_data_avg
01.10.2017	3.263.632	3.255.311	3.255.311	3.255.311	10.371.969	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
02.10.2017	3.257.748	3.243.028	3.243.028	3.243.028	10.534.418	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
03.10.2017	3.816.281	3.812.668	3.812.668	3.812.668	10.718.389	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
04.10.2017	3.287.292	3.280.010	3.280.010	3.280.010	10.887.788	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
05.10.2017	3.324.508	3.321.578	3.321.578	3.321.578	11.056.487	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
06.10.2017	3.026.467	3.017.196	3.017.196	3.017.196	11.210.145	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
07.10.2017	2.995.925	2.981.313	2.981.313	2.981.313	11.352.483	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
08.10.2017	3.756.054	3.742.851	3.742.851	3.742.851	11.533.997	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
09.10.2017	3.530.668	3.523.808	3.523.808	3.523.808	11.717.427	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
10.10.2017		3.434.353	3.434.353	3.434.353	14.892.963	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
11.10.2017	3.348.312	3.340.174	3.340.174	3.340.174	12.062.773	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
12.10.2017	3.121.865	3.118.876	3.118.876	3.118.876	12.219.688	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
13.10.2017	2.884.313	2.880.529	2.880.529	2.880.529	12.363.355	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
14.10.2017	2.624.949	2.613.330	2.613.330	2.613.330	12.488.539	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
15.10.2017	3.309.191	3.294.618	3.294.618	3.294.618	12.647.318	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
16.10.2017	3.376.679	3.368.011	3.368.011	3.368.011	12.822.929	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
17.10.2017	3.312.079	3.305.549	3.305.549	3.305.549	12.994.294	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
18.10.2017	3.249.247	3.243.810	3.243.810	3.243.810	13.158.840	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
19.10.2017	3.100.149	3.091.790	3.091.790	3.091.790	13.314.048	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
20.10.2017	2.985.784	2.979.140	2.979.140	2.979.140	13.458.805	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
21.10.2017	2.892.229	2.886.577	2.886.577	2.886.577	13.593.582	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
22.10.2017	3.772.509	3.764.634	3.764.634	3.764.634	13.770.928	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
23.10.2017	3.601.883	3.591.036	3.591.036	3.591.036	13.952.441	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586
24.10.2017	3.484.076	3.480.078	3.480.078	3.480.078	14.122.945	4.628.567	4.628.567	4.628.567	4.272.586	4.272.586

	WF [04]				WF [04A]		WF [05]			WF [07]				
sum / max / %	49.585.859	43.275.972	43.275.972	43.275.972	43.243.252	99,92%	20	20	20	195.992.440	89,51%	29.259.855	67,61%	85,89%
	MAX	SUM	SUM	SUM	SUM	AVG %	MAX	MAX	MAX	SUM	AVG %	SUM	AVG %	AVG %
[unit]	[off hit]	[off hit]	[off hit]	[off hit]	[cost data]	[%]	[product]	[product]	[product]	[ON-page hits]	[%]	[off-page hits]	[%]	[%]
Date	[18] stg_intelliad_click_ report_ext	[19] stg_intelliad_ click_report	[20] cls_intelliad_ click_report	[21] co_intelliad_ click_report	[22] co_intelliad_click_ report_with_prices	[22A] price mappings ([22]/[21])	[23] stg_tl_product_ prices	[24] cls_tl_product_prices	[25] co_tl_product_ prices	[26] co_final_customer _journey [ONPAGE]	[26A] used ON hits ([26] - [07])	[27] co_final_customer _journey [OFFPAGE]	[27A] used OFF hits ([27]/[22])	[27B] used hits over all ([26A] + [27A] / [07] + [22])
01.10.2017	49.585.859	529.934	529.934	529.934	529.499	99,92%	20	20	20	2.526.645	89,95%	357.913	67,59%	86,41%
02.10.2017	49.585.859	541.378	541.378	541.378	540.891	99,91%	20	20	20	2.484.935	89,69%	364.171	67,33%	86,04%
03.10.2017	49.585.859	615.091	615.091	615.091	614.658	99,93%	20	20	20	2.965.773	89,91%	415.907	67,66%	86,41%
04.10.2017	49.585.859	563.400	563.400	563.400	563.012	99,93%	20	20	20	2.518.992	89,46%	377.665	67,08%	85,73%
05.10.2017	49.585.859	561.411	561.411	561.411	561.040	99,93%	20	20	20	2.477.034	89,59%	372.168	66,34%	85,67%
06.10.2017	49.585.859	519.636	519.636	519.636	519.184	99,91%	20	20	20	2.235.226	89,37%	338.155	65,13%	85,20%
07.10.2017	49.585.859	486.843	486.843	486.843	486.305	99,89%	20	20	20	2.222.438	89,59%	314.849	64,74%	85,51%
08.10.2017	49.585.859	597.941	597.941	597.941	597.522	99,93%	20	20	20	2.855.376	89,87%	394.767	66,07%	86,10%
09.10.2017	49.585.859	590.295	590.295	590.295	589.899	99,93%	20	20	20	2.637.392	89,57%	386.892	65,59%	85,57%
10.10.2017	49.585.859	583.838	583.838	583.838	583.485	99,94%	20	20	20	2.560.554	89,41%	384.392	65,88%	85,42%
11.10.2017	49.585.859	570.053	570.053	570.053	569.663	99,93%	20	20	20	2.478.099	89,44%	372.224	65,34%	85,33%
12.10.2017	49.585.859	529.147	529.147	529.147	528.777	99,93%	20	20	20	2.317.742	89,42%	346.191	65,47%	85,36%
13.10.2017	49.585.859	499.826	499.826	499.826	499.368	99,91%	20	20	20	2.116.365	89,28%	322.583	64,60%	84,99%
14.10.2017	49.585.859	438.668	438.668	438.668	438.182	99,89%	20	20	20	1.936.526	89,51%	279.639	63,82%	85,18%
15.10.2017	49.585.859	535.147	535.147	535.147	534.798	99,93%	20	20	20	2.487.997	89,56%	350.193	65,48%	85,67%
16.10.2017	49.585.859	576.110	576.110	576.110	575.795	99,95%	20	20	20	2.474.912	89,26%	374.865	65,10%	85,11%
17.10.2017	49.585.859	570.247	570.247	570.247	569.918	99,94%	20	20	20	2.439.802	89,08%	370.203	64,96%	84,92%
18.10.2017	49.585.859	554.128	554.128	554.128	553.776	99,94%	20	20	20	2.386.047	89,20%	360.416	65,08%	85,07%
19.10.2017	49.585.859	528.997	528.997	528.997	528.688	99,94%	20	20	20	2.279.729	89,36%	346.052	65,45%	85,26%
20.10.2017	49.585.859	506.914	506.914	506.914	506.506	99,92%	20	20	20	2.189.551	89,07%	328.299	64,82%	84,93%
21.10.2017	49.585.859	476.707	476.707	476.707	476.256	99,91%	20	20	20	2.136.527	89,18%	306.647	64,39%	85,07%
22.10.2017	49.585.859	604.250	604.250	604.250	603.842	99,93%	20	20	20	2.852.815	89,52%	395.764	65,54%	85,70%
23.10.2017	49.585.859	602.344	602.344	602.344	601.968	99,94%	20	20	20	2.668.900	89,05%	393.147	65,31%	85,08%
24.10.2017	49.585.859	589.703	589.703	589.703	589.358	99,94%	20	20	20	2.585.373	89,25%	385.882	65,47%	85,23%

Appendix 6: WF[01]_stage_run

```
----------------------------------------------------------------
-- Workflow [01]: GOOGLE_ANALYTICS_STAGE_2_CORE
-- FILE: df_stg_google_analytics.sql
-- AREA: stage
----------------------------------------------------------------


SET hive.exec.dynamic.partition.mode=nonstrict;
SET hive.execution.engine=mr;
DROP   TABLE dip_immonet_stage.stg_googleanalytics;
CREATE TABLE dip_immonet_stage.stg_googleanalytics
STORED AS PARQUET TBLPROPERTIES ('PARQUET.COMPRESS'='SNAPPY')
AS
SELECT json_string                                              AS json_string,
       SUBSTR(CAST(from_unixtime(unix_timestamp(get_json_object(json_string, '$.date') , 'yyyyMMdd')) AS string),1,10) AS file_date,
       REGEXP_EXTRACT(INPUT__FILE__NAME, '.*/(.*)/(.*)', 2)     AS file_name
FROM   dip_immonet_stage.stg_googleanalytics_ext
WHERE  file_date =  CAST(CONCAT(SUBSTR(${DATE},1,8),"01") AS DATE)
AND       SUBSTR(CAST(from_unixtime(unix_timestamp(get_json_object(json_string, '$.date') , 'yyyyMMdd')) AS string),1,10)  = ${DATE} ;
```

Appendix 7: WF[01]_cleanse_run

```sql
-----------------------------------------------------------------
-- Workflow [01]: GOOGLE_ANALYTICS_STAGE_2_CORE
-- FILE: df_cls_google_analytics.sql
-- AREA: cleanse
-----------------------------------------------------------------

SET hive.exec.dynamic.partition.mode=nonstrict;
SET hive.execution.engine=mr;
DROP   TABLE dip_portal_cleanse.cls_googleanalytics;
CREATE TABLE dip_portal_cleanse.cls_googleanalytics
STORED AS PARQUET TBLPROPERTIES ('PARQUET.COMPRESS'='SNAPPY')
AS
SELECT get_json_object(json_string, '$.visitorId')                AS `visitorId`,
get_json_object(json_string, '$.visitNumber')                     AS `visitNumber`,
get_json_object(json_string, '$.visitId')                         AS `visitId`,
get_json_object(json_string, '$.visitStartTime')                  AS `visitStartTime`,
get_json_object(json_string, '$.date')                            AS `date`,
get_json_object(json_string, '$.totals')                          AS `totals`,
get_json_object(json_string, '$.totals.visits')                   AS `totals_visits`,
get_json_object(json_string, '$.totals.hits')                     AS `totals_hits`,
get_json_object(json_string, '$.totals.pageviews')                AS `totals_pageviews`,
get_json_object(json_string, '$.totals.timeOnSite')               AS `totals_timeOnSite`,
get_json_object(json_string, '$.totals.bounces')                  AS `totals_bounces`,
get_json_object(json_string, '$.totals.transactions')             AS `totals_transactions`,
get_json_object(json_string, '$.totals.transactionRevenue')       AS `totals_transactionRevenue`,
get_json_object(json_string, '$.totals.newVisits')                AS `totals_newVisits`,
get_json_object(json_string, '$.totals.screenviews')              AS `totals_screenviews`,
get_json_object(json_string, '$.totals.uniqueScreenviews')        AS `totals_uniqueScreenviews`,
get_json_object(json_string, '$.totals.timeOnScreen')             AS `totals_timeOnScreen`,
get_json_object(json_string, '$.totals.totalTransactionRevenue')  AS `totals_totalTransactionRevenue`,
get_json_object(json_string, '$.totals.sessionQualityDim')        AS `totals_sessionQualityDim`,
get_json_object(json_string, '$.trafficSource')                   AS `trafficSource`,
get_json_object(json_string, '$.trafficSource.referralPath')      AS `trafficSource_referralPath`,
get_json_object(json_string, '$.trafficSource.campaign')          AS `trafficSource_campaign`,
get_json_object(json_string, '$.trafficSource.source')            AS `trafficSource_source`,
get_json_object(json_string, '$.trafficSource.medium')            AS `trafficSource_medium`,
get_json_object(json_string, '$.trafficSource.keyword')           AS `trafficSource_keyword`,
get_json_object(json_string, '$.trafficSource.adContent')         AS `trafficSource_adContent`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo')                        AS
`trafficSource_adwordsClickInfo`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.campaignId')             AS
`trafficSource_adwordsClickInfo_campaignId`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.adGroupId')              AS
`trafficSource_adwordsClickInfo_adGroupId`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.creativeId')             AS
`trafficSource_adwordsClickInfo_creativeId`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.criteriaId')             AS
`trafficSource_adwordsClickInfo_criteriaId`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.page')                   AS
`trafficSource_adwordsClickInfo_page`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.slot')                   AS
`trafficSource_adwordsClickInfo_slot`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.criteriaParameters')     AS
`trafficSource_adwordsClickInfo_criteriaParameters`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.gclId')                  AS
`trafficSource_adwordsClickInfo_gclId`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.customerId')             AS
`trafficSource_adwordsClickInfo_customerId`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.adNetworkType')          AS
`trafficSource_adwordsClickInfo_adNetworkType`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.targetingCriteria')      AS
`trafficSource_adwordsClickInfo_targetingCriteria`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.targetingCriteria.boomUserlistId')

    AS `trafficSource_adwordsClickInfo_targetingCriteria_boomUserlistId`,
get_json_object(json_string,
'$.trafficSource.adwordsClickInfo.isVideoAd')              AS
`trafficSource_adwordsClickInfo_isVideoAd`,
get_json_object(json_string,
'$.trafficSource.isTrueDirect')                            AS
`trafficSource_isTrueDirect`,
get_json_object(json_string,
'$.trafficSource.campaignCode')                            AS
`trafficSource_campaignCode`,
get_json_object(json_string,
'$.device')                                                AS `device`,
get_json_object(json_string,
'$.device.browser')                                        AS `device_browser`,
get_json_object(json_string,
'$.device.browserVersion')                                 AS `device_browserVersion`,
get_json_object(json_string,
'$.device.browserSize')                                    AS `device_browserSize`,
get_json_object(json_string,
'$.device.operatingSystem')                                AS
`device_operatingSystem`,
get_json_object(json_string,
'$.device.operatingSystemVersion')                         AS
`device_operatingSystemVersion`,
get_json_object(json_string,
'$.device.isMobile')                                       AS `device_isMobile`,
get_json_object(json_string,
'$.device.mobileDeviceBranding')                           AS
`device_mobileDeviceBranding`,
get_json_object(json_string,
'$.device.mobileDeviceModel')                              AS
`device_mobileDeviceModel`,
get_json_object(json_string, '$.device.mobileInputSelector')            AS `device_mobileInputSelector`,
get_json_object(json_string, '$.device.mobileDeviceInfo')               AS `device_mobileDeviceInfo`,
get_json_object(json_string, '$.device.mobileDeviceMarketingName')      AS `device_mobileDeviceMarketingName`,
get_json_object(json_string, '$.device.flAShVersion')                   AS `device_flAShVersion`,
get_json_object(json_string, '$.device.javaEnabled')                    AS `device_javaEnabled`,
get_json_object(json_string, '$.device.language')                       AS `device_language`,
get_json_object(json_string, '$.device.screenColors')                   AS `device_screenColors`,
get_json_object(json_string, '$.device.screenResolution')               AS `device_screenResolution`,
get_json_object(json_string, '$.device.deviceCategory')                 AS `device_deviceCategory`,
get_json_object(json_string, '$.geoNetwork')                            AS `geoNetwork`,
get_json_object(json_string, '$.geoNetwork.continent')                  AS `geoNetwork_continent`,
get_json_object(json_string, '$.geoNetwork.subContinent')               AS `geoNetwork_subContinent`,
get_json_object(json_string, '$.geoNetwork.country')                    AS `geoNetwork_country`,
get_json_object(json_string, '$.geoNetwork.region')                     AS `geoNetwork_region`,
get_json_object(json_string, '$.geoNetwork.metro')                      AS `geoNetwork_metro`,
get_json_object(json_string, '$.geoNetwork.city')                       AS `geoNetwork_city`,
get_json_object(json_string, '$.geoNetwork.cityId')                     AS `geoNetwork_cityId`,
get_json_object(json_string, '$.geoNetwork.networkDomain')              AS `geoNetwork_networkDomain`,
get_json_object(json_string, '$.geoNetwork.latitude')                   AS `geoNetwork_latitude`,
get_json_object(json_string, '$.geoNetwork.longitude')                  AS `geoNetwork_longitude`,
get_json_object(json_string, '$.geoNetwork.networkLocation')            AS `geoNetwork_networkLocation`,
get_json_object(json_string, '$.customDimensions')                      AS `customDimensions`,
get_json_object(json_string, '$.customDimensions.index')                AS `customDimensions_index`,
get_json_object(json_string, '$.customDimensions.value')                AS `customDimensions_value`,
get_json_object(json_string, '$.hits')                                  AS `hits`,
get_json_object(json_string, '$.hits.hitNumber')                        AS `hits_hitNumber`,
get_json_object(json_string, '$.hits.time')                             AS `hits_time`,
get_json_object(json_string, '$.hits.hour')                             AS `hits_hour`,
get_json_object(json_string, '$.hits.minute')                           AS `hits_minute`,
get_json_object(json_string, '$.hits.isSecure')                         AS `hits_isSecure`,
get_json_object(json_string, '$.hits.isInteraction')                    AS `hits_isInteraction`,
get_json_object(json_string,
'$.hits.isEntrance')                                    AS `hits_isEntrance`,
get_json_object(json_string,
'$.hits.isExit')                                        AS `hits_isExit`,
get_json_object(json_string,
'$.hits.referer')                                       AS `hits_referer`,
get_json_object(json_string,
'$.hits.page')                                          AS `hits_page`,
get_json_object(json_string,
'$.hits.page.pagePath')                                 AS `hits_page_pagePath`,
get_json_object(json_string,
'$.hits.page.hostname')                                 AS `hits_page_hostname`,
get_json_object(json_string,
'$.hits.page.pageTitle')                                AS `hits_page_pageTitle`,
get_json_object(json_string,
'$.hits.page.searchKeyword')                            AS
`hits_page_searchKeyword`,
get_json_object(json_string,
'$.hits.page.searchCategory')                           AS
`hits_page_searchCategory`,
get_json_object(json_string,
'$.hits.page.pagePathLevel1')                           AS
`hits_page_pagePathLevel1`,
get_json_object(json_string,
'$.hits.page.pagePathLevel2')                           AS
`hits_page_pagePathLevel2`,
get_json_object(json_string,
'$.hits.page.pagePathLevel3')                           AS
`hits_page_pagePathLevel3`,
get_json_object(json_string,
'$.hits.page.pagePathLevel4')                           AS
`hits_page_pagePathLevel4`,
get_json_object(json_string,
'$.hits.transaction')                                   AS `hits_transaction`,
get_json_object(json_string,
'$.hits.transaction.transactionId')                     AS
`hits_transaction_transactionId`,
get_json_object(json_string,
'$.hits.transaction.transactionRevenue')                AS
`hits_transaction_transactionRevenue`,
get_json_object(json_string,
'$.hits.transaction.transactionTax')                    AS
`hits_transaction_transactionTax`,
get_json_object(json_string,
'$.hits.transaction.transactionShipping')               AS
`hits_transaction_transactionShipping`,
get_json_object(json_string,
'$.hits.transaction.affiliation')                       AS
`hits_transaction_affiliation`,
get_json_object(json_string,
'$.hits.transaction.currencyCode')                      AS
`hits_transaction_currencyCode`,
get_json_object(json_string,
'$.hits.transaction.localTransactionRevenue')           AS
`hits_transaction_localTransactionRevenue`,
get_json_object(json_string,
'$.hits.transaction.localTransactionTax')               AS
`hits_transaction_localTransactionTax`,
get_json_object(json_string,
'$.hits.transaction.localTransactionShipping')          AS
`hits_transaction_localTransactionShipping`,
get_json_object(json_string,
'$.hits.transaction.transactionCoupon')                 AS
`hits_transaction_transactionCoupon`,
get_json_object(json_string,
'$.hits.item')                                          AS `hits_item`,
get_json_object(json_string,
'$.hits.item.transactionId')                            AS
`hits_item_transactionId`,
get_json_object(json_string,
'$.hits.item.productName')                              AS `hits_item_productName`,
get_json_object(json_string, '$.hits.item.productCategory') AS `hits_item_productCategory`,
get_json_object(json_string, '$.hits.item.productSku') AS `hits_item_productSku`,
get_json_object(json_string, '$.hits.item.itemQuantity') AS `hits_item_itemQuantity`,
get_json_object(json_string, '$.hits.item.itemRevenue') AS `hits_item_itemRevenue`,
get_json_object(json_string, '$.hits.item.currencyCode') AS `hits_item_currencyCode`,
get_json_object(json_string, '$.hits.item.localItemRevenue') AS `hits_item_localItemRevenue`,
get_json_object(json_string, '$.hits.contentInfo') AS `hits_contentInfo`,
get_json_object(json_string, '$.hits.contentInfo.contentDescription') AS `hits_contentInfo_contentDescription`,
get_json_object(json_string, '$.hits.appInfo') AS `hits_appInfo`,
get_json_object(json_string, '$.hits.appInfo.name') AS `hits_appInfo_name`,
get_json_object(json_string, '$.hits.appInfo.version') AS `hits_appInfo_version`,
get_json_object(json_string, '$.hits.appInfo.id') AS `hits_appInfo_id`,
get_json_object(json_string, '$.hits.appInfo.installerId') AS `hits_appInfo_installerId`,
get_json_object(json_string, '$.hits.appInfo.appInstallerId') AS `hits_appInfo_appInstallerId`,
get_json_object(json_string, '$.hits.appInfo.appName') AS `hits_appInfo_appName`,
get_json_object(json_string, '$.hits.appInfo.appVersion') AS `hits_appInfo_appVersion`,
get_json_object(json_string, '$.hits.appInfo.appId') AS `hits_appInfo_appId`,
get_json_object(json_string, '$.hits.appInfo.screenName') AS `hits_appInfo_screenName`,
get_json_object(json_string, '$.hits.appInfo.landingScreenName') AS `hits_appInfo_landingScreenName`,
get_json_object(json_string, '$.hits.appInfo.exitScreenName') AS `hits_appInfo_exitScreenName`,
get_json_object(json_string, '$.hits.appInfo.screenDepth') AS `hits_appInfo_screenDepth`,
get_json_object(json_string, '$.hits.exceptionInfo') AS `hits_exceptionInfo`,
get_json_object(json_string, '$.hits.exceptionInfo.description') AS `hits_exceptionInfo_description`,
get_json_object(json_string, '$.hits.exceptionInfo.isFatal') AS `hits_exceptionInfo_isFatal`,
get_json_object(json_string, '$.hits.exceptionInfo.exceptions') AS `hits_exceptionInfo_exceptions`,
get_json_object(json_string, '$.hits.exceptionInfo.fatalExceptions') AS `hits_exceptionInfo_fatalExceptions`,
get_json_object(json_string, '$.hits.eventInfo') AS `hits_eventInfo`,
get_json_object(json_string,
'$.hits.eventInfo.eventCategory')                                AS
`hits_eventInfo_eventCategory`,
get_json_object(json_string,
'$.hits.eventInfo.eventAction')                                  AS
`hits_eventInfo_eventAction`,
get_json_object(json_string,
'$.hits.eventInfo.eventLabel')                                   AS
`hits_eventInfo_eventLabel`,
get_json_object(json_string,
'$.hits.eventInfo.eventValue')                                   AS
`hits_eventInfo_eventValue`,
get_json_object(json_string,
'$.hits.product')                                                AS `hits_product`,
get_json_object(json_string,
'$.hits.product.productSKU')                                     AS
`hits_product_productSKU`,
get_json_object(json_string,
'$.hits.product.v2ProductName')                                  AS
`hits_product_v2ProductName`,
get_json_object(json_string,
'$.hits.product.v2ProductCategory')                              AS
`hits_product_v2ProductCategory`,
get_json_object(json_string,
'$.hits.product.productVariant')                                 AS
`hits_product_productVariant`,
get_json_object(json_string,
'$.hits.product.productBrand')                                   AS
`hits_product_productBrand`,
get_json_object(json_string,
'$.hits.product.productRevenue')                                 AS
`hits_product_productRevenue`,
get_json_object(json_string,
'$.hits.product.localProductRevenue')                            AS
`hits_product_localProductRevenue`,
get_json_object(json_string,
'$.hits.product.productPrice')                                   AS
`hits_product_productPrice`,
get_json_object(json_string,
'$.hits.product.localProductPrice')                              AS
`hits_product_localProductPrice`,
get_json_object(json_string,
'$.hits.product.productQuantity')                                AS
`hits_product_productQuantity`,
get_json_object(json_string,
'$.hits.product.productRefundAmount')                            AS
`hits_product_productRefundAmount`,
get_json_object(json_string,
'$.hits.product.localProductRefundAmount')                       AS
`hits_product_localProductRefundAmount`,
get_json_object(json_string,
'$.hits.product.isImpression')                                   AS
`hits_product_isImpression`,
get_json_object(json_string,
'$.hits.product.isClick')                                        AS `hits_product_isClick`,
get_json_object(json_string,
'$.hits.product.customDimensions')                               AS
`hits_product_customDimensions`,
get_json_object(json_string,
'$.hits.product.customDimensions.index')                         AS
`hits_product_customDimensions_index`,
get_json_object(json_string,
'$.hits.product.customDimensions.value')                         AS
`hits_product_customDimensions_value`,
get_json_object(json_string,
'$.hits.product.customMetrics')                                  AS
`hits_product_customMetrics`,
get_json_object(json_string,
'$.hits.product.customMetrics.index')                            AS
`hits_product_customMetrics_index`,
get_json_object(json_string,
'$.hits.product.customMetrics.value')                            AS
```

```
`hits_product_customMetrics_value`,
get_json_object(json_string, '$.hits.product.productListName') AS `hits_product_productListName`,
get_json_object(json_string, '$.hits.product.productListPosition') AS `hits_product_productListPosition`,
get_json_object(json_string, '$.hits.promotion') AS `hits_promotion`,
get_json_object(json_string, '$.hits.promotion.promoId') AS `hits_promotion_promoId`,
get_json_object(json_string, '$.hits.promotion.promoName') AS `hits_promotion_promoName`,
get_json_object(json_string, '$.hits.promotion.promoCreative') AS `hits_promotion_promoCreative`,
get_json_object(json_string, '$.hits.promotion.promoPosition') AS `hits_promotion_promoPosition`,
get_json_object(json_string, '$.hits.promotionActionInfo') AS `hits_promotionActionInfo`,
get_json_object(json_string, '$.hits.promotionActionInfo.promoIsView') AS `hits_promotionActionInfo_promoIsView`,
get_json_object(json_string, '$.hits.promotionActionInfo.promoIsClick') AS `hits_promotionActionInfo_promoIsClick`,
get_json_object(json_string, '$.hits.refund') AS `hits_refund`,
get_json_object(json_string, '$.hits.refund.refundAmount') AS `hits_refund_refundAmount`,
get_json_object(json_string, '$.hits.refund.localRefundAmount') AS `hits_refund_localRefundAmount`,
get_json_object(json_string, '$.hits.eCommerceAction') AS `hits_eCommerceAction`,
get_json_object(json_string, '$.hits.eCommerceAction.action_type') AS `hits_eCommerceAction_action_type`,
get_json_object(json_string, '$.hits.eCommerceAction.step') AS `hits_eCommerceAction_step`,
get_json_object(json_string, '$.hits.eCommerceAction.option') AS `hits_eCommerceAction_option`,
get_json_object(json_string, '$.hits.experiment') AS `hits_experiment`,
get_json_object(json_string, '$.hits.experiment.experimentId') AS `hits_experiment_experimentId`,
get_json_object(json_string, '$.hits.experiment.experimentVariant') AS `hits_experiment_experimentVariant`,
get_json_object(json_string, '$.hits.publisher') AS `hits_publisher`,
get_json_object(json_string, '$.hits.publisher.dfpClicks') AS `hits_publisher_dfpClicks`,
get_json_object(json_string, '$.hits.publisher.dfpImpressions') AS `hits_publisher_dfpImpressions`,
get_json_object(json_string, '$.hits.publisher.dfpMatchedQueries') AS `hits_publisher_dfpMatchedQueries`,
get_json_object(json_string, '$.hits.publisher.dfpMeASurableImpressions') AS `hits_publisher_dfpMeASurableImpressions`,
get_json_object(json_string,
 '$.hits.publisher.dfpQueries')                            AS
 `hits_publisher_dfpQueries`,
get_json_object(json_string,
 '$.hits.publisher.dfpRevenueCpm')                         AS
 `hits_publisher_dfpRevenueCpm`,
get_json_object(json_string,
 '$.hits.publisher.dfpRevenueCpc')                         AS
 `hits_publisher_dfpRevenueCpc`,
get_json_object(json_string,
 '$.hits.publisher.dfpViewableImpressions')                AS
 `hits_publisher_dfpViewableImpressions`,
get_json_object(json_string,
 '$.hits.publisher.dfpPagesViewed')                        AS
 `hits_publisher_dfpPagesViewed`,
get_json_object(json_string,
 '$.hits.publisher.adsenseBackfillDfpClicks')              AS
 `hits_publisher_adsenseBackfillDfpClicks`,
get_json_object(json_string,
 '$.hits.publisher.adsenseBackfillDfpImpressions')         AS
 `hits_publisher_adsenseBackfillDfpImpressions`,
get_json_object(json_string,
 '$.hits.publisher.adsenseBackfillDfpMatchedQueries')      AS
 `hits_publisher_adsenseBackfillDfpMatchedQueries`,
get_json_object(json_string,
 '$.hits.publisher.adsenseBackfillDfpMeASurableImpressions')

    AS `hits_publisher_adsenseBackfillDfpMeASurableImpressions`,
get_json_object(json_string,
 '$.hits.publisher.adsenseBackfillDfpQueries')             AS
 `hits_publisher_adsenseBackfillDfpQueries`,
get_json_object(json_string,
 '$.hits.publisher.adsenseBackfillDfpRevenueCpm')          AS
 `hits_publisher_adsenseBackfillDfpRevenueCpm`,
get_json_object(json_string,
 '$.hits.publisher.adsenseBackfillDfpRevenueCpc')          AS
 `hits_publisher_adsenseBackfillDfpRevenueCpc`,
get_json_object(json_string,
 '$.hits.publisher.adsenseBackfillDfpViewableImpressions') AS
 `hits_publisher_adsenseBackfillDfpViewableImpressions`,
get_json_object(json_string,
 '$.hits.publisher.adsenseBackfillDfpPagesViewed')         AS
 `hits_publisher_adsenseBackfillDfpPagesViewed`,
get_json_object(json_string,
 '$.hits.publisher.adxBackfillDfpClicks')                  AS
 `hits_publisher_adxBackfillDfpClicks`,
get_json_object(json_string,
 '$.hits.publisher.adxBackfillDfpImpressions')             AS
 `hits_publisher_adxBackfillDfpImpressions`,
get_json_object(json_string,
 '$.hits.publisher.adxBackfillDfpMatchedQueries')          AS
 `hits_publisher_adxBackfillDfpMatchedQueries`,
get_json_object(json_string,
 '$.hits.publisher.adxBackfillDfpMeASurableImpressions')   AS
 `hits_publisher_adxBackfillDfpMeASurableImpressions`,
get_json_object(json_string,
 '$.hits.publisher.adxBackfillDfpQueries')                 AS
 `hits_publisher_adxBackfillDfpQueries`,
get_json_object(json_string,
 '$.hits.publisher.adxBackfillDfpRevenueCpm')              AS
 `hits_publisher_adxBackfillDfpRevenueCpm`,
get_json_object(json_string,
 '$.hits.publisher.adxBackfillDfpRevenueCpc')              AS
 `hits_publisher_adxBackfillDfpRevenueCpc`,
get_json_object(json_string,
 '$.hits.publisher.adxBackfillDfpViewableImpressions')     AS
 `hits_publisher_adxBackfillDfpViewableImpressions`,
get_json_object(json_string,
 '$.hits.publisher.adxBackfillDfpPagesViewed')             AS
 `hits_publisher_adxBackfillDfpPagesViewed`,
get_json_object(json_string,
```

^^^v

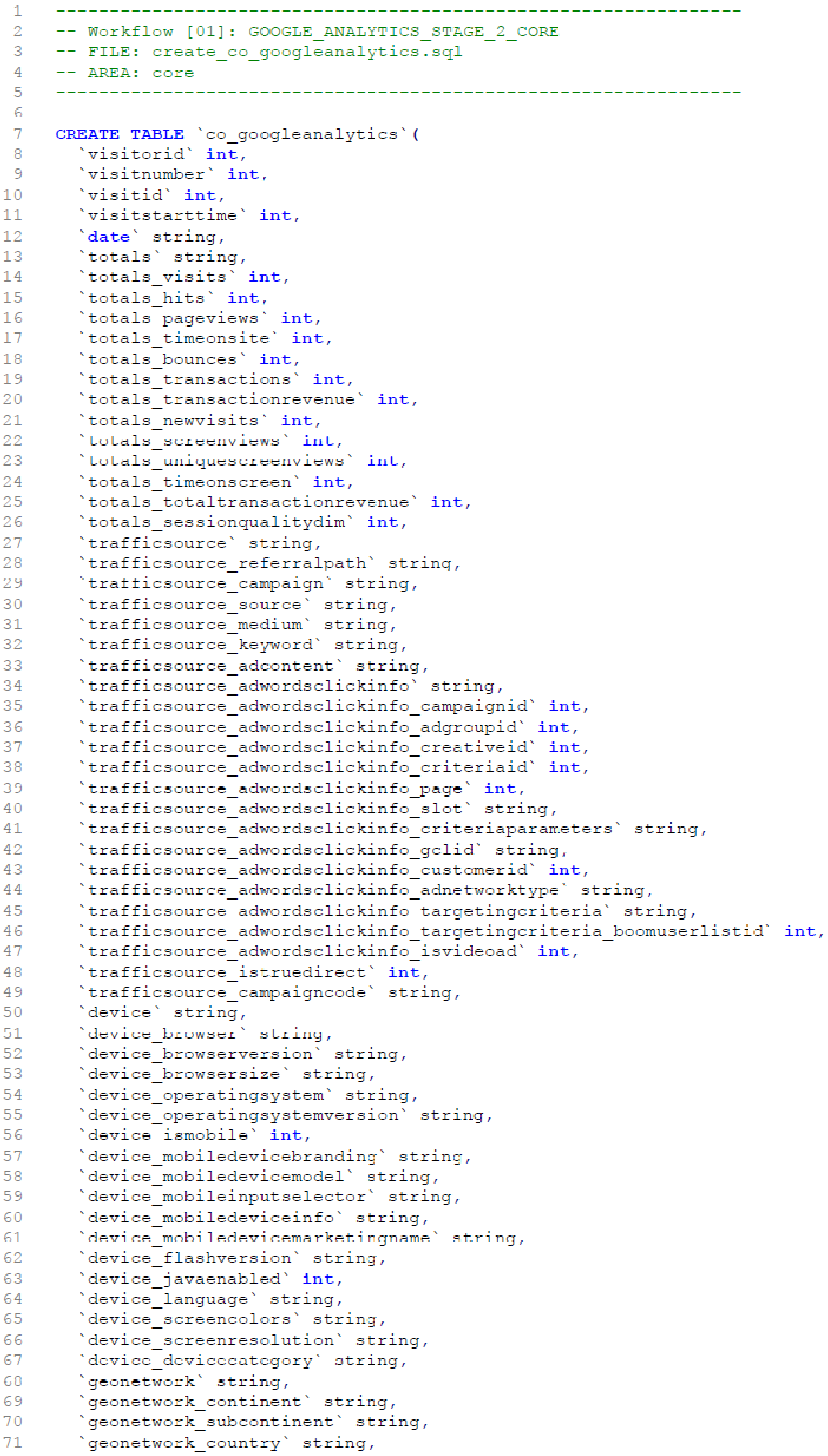
```
'$.hits.publisher.adxClicks')                               AS
`hits_publisher_adxClicks`,
get_json_object(json_string,
'$.hits.publisher.adxImpressions')                          AS
`hits_publisher_adxImpressions`,
get_json_object(json_string,
'$.hits.publisher.adxMatchedQueries')                       AS
`hits_publisher_adxMatchedQueries`,
get_json_object(json_string,
'$.hits.publisher.adxMeASurableImpressions')                AS
`hits_publisher_adxMeASurableImpressions`,
get_json_object(json_string,
'$.hits.publisher.adxQueries')                              AS
`hits_publisher_adxQueries`,
get_json_object(json_string,
'$.hits.publisher.adxRevenue')                              AS
`hits_publisher_adxRevenue`,
get_json_object(json_string,
'$.hits.publisher.adxViewableImpressions')                  AS
`hits_publisher_adxViewableImpressions`,
get_json_object(json_string,
'$.hits.publisher.adxPagesViewed')                          AS
`hits_publisher_adxPagesViewed`,
get_json_object(json_string,
'$.hits.publisher.adsViewed')                               AS
`hits_publisher_adsViewed`,
get_json_object(json_string,
'$.hits.publisher.adsUnitsViewed')                          AS
`hits_publisher_adsUnitsViewed`,
get_json_object(json_string,
'$.hits.publisher.adsUnitsMatched')                         AS
`hits_publisher_adsUnitsMatched`,
get_json_object(json_string,
'$.hits.publisher.viewableAdsViewed')                       AS
`hits_publisher_viewableAdsViewed`,
get_json_object(json_string,
'$.hits.publisher.meASurableAdsViewed')                     AS
`hits_publisher_meASurableAdsViewed`,
get_json_object(json_string,
'$.hits.publisher.adsPagesViewed')                          AS
`hits_publisher_adsPagesViewed`,
get_json_object(json_string,
'$.hits.publisher.adsClicked')                              AS
`hits_publisher_adsClicked`,
get_json_object(json_string,
'$.hits.publisher.adsRevenue')                              AS
`hits_publisher_adsRevenue`,
get_json_object(json_string,
'$.hits.publisher.dfpAdGroup')                              AS
`hits_publisher_dfpAdGroup`,
get_json_object(json_string,
'$.hits.publisher.dfpAdUnits')                              AS
`hits_publisher_dfpAdUnits`,
get_json_object(json_string,
'$.hits.publisher.dfpNetworkId')                            AS
`hits_publisher_dfpNetworkId`,
get_json_object(json_string,
'$.hits.customVariables')                                   AS `hits_customVariables`,
get_json_object(json_string,
'$.hits.customVariables.index')                             AS
`hits_customVariables_index`,
get_json_object(json_string,
'$.hits.customVariables.customVarName')                     AS
`hits_customVariables_customVarName`,
get_json_object(json_string,
'$.hits.customVariables.customVarValue')                    AS
`hits_customVariables_customVarValue`,
get_json_object(json_string,
'$.hits.customDimensions')                                  AS `hits_customDimensions`,
get_json_object(json_string,
'$.hits.customDimensions.index')                            AS
```

```
`hits_customDimensions_index`,
get_json_object(json_string,
'$.hits.customDimensions.value')                          AS
`hits_customDimensions_value`,
get_json_object(json_string,
'$.hits.customMetrics')                                   AS `hits_customMetrics`,
get_json_object(json_string,
'$.hits.customMetrics.index')                             AS
`hits_customMetrics_index`,
get_json_object(json_string,
'$.hits.customMetrics.value')                             AS
`hits_customMetrics_value`,
get_json_object(json_string,
'$.hits.type')                                            AS `hits_type`,
get_json_object(json_string,
'$.hits.social')                                          AS `hits_social`,
get_json_object(json_string,
'$.hits.social.socialInteractionNetwork')                 AS
`hits_social_socialInteractionNetwork`,
get_json_object(json_string,
'$.hits.social.socialInteractionAction')                  AS
`hits_social_socialInteractionAction`,
get_json_object(json_string,
'$.hits.social.socialInteractions')                       AS
`hits_social_socialInteractions`,
get_json_object(json_string,
'$.hits.social.socialInteractionTarget')                  AS
`hits_social_socialInteractionTarget`,
get_json_object(json_string,
'$.hits.social.socialNetwork')                            AS
`hits_social_socialNetwork`,
get_json_object(json_string,
'$.hits.social.uniqueSocialInteractions')                 AS
`hits_social_uniqueSocialInteractions`,
get_json_object(json_string,
'$.hits.social.hASSocialSourceReferral')                  AS
`hits_social_hASSocialSourceReferral`,
get_json_object(json_string,
'$.hits.social.socialInteractionNetworkAction')           AS
`hits_social_socialInteractionNetworkAction`,
get_json_object(json_string,
'$.hits.latencyTracking')                                 AS `hits_latencyTracking`,
get_json_object(json_string,
'$.hits.latencyTracking.pageLoadSample')                  AS
`hits_latencyTracking_pageLoadSample`,
get_json_object(json_string,
'$.hits.latencyTracking.pageLoadTime')                    AS
`hits_latencyTracking_pageLoadTime`,
get_json_object(json_string,
'$.hits.latencyTracking.pageDownloadTime')                AS
`hits_latencyTracking_pageDownloadTime`,
get_json_object(json_string,
'$.hits.latencyTracking.redirectionTime')                 AS
`hits_latencyTracking_redirectionTime`,
get_json_object(json_string,
'$.hits.latencyTracking.speedMetricsSample')              AS
`hits_latencyTracking_speedMetricsSample`,
get_json_object(json_string,
'$.hits.latencyTracking.domainLookupTime')                AS
`hits_latencyTracking_domainLookupTime`,
get_json_object(json_string,
'$.hits.latencyTracking.serverConnectionTime')            AS
`hits_latencyTracking_serverConnectionTime`,
get_json_object(json_string,
'$.hits.latencyTracking.serverResponseTime')              AS
`hits_latencyTracking_serverResponseTime`,
get_json_object(json_string,
'$.hits.latencyTracking.domLatencyMetricsSample')         AS
`hits_latencyTracking_domLatencyMetricsSample`,
get_json_object(json_string,
'$.hits.latencyTracking.domInteractiveTime')              AS
```

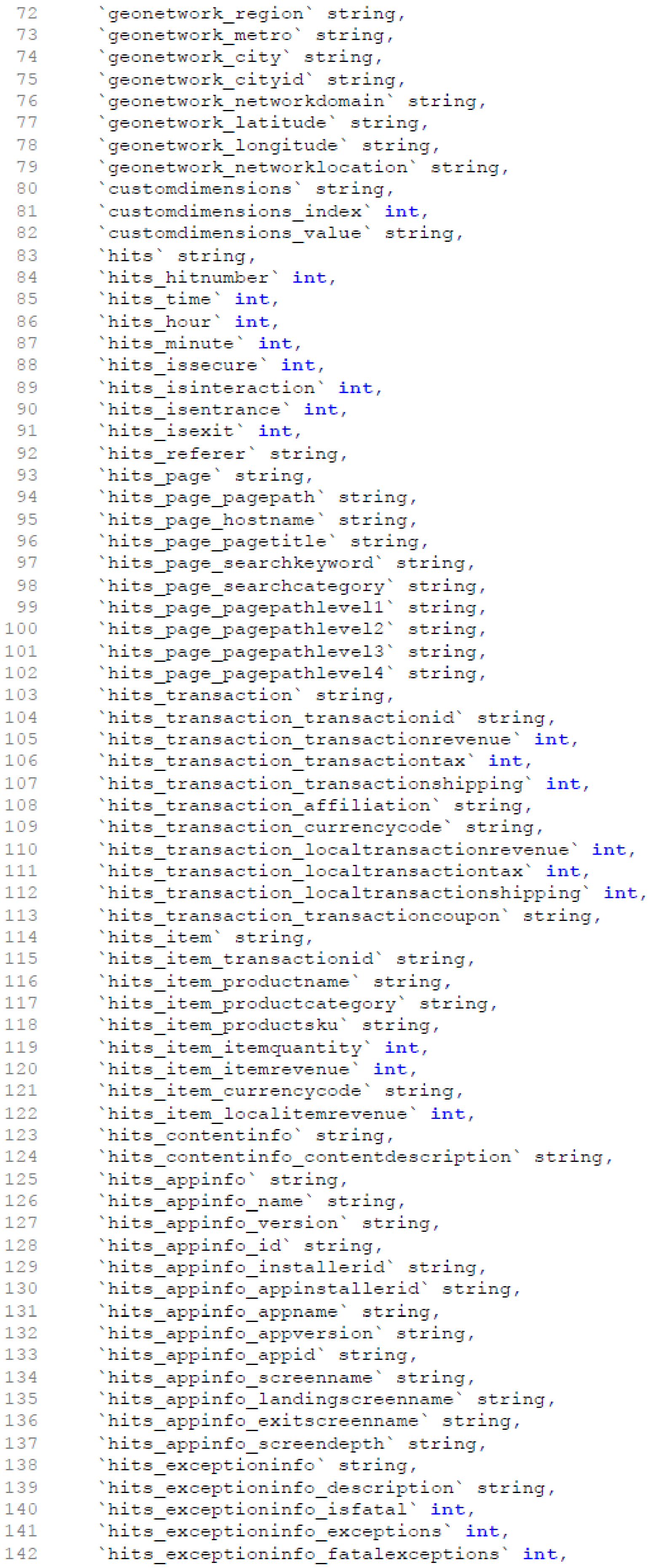
```
`hits_latencyTracking_domInteractiveTime`,
get_json_object(json_string, '$.hits.latencyTracking.domContentLoadedTime') AS `hits_latencyTracking_domContentLoadedTime`,
get_json_object(json_string, '$.hits.latencyTracking.userTimingValue') AS `hits_latencyTracking_userTimingValue`,
get_json_object(json_string, '$.hits.latencyTracking.userTimingSample') AS `hits_latencyTracking_userTimingSample`,
get_json_object(json_string, '$.hits.latencyTracking.userTimingVariable') AS `hits_latencyTracking_userTimingVariable`,
get_json_object(json_string, '$.hits.latencyTracking.userTimingCategory') AS `hits_latencyTracking_userTimingCategory`,
get_json_object(json_string, '$.hits.latencyTracking.userTimingLabel') AS `hits_latencyTracking_userTimingLabel`,
get_json_object(json_string, '$.hits.sourcePropertyInfo') AS `hits_sourcePropertyInfo`,
get_json_object(json_string, '$.hits.sourcePropertyInfo.sourcePropertyDisplayName') AS `hits_sourcePropertyInfo_sourcePropertyDisplayName`,
get_json_object(json_string, '$.hits.sourcePropertyInfo.sourcePropertyTrackingId') AS `hits_sourcePropertyInfo_sourcePropertyTrackingId`,
get_json_object(json_string, '$.hits.contentGroup') AS `hits_contentGroup`,
get_json_object(json_string, '$.hits.contentGroup.contentGroup1') AS `hits_contentGroup_contentGroup1`,
get_json_object(json_string, '$.hits.contentGroup.contentGroup2') AS `hits_contentGroup_contentGroup2`,
get_json_object(json_string, '$.hits.contentGroup.contentGroup3') AS `hits_contentGroup_contentGroup3`,
get_json_object(json_string, '$.hits.contentGroup.contentGroup4') AS `hits_contentGroup_contentGroup4`,
get_json_object(json_string, '$.hits.contentGroup.contentGroup5') AS `hits_contentGroup_contentGroup5`,
get_json_object(json_string, '$.hits.contentGroup.previousContentGroup1') AS `hits_contentGroup_previousContentGroup1`,
get_json_object(json_string, '$.hits.contentGroup.previousContentGroup2') AS `hits_contentGroup_previousContentGroup2`,
get_json_object(json_string, '$.hits.contentGroup.previousContentGroup3') AS `hits_contentGroup_previousContentGroup3`,
get_json_object(json_string, '$.hits.contentGroup.previousContentGroup4') AS `hits_contentGroup_previousContentGroup4`,
get_json_object(json_string, '$.hits.contentGroup.previousContentGroup5') AS `hits_contentGroup_previousContentGroup5`,
get_json_object(json_string, '$.hits.contentGroup.contentGroupUniqueViews1') AS `hits_contentGroup_contentGroupUniqueViews1`,
get_json_object(json_string, '$.hits.contentGroup.contentGroupUniqueViews2') AS `hits_contentGroup_contentGroupUniqueViews2`,
get_json_object(json_string, '$.hits.contentGroup.contentGroupUniqueViews3') AS `hits_contentGroup_contentGroupUniqueViews3`,
get_json_object(json_string, '$.hits.contentGroup.contentGroupUniqueViews4') AS
```

```
`hits_contentGroup_contentGroupUniqueViews4`,
get_json_object(json_string,
'$.hits.contentGroup.contentGroupUniqueViews5')                    AS
`hits_contentGroup_contentGroupUniqueViews5`,
get_json_object(json_string,
'$.hits.datASource')                                               AS `hits_datASource`,
get_json_object(json_string,
'$.fullVisitorId')                                                 AS `fullVisitorId`,
get_json_object(json_string,
'$.userId')                                                        AS `userId`,
get_json_object(json_string,
'$.channelGrouping')                                               AS `channelGrouping`,
get_json_object(json_string,
'$.socialEngagementType')                                          AS `socialEngagementType`,
file_date,
file name
FROM   dip_portal_stage.stg_googleanalytics;
```

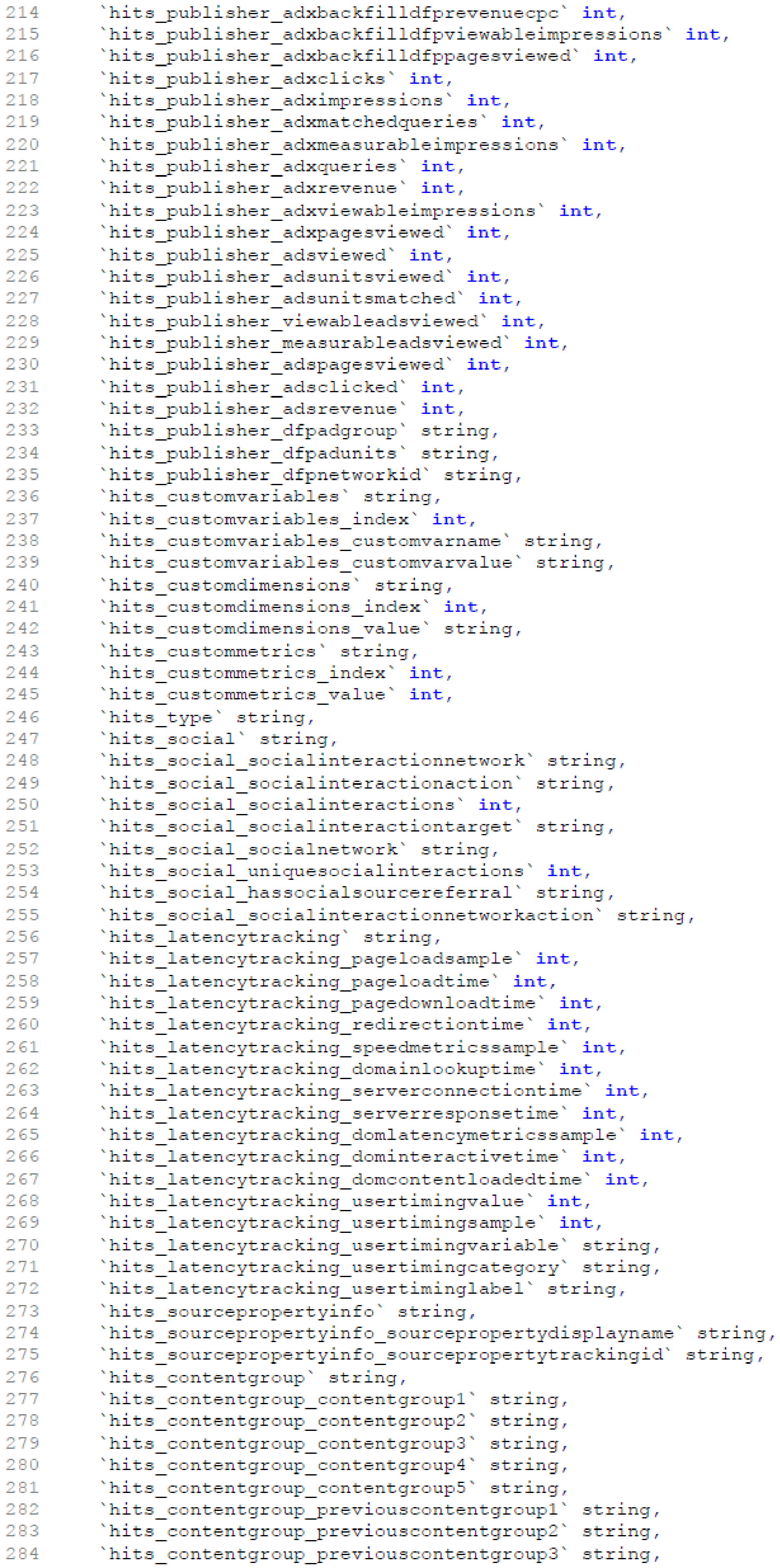
Appendix 8: WF[01]_core_create



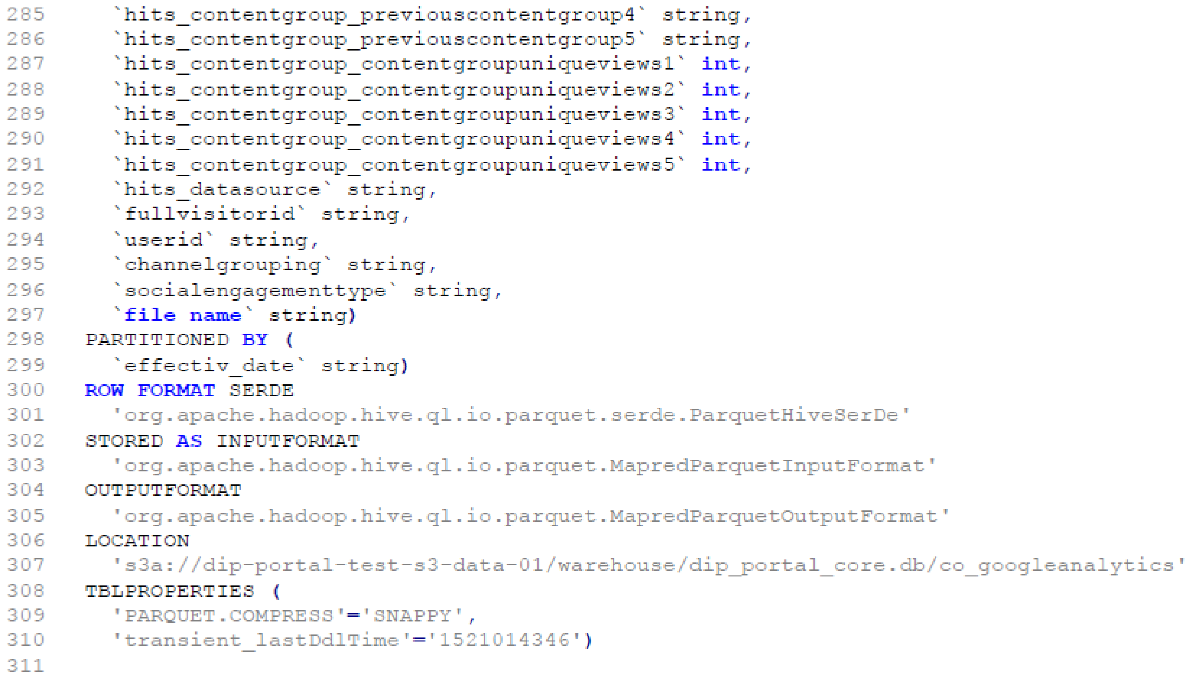
```
----------------------------------------------------------------
-- Workflow [01]: GOOGLE_ANALYTICS_STAGE_2_CORE
-- FILE: create_co_googleanalytics.sql
-- AREA: core
----------------------------------------------------------------

CREATE TABLE `co_googleanalytics`(
  `visitorid` int,
  `visitnumber` int,
  `visitid` int,
  `visitstarttime` int,
  `date` string,
  `totals` string,
  `totals_visits` int,
  `totals_hits` int,
  `totals_pageviews` int,
  `totals_timeonsite` int,
  `totals_bounces` int,
  `totals_transactions` int,
  `totals_transactionrevenue` int,
  `totals_newvisits` int,
  `totals_screenviews` int,
  `totals_uniquescreenviews` int,
  `totals_timeonscreen` int,
  `totals_totaltransactionrevenue` int,
  `totals_sessionqualitydim` int,
  `trafficsource` string,
  `trafficsource_referralpath` string,
  `trafficsource_campaign` string,
  `trafficsource_source` string,
  `trafficsource_medium` string,
  `trafficsource_keyword` string,
  `trafficsource_adcontent` string,
  `trafficsource_adwordsclickinfo` string,
  `trafficsource_adwordsclickinfo_campaignid` int,
  `trafficsource_adwordsclickinfo_adgroupid` int,
  `trafficsource_adwordsclickinfo_creativeid` int,
  `trafficsource_adwordsclickinfo_criteriaid` int,
  `trafficsource_adwordsclickinfo_page` int,
  `trafficsource_adwordsclickinfo_slot` string,
  `trafficsource_adwordsclickinfo_criteriaparameters` string,
  `trafficsource_adwordsclickinfo_gclid` string,
  `trafficsource_adwordsclickinfo_customerid` int,
  `trafficsource_adwordsclickinfo_adnetworktype` string,
  `trafficsource_adwordsclickinfo_targetingcriteria` string,
  `trafficsource_adwordsclickinfo_targetingcriteria_boomuserlistid` int,
  `trafficsource_adwordsclickinfo_isvideoad` int,
  `trafficsource_istruedirect` int,
  `trafficsource_campaigncode` string,
  `device` string,
  `device_browser` string,
  `device_browserversion` string,
  `device_browsersize` string,
  `device_operatingsystem` string,
  `device_operatingsystemversion` string,
  `device_ismobile` int,
  `device_mobiledevicebranding` string,
  `device_mobiledevicemodel` string,
  `device_mobileinputselector` string,
  `device_mobiledeviceinfo` string,
  `device_mobiledevicemarketingname` string,
  `device_flashversion` string,
  `device_javaenabled` int,
  `device_language` string,
  `device_screencolors` string,
  `device_screenresolution` string,
  `device_devicecategory` string,
  `geonetwork` string,
  `geonetwork_continent` string,
  `geonetwork_subcontinent` string,
  `geonetwork_country` string,
```



```
  `geonetwork_region` string,
  `geonetwork_metro` string,
  `geonetwork_city` string,
  `geonetwork_cityid` string,
  `geonetwork_networkdomain` string,
  `geonetwork_latitude` string,
  `geonetwork_longitude` string,
  `geonetwork_networklocation` string,
  `customdimensions` string,
  `customdimensions_index` int,
  `customdimensions_value` string,
  `hits` string,
  `hits_hitnumber` int,
  `hits_time` int,
  `hits_hour` int,
  `hits_minute` int,
  `hits_issecure` int,
  `hits_isinteraction` int,
  `hits_isentrance` int,
  `hits_isexit` int,
  `hits_referer` string,
  `hits_page` string,
  `hits_page_pagepath` string,
  `hits_page_hostname` string,
  `hits_page_pagetitle` string,
  `hits_page_searchkeyword` string,
  `hits_page_searchcategory` string,
  `hits_page_pagepathlevel1` string,
  `hits_page_pagepathlevel2` string,
  `hits_page_pagepathlevel3` string,
  `hits_page_pagepathlevel4` string,
  `hits_transaction` string,
  `hits_transaction_transactionid` string,
  `hits_transaction_transactionrevenue` int,
  `hits_transaction_transactiontax` int,
  `hits_transaction_transactionshipping` int,
  `hits_transaction_affiliation` string,
  `hits_transaction_currencycode` string,
  `hits_transaction_localtransactionrevenue` int,
  `hits_transaction_localtransactiontax` int,
  `hits_transaction_localtransactionshipping` int,
  `hits_transaction_transactioncoupon` string,
  `hits_item` string,
  `hits_item_transactionid` string,
  `hits_item_productname` string,
  `hits_item_productcategory` string,
  `hits_item_productsku` string,
  `hits_item_itemquantity` int,
  `hits_item_itemrevenue` int,
  `hits_item_currencycode` string,
  `hits_item_localitemrevenue` int,
  `hits_contentinfo` string,
  `hits_contentinfo_contentdescription` string,
  `hits_appinfo` string,
  `hits_appinfo_name` string,
  `hits_appinfo_version` string,
  `hits_appinfo_id` string,
  `hits_appinfo_installerid` string,
  `hits_appinfo_appinstallerid` string,
  `hits_appinfo_appname` string,
  `hits_appinfo_appversion` string,
  `hits_appinfo_appid` string,
  `hits_appinfo_screenname` string,
  `hits_appinfo_landingscreenname` string,
  `hits_appinfo_exitscreenname` string,
  `hits_appinfo_screendepth` string,
  `hits_exceptioninfo` string,
  `hits_exceptioninfo_description` string,
  `hits_exceptioninfo_isfatal` int,
  `hits_exceptioninfo_exceptions` int,
  `hits_exceptioninfo_fatalexceptions` int,
`hits_eventinfo` string,
`hits_eventinfo_eventcategory` string,
`hits_eventinfo_eventaction` string,
`hits_eventinfo_eventlabel` string,
`hits_eventinfo_eventvalue` int,
`hits_product` string,
`hits_product_productsku` string,
`hits_product_v2productname` string,
`hits_product_v2productcategory` string,
`hits_product_productvariant` string,
`hits_product_productbrand` string,
`hits_product_productrevenue` int,
`hits_product_localproductrevenue` int,
`hits_product_productprice` int,
`hits_product_localproductprice` int,
`hits_product_productquantity` int,
`hits_product_productrefundamount` int,
`hits_product_localproductrefundamount` int,
`hits_product_isimpression` int,
`hits_product_isclick` int,
`hits_product_customdimensions` string,
`hits_product_customdimensions_index` int,
`hits_product_customdimensions_value` string,
`hits_product_custommetrics` string,
`hits_product_custommetrics_index` int,
`hits_product_custommetrics_value` int,
`hits_product_productlistname` string,
`hits_product_productlistposition` int,
`hits_promotion` string,
`hits_promotion_promoid` string,
`hits_promotion_promoname` string,
`hits_promotion_promocreative` string,
`hits_promotion_promoposition` string,
`hits_promotionactioninfo` string,
`hits_promotionactioninfo_promoisview` int,
`hits_promotionactioninfo_promoisclick` int,
`hits_refund` string,
`hits_refund_refundamount` int,
`hits_refund_localrefundamount` int,
`hits_ecommerceaction` string,
`hits_ecommerceaction_action_type` string,
`hits_ecommerceaction_step` int,
`hits_ecommerceaction_option` string,
`hits_experiment` string,
`hits_experiment_experimentid` string,
`hits_experiment_experimentvariant` string,
`hits_publisher` string,
`hits_publisher_dfpclicks` int,
`hits_publisher_dfpimpressions` int,
`hits_publisher_dfpmatchedqueries` int,
`hits_publisher_dfpmeasurableimpressions` int,
`hits_publisher_dfpqueries` int,
`hits_publisher_dfprevenuecpm` int,
`hits_publisher_dfprevenuecpc` int,
`hits_publisher_dfpviewableimpressions` int,
`hits_publisher_dfppagesviewed` int,
`hits_publisher_adsensebackfilldfpclicks` int,
`hits_publisher_adsensebackfilldfpimpressions` int,
`hits_publisher_adsensebackfilldfpmatchedqueries` int,
`hits_publisher_adsensebackfilldfpmeasurableimpressions` int,
`hits_publisher_adsensebackfilldfpqueries` int,
`hits_publisher_adsensebackfilldfprevenuecpm` int,
`hits_publisher_adsensebackfilldfprevenuecpc` int,
`hits_publisher_adsensebackfilldfpviewableimpressions` int,
`hits_publisher_adsensebackfilldfppagesviewed` int,
`hits_publisher_adxbackfilldfpclicks` int,
`hits_publisher_adxbackfilldfpimpressions` int,
`hits_publisher_adxbackfilldfpmatchedqueries` int,
`hits_publisher_adxbackfilldfpmeasurableimpressions` int,
`hits_publisher_adxbackfilldfpqueries` int,
`hits_publisher_adxbackfilldfprevenuecpm` int,
```

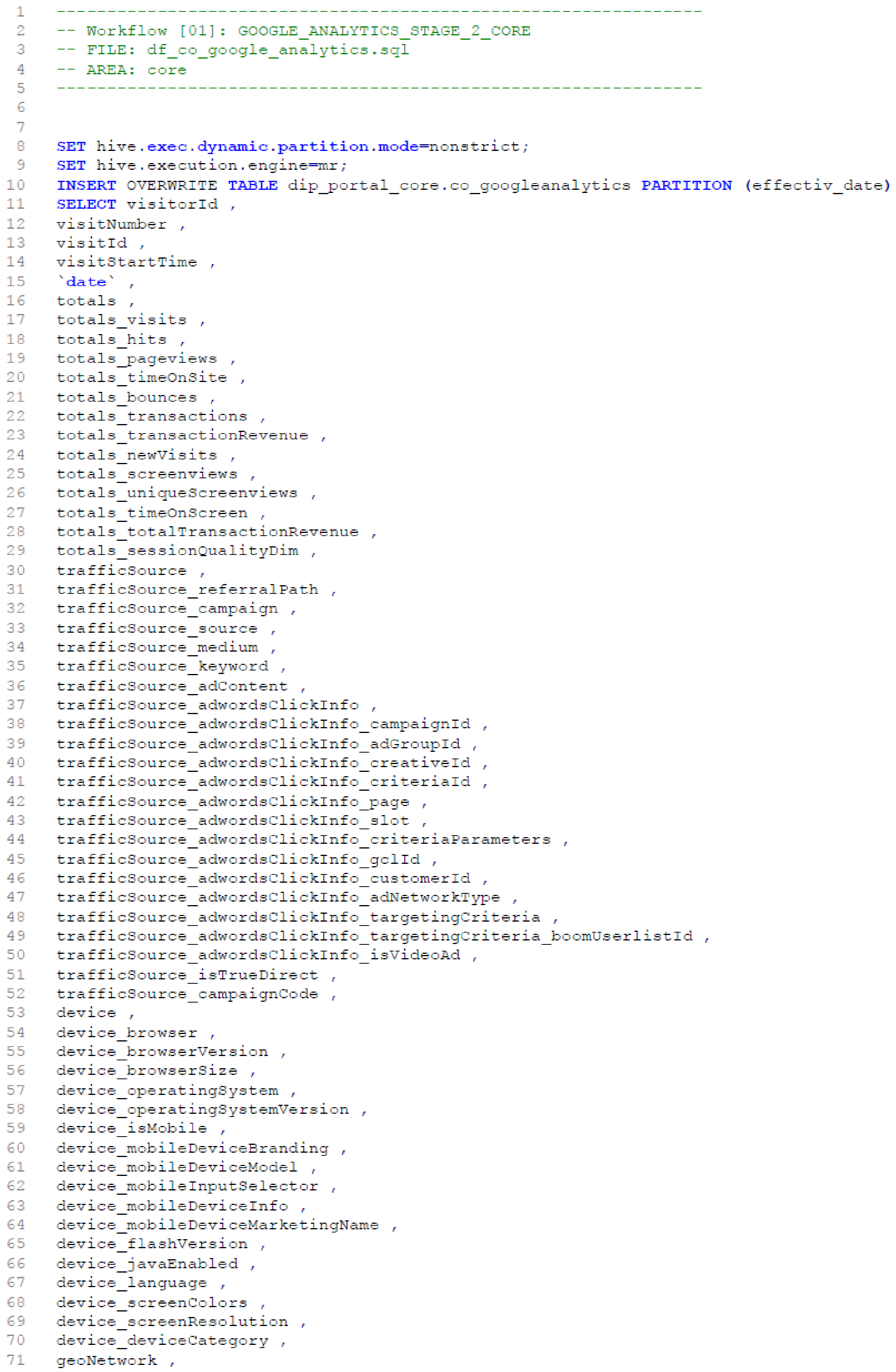


```
    `hits_publisher_adxbackfilldfprevenuecpc` int,
    `hits_publisher_adxbackfilldfpviewableimpressions` int,
    `hits_publisher_adxbackfilldfppagesviewed` int,
    `hits_publisher_adxclicks` int,
    `hits_publisher_adximpressions` int,
    `hits_publisher_adxmatchedqueries` int,
    `hits_publisher_adxmeasurableimpressions` int,
    `hits_publisher_adxqueries` int,
    `hits_publisher_adxrevenue` int,
    `hits_publisher_adxviewableimpressions` int,
    `hits_publisher_adxpagesviewed` int,
    `hits_publisher_adsviewed` int,
    `hits_publisher_adsunitsviewed` int,
    `hits_publisher_adsunitsmatched` int,
    `hits_publisher_viewableadsviewed` int,
    `hits_publisher_measurableadsviewed` int,
    `hits_publisher_adspagesviewed` int,
    `hits_publisher_adsclicked` int,
    `hits_publisher_adsrevenue` int,
    `hits_publisher_dfpadgroup` string,
    `hits_publisher_dfpadunits` string,
    `hits_publisher_dfpnetworkid` string,
    `hits_customvariables` string,
    `hits_customvariables_index` int,
    `hits_customvariables_customvarname` string,
    `hits_customvariables_customvarvalue` string,
    `hits_customdimensions` string,
    `hits_customdimensions_index` int,
    `hits_customdimensions_value` string,
    `hits_custommetrics` string,
    `hits_custommetrics_index` int,
    `hits_custommetrics_value` int,
    `hits_type` string,
    `hits_social` string,
    `hits_social_socialinteractionnetwork` string,
    `hits_social_socialinteractionaction` string,
    `hits_social_socialinteractions` int,
    `hits_social_socialinteractiontarget` string,
    `hits_social_socialnetwork` string,
    `hits_social_uniquesocialinteractions` int,
    `hits_social_hassocialsourcereferral` string,
    `hits_social_socialinteractionnetworkaction` string,
    `hits_latencytracking` string,
    `hits_latencytracking_pageloadsample` int,
    `hits_latencytracking_pageloadtime` int,
    `hits_latencytracking_pagedownloadtime` int,
    `hits_latencytracking_redirectiontime` int,
    `hits_latencytracking_speedmetricssample` int,
    `hits_latencytracking_domainlookuptime` int,
    `hits_latencytracking_serverconnectiontime` int,
    `hits_latencytracking_serverresponsetime` int,
    `hits_latencytracking_domlatencymetricssample` int,
    `hits_latencytracking_dominteractivetime` int,
    `hits_latencytracking_domcontentloadedtime` int,
    `hits_latencytracking_usertimingvalue` int,
    `hits_latencytracking_usertimingsample` int,
    `hits_latencytracking_usertimingvariable` string,
    `hits_latencytracking_usertimingcategory` string,
    `hits_latencytracking_usertiminglabel` string,
    `hits_sourcepropertyinfo` string,
    `hits_sourcepropertyinfo_sourcepropertydisplayname` string,
    `hits_sourcepropertyinfo_sourcepropertytrackingid` string,
    `hits_contentgroup` string,
    `hits_contentgroup_contentgroup1` string,
    `hits_contentgroup_contentgroup2` string,
    `hits_contentgroup_contentgroup3` string,
    `hits_contentgroup_contentgroup4` string,
    `hits_contentgroup_contentgroup5` string,
    `hits_contentgroup_previouscontentgroup1` string,
    `hits_contentgroup_previouscontentgroup2` string,
    `hits_contentgroup_previouscontentgroup3` string,
```

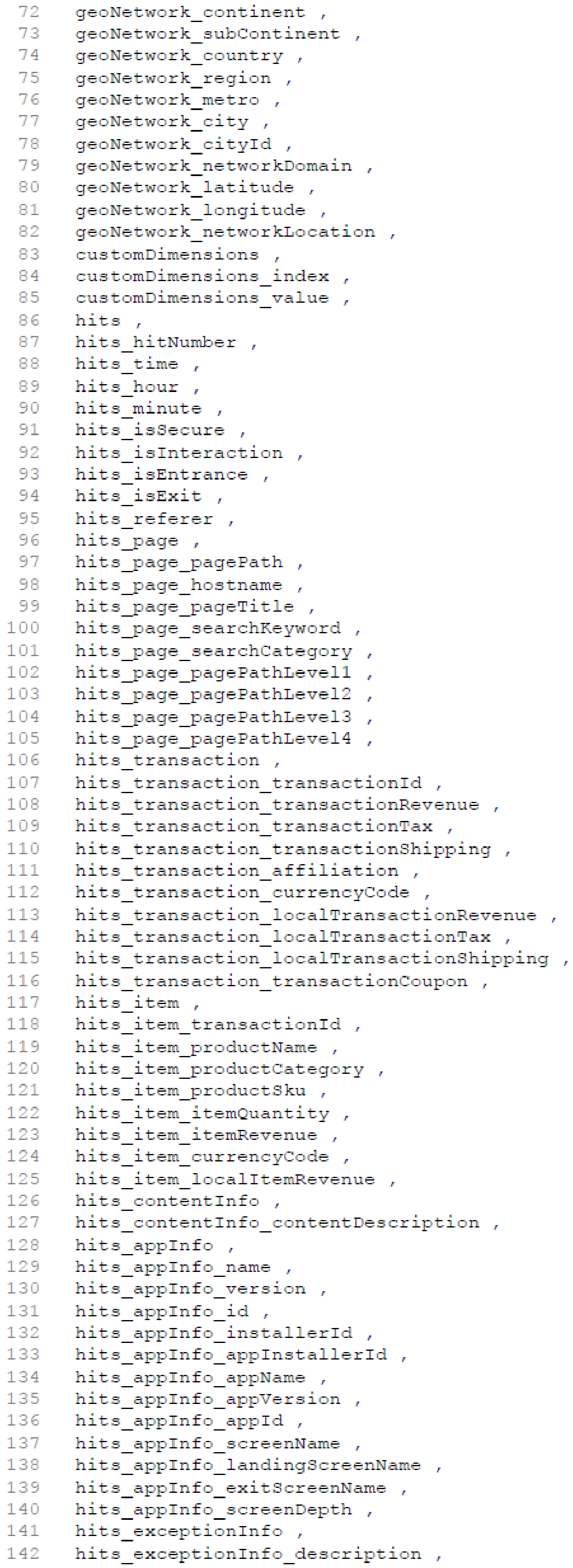


```
  `hits_contentgroup_previouscontentgroup4` string,
  `hits_contentgroup_previouscontentgroup5` string,
  `hits_contentgroup_contentgroupuniqueviews1` int,
  `hits_contentgroup_contentgroupuniqueviews2` int,
  `hits_contentgroup_contentgroupuniqueviews3` int,
  `hits_contentgroup_contentgroupuniqueviews4` int,
  `hits_contentgroup_contentgroupuniqueviews5` int,
  `hits_datasource` string,
  `fullvisitorid` string,
  `userid` string,
  `channelgrouping` string,
  `socialengagementtype` string,
  `file name` string)
PARTITIONED BY (
  `effectiv_date` string)
ROW FORMAT SERDE
  'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
STORED AS INPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'
LOCATION
  's3a://dip-portal-test-s3-data-01/warehouse/dip_portal_core.db/co_googleanalytics'
TBLPROPERTIES (
  'PARQUET.COMPRESS'='SNAPPY',
  'transient_lastDdlTime'='1521014346')
```

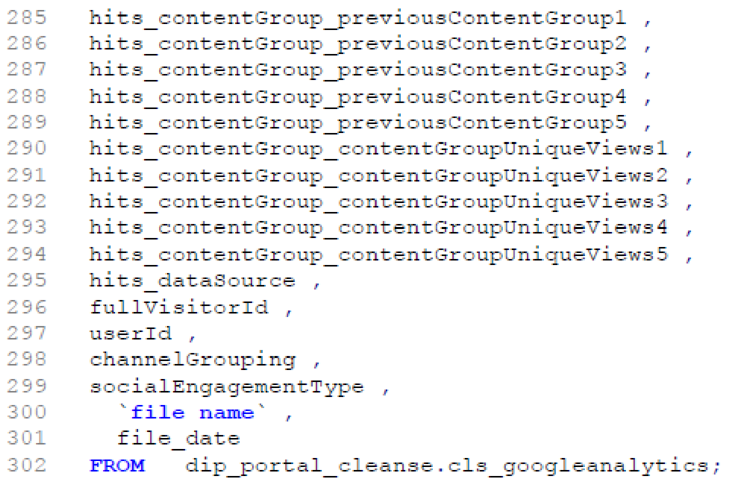
Appendix 9: WF[01]_core_run



```
-----------------------------------------------------------------
-- Workflow [01]: GOOGLE_ANALYTICS_STAGE_2_CORE
-- FILE: df_co_google_analytics.sql
-- AREA: core
-----------------------------------------------------------------


SET hive.exec.dynamic.partition.mode=nonstrict;
SET hive.execution.engine=mr;
INSERT OVERWRITE TABLE dip_portal_core.co_googleanalytics PARTITION (effectiv_date)
SELECT visitorId ,
visitNumber ,
visitId ,
visitStartTime ,
`date` ,
totals ,
totals_visits ,
totals_hits ,
totals_pageviews ,
totals_timeOnSite ,
totals_bounces ,
totals_transactions ,
totals_transactionRevenue ,
totals_newVisits ,
totals_screenviews ,
totals_uniqueScreenviews ,
totals_timeOnScreen ,
totals_totalTransactionRevenue ,
totals_sessionQualityDim ,
trafficSource ,
trafficSource_referralPath ,
trafficSource_campaign ,
trafficSource_source ,
trafficSource_medium ,
trafficSource_keyword ,
trafficSource_adContent ,
trafficSource_adwordsClickInfo ,
trafficSource_adwordsClickInfo_campaignId ,
trafficSource_adwordsClickInfo_adGroupId ,
trafficSource_adwordsClickInfo_creativeId ,
trafficSource_adwordsClickInfo_criteriaId ,
trafficSource_adwordsClickInfo_page ,
trafficSource_adwordsClickInfo_slot ,
trafficSource_adwordsClickInfo_criteriaParameters ,
trafficSource_adwordsClickInfo_gclId ,
trafficSource_adwordsClickInfo_customerId ,
trafficSource_adwordsClickInfo_adNetworkType ,
trafficSource_adwordsClickInfo_targetingCriteria ,
trafficSource_adwordsClickInfo_targetingCriteria_boomUserlistId ,
trafficSource_adwordsClickInfo_isVideoAd ,
trafficSource_isTrueDirect ,
trafficSource_campaignCode ,
device ,
device_browser ,
device_browserVersion ,
device_browserSize ,
device_operatingSystem ,
device_operatingSystemVersion ,
device_isMobile ,
device_mobileDeviceBranding ,
device_mobileDeviceModel ,
device_mobileInputSelector ,
device_mobileDeviceInfo ,
device_mobileDeviceMarketingName ,
device_flashVersion ,
device_javaEnabled ,
device_language ,
device_screenColors ,
device_screenResolution ,
device_deviceCategory ,
geoNetwork ,
```



```
geoNetwork_continent ,
geoNetwork_subContinent ,
geoNetwork_country ,
geoNetwork_region ,
geoNetwork_metro ,
geoNetwork_city ,
geoNetwork_cityId ,
geoNetwork_networkDomain ,
geoNetwork_latitude ,
geoNetwork_longitude ,
geoNetwork_networkLocation ,
customDimensions ,
customDimensions_index ,
customDimensions_value ,
hits ,
hits_hitNumber ,
hits_time ,
hits_hour ,
hits_minute ,
hits_isSecure ,
hits_isInteraction ,
hits_isEntrance ,
hits_isExit ,
hits_referer ,
hits_page ,
hits_page_pagePath ,
hits_page_hostname ,
hits_page_pageTitle ,
hits_page_searchKeyword ,
hits_page_searchCategory ,
hits_page_pagePathLevel1 ,
hits_page_pagePathLevel2 ,
hits_page_pagePathLevel3 ,
hits_page_pagePathLevel4 ,
hits_transaction ,
hits_transaction_transactionId ,
hits_transaction_transactionRevenue ,
hits_transaction_transactionTax ,
hits_transaction_transactionShipping ,
hits_transaction_affiliation ,
hits_transaction_currencyCode ,
hits_transaction_localTransactionRevenue ,
hits_transaction_localTransactionTax ,
hits_transaction_localTransactionShipping ,
hits_transaction_transactionCoupon ,
hits_item ,
hits_item_transactionId ,
hits_item_productName ,
hits_item_productCategory ,
hits_item_productSku ,
hits_item_itemQuantity ,
hits_item_itemRevenue ,
hits_item_currencyCode ,
hits_item_localItemRevenue ,
hits_contentInfo ,
hits_contentInfo_contentDescription ,
hits_appInfo ,
hits_appInfo_name ,
hits_appInfo_version ,
hits_appInfo_id ,
hits_appInfo_installerId ,
hits_appInfo_appInstallerId ,
hits_appInfo_appName ,
hits_appInfo_appVersion ,
hits_appInfo_appId ,
hits_appInfo_screenName ,
hits_appInfo_landingScreenName ,
hits_appInfo_exitScreenName ,
hits_appInfo_screenDepth ,
hits_exceptionInfo ,
hits_exceptionInfo_description ,
hits_exceptionInfo_isFatal ,
hits_exceptionInfo_exceptions ,
hits_exceptionInfo_fatalExceptions ,
hits_eventInfo ,
hits_eventInfo_eventCategory ,
hits_eventInfo_eventAction ,
hits_eventInfo_eventLabel ,
hits_eventInfo_eventValue ,
hits_product ,
hits_product_productSKU ,
hits_product_v2ProductName ,
hits_product_v2ProductCategory ,
hits_product_productVariant ,
hits_product_productBrand ,
hits_product_productRevenue ,
hits_product_localProductRevenue ,
hits_product_productPrice ,
hits_product_localProductPrice ,
hits_product_productQuantity ,
hits_product_productRefundAmount ,
hits_product_localProductRefundAmount ,
hits_product_isImpression ,
hits_product_isClick ,
hits_product_customDimensions ,
hits_product_customDimensions_index ,
hits_product_customDimensions_value ,
hits_product_customMetrics ,
hits_product_customMetrics_index ,
hits_product_customMetrics_value ,
hits_product_productListName ,
hits_product_productListPosition ,
hits_promotion ,
hits_promotion_promoId ,
hits_promotion_promoName ,
hits_promotion_promoCreative ,
hits_promotion_promoPosition ,
hits_promotionActionInfo ,
hits_promotionActionInfo_promoIsView ,
hits_promotionActionInfo_promoIsClick ,
hits_refund ,
hits_refund_refundAmount ,
hits_refund_localRefundAmount ,
hits_eCommerceAction ,
hits_eCommerceAction_action_type ,
hits_eCommerceAction_step ,
hits_eCommerceAction_option ,
hits_experiment ,
hits_experiment_experimentId ,
hits_experiment_experimentVariant ,
hits_publisher ,
hits_publisher_dfpClicks ,
hits_publisher_dfpImpressions ,
hits_publisher_dfpMatchedQueries ,
hits_publisher_dfpMeasurableImpressions ,
hits_publisher_dfpQueries ,
hits_publisher_dfpRevenueCpm ,
hits_publisher_dfpRevenueCpc ,
hits_publisher_dfpViewableImpressions ,
hits_publisher_dfpPagesViewed ,
hits_publisher_adsenseBackfillDfpClicks ,
hits_publisher_adsenseBackfillDfpImpressions ,
hits_publisher_adsenseBackfillDfpMatchedQueries ,
hits_publisher_adsenseBackfillDfpMeasurableImpressions ,
hits_publisher_adsenseBackfillDfpQueries ,
hits_publisher_adsenseBackfillDfpRevenueCpm ,
hits_publisher_adsenseBackfillDfpRevenueCpc ,
hits_publisher_adsenseBackfillDfpViewableImpressions ,
hits_publisher_adsenseBackfillDfpPagesViewed ,
hits_publisher_adxBackfillDfpClicks ,
hits_publisher_adxBackfillDfpImpressions ,
hits_publisher_adxBackfillDfpMatchedQueries ,
hits_publisher_adxBackfillDfpMeasurableImpressions ,
hits_publisher_adxBackfillDfpQueries ,
hits_publisher_adxBackfillDfpRevenueCpm ,
hits_publisher_adxBackfillDfpRevenueCpc ,
hits_publisher_adxBackfillDfpViewableImpressions ,
hits_publisher_adxBackfillDfpPagesViewed ,
hits_publisher_adxClicks ,
hits_publisher_adxImpressions ,
hits_publisher_adxMatchedQueries ,
hits_publisher_adxMeasurableImpressions ,
hits_publisher_adxQueries ,
hits_publisher_adxRevenue ,
hits_publisher_adxViewableImpressions ,
hits_publisher_adxPagesViewed ,
hits_publisher_adsViewed ,
hits_publisher_adsUnitsViewed ,
hits_publisher_adsUnitsMatched ,
hits_publisher_viewableAdsViewed ,
hits_publisher_measurableAdsViewed ,
hits_publisher_adsPagesViewed ,
hits_publisher_adsClicked ,
hits_publisher_adsRevenue ,
hits_publisher_dfpAdGroup ,
hits_publisher_dfpAdUnits ,
hits_publisher_dfpNetworkId ,
hits_customVariables ,
hits_customVariables_index ,
hits_customVariables_customVarName ,
hits_customVariables_customVarValue ,
hits_customDimensions ,
hits_customDimensions_index ,
hits_customDimensions_value ,
hits_customMetrics ,
hits_customMetrics_index ,
hits_customMetrics_value ,
hits_type ,
hits_social ,
hits_social_socialInteractionNetwork ,
hits_social_socialInteractionAction ,
hits_social_socialInteractions ,
hits_social_socialInteractionTarget ,
hits_social_socialNetwork ,
hits_social_uniqueSocialInteractions ,
hits_social_hasSocialSourceReferral ,
hits_social_socialInteractionNetworkAction ,
hits_latencyTracking ,
hits_latencyTracking_pageLoadSample ,
hits_latencyTracking_pageLoadTime ,
hits_latencyTracking_pageDownloadTime ,
hits_latencyTracking_redirectionTime ,
hits_latencyTracking_speedMetricsSample ,
hits_latencyTracking_domainLookupTime ,
hits_latencyTracking_serverConnectionTime ,
hits_latencyTracking_serverResponseTime ,
hits_latencyTracking_domLatencyMetricsSample ,
hits_latencyTracking_domInteractiveTime ,
hits_latencyTracking_domContentLoadedTime ,
hits_latencyTracking_userTimingValue ,
hits_latencyTracking_userTimingSample ,
hits_latencyTracking_userTimingVariable ,
hits_latencyTracking_userTimingCategory ,
hits_latencyTracking_userTimingLabel ,
hits_sourcePropertyInfo ,
hits_sourcePropertyInfo_sourcePropertyDisplayName ,
hits_sourcePropertyInfo_sourcePropertyTrackingId ,
hits_contentGroup ,
hits_contentGroup_contentGroup1 ,
hits_contentGroup_contentGroup2 ,
hits_contentGroup_contentGroup3 ,
hits_contentGroup_contentGroup4 ,
hits_contentGroup_contentGroup5 ,
```



```
hits_contentGroup_previousContentGroup1 ,
hits_contentGroup_previousContentGroup2 ,
hits_contentGroup_previousContentGroup3 ,
hits_contentGroup_previousContentGroup4 ,
hits_contentGroup_previousContentGroup5 ,
hits_contentGroup_contentGroupUniqueViews1 ,
hits_contentGroup_contentGroupUniqueViews2 ,
hits_contentGroup_contentGroupUniqueViews3 ,
hits_contentGroup_contentGroupUniqueViews4 ,
hits_contentGroup_contentGroupUniqueViews5 ,
hits_dataSource ,
fullVisitorId ,
userId ,
channelGrouping ,
socialEngagementType ,
  `file name` ,
  file_date
FROM   dip_portal_cleanse.cls_googleanalytics;
```

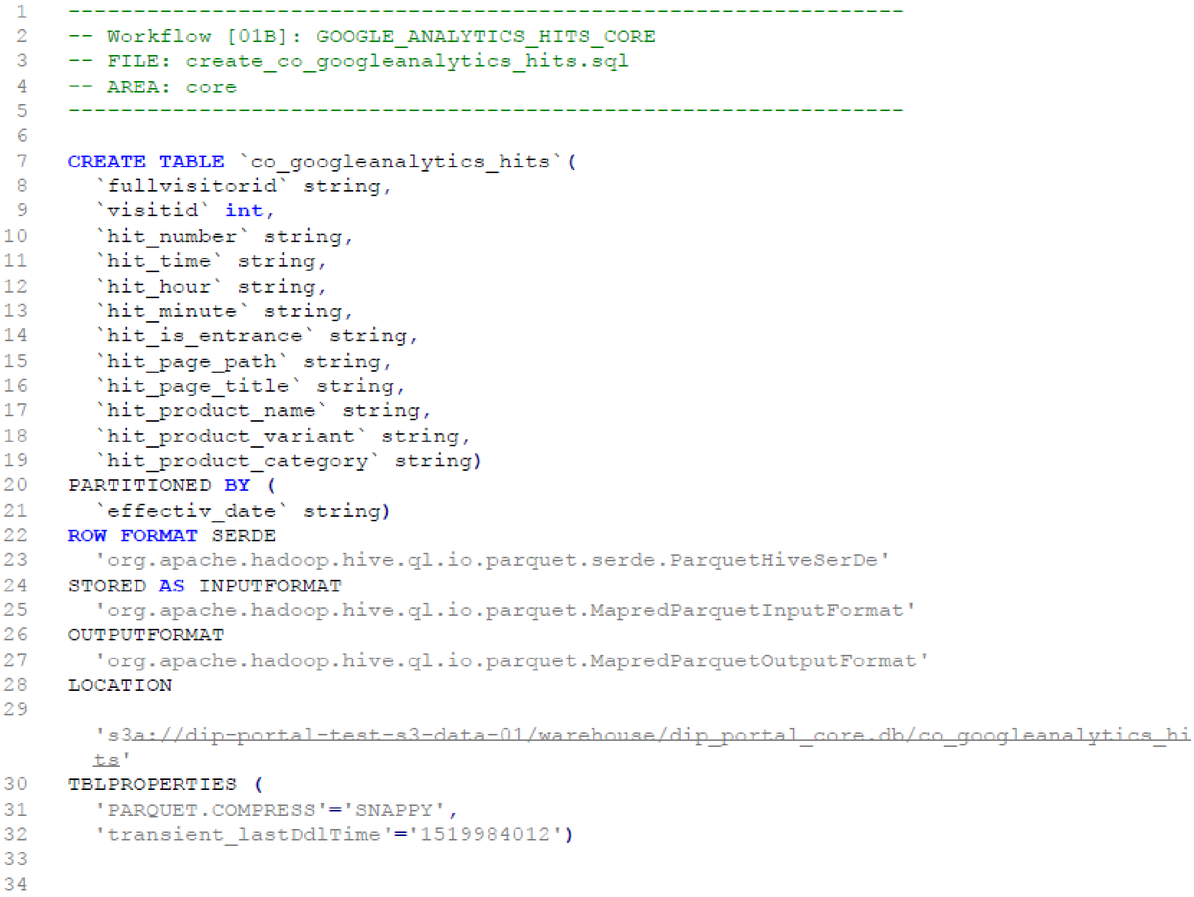
Appendix 10: WF[01A]_core_create

```sql
-----------------------------------------------------------------
-- Workflow [01A]: GOOGLE_ANALYTICS_HITS_CUSTOMDIM_CORE
-- FILE: create_co_googleanalytics_hits_customdim.sql
-- AREA: core
-----------------------------------------------------------------

CREATE TABLE `co_googleanalytics_hits_customdim`(
  `fullvisitorid` string,
  `hits_customdim_68` string,
  `hits_customdim_69` string)
PARTITIONED BY (
  `effectiv_date` string)
ROW FORMAT SERDE
  'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
STORED AS INPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'
LOCATION
  's3a://dip-portal-test-s3-data-01/warehouse/dip_portal_core.db/co_googleanalytics_hits_customdim'
TBLPROPERTIES (
  'PARQUET.COMPRESS'='SNAPPY',
  'numFiles'='1',
  'numRows'='100',
  'rawDataSize'='600',
  'totalSize'='18119',
  'transient_lastDdlTime'='1522232631')

```

Appendix 11: WF[01A]_core_run

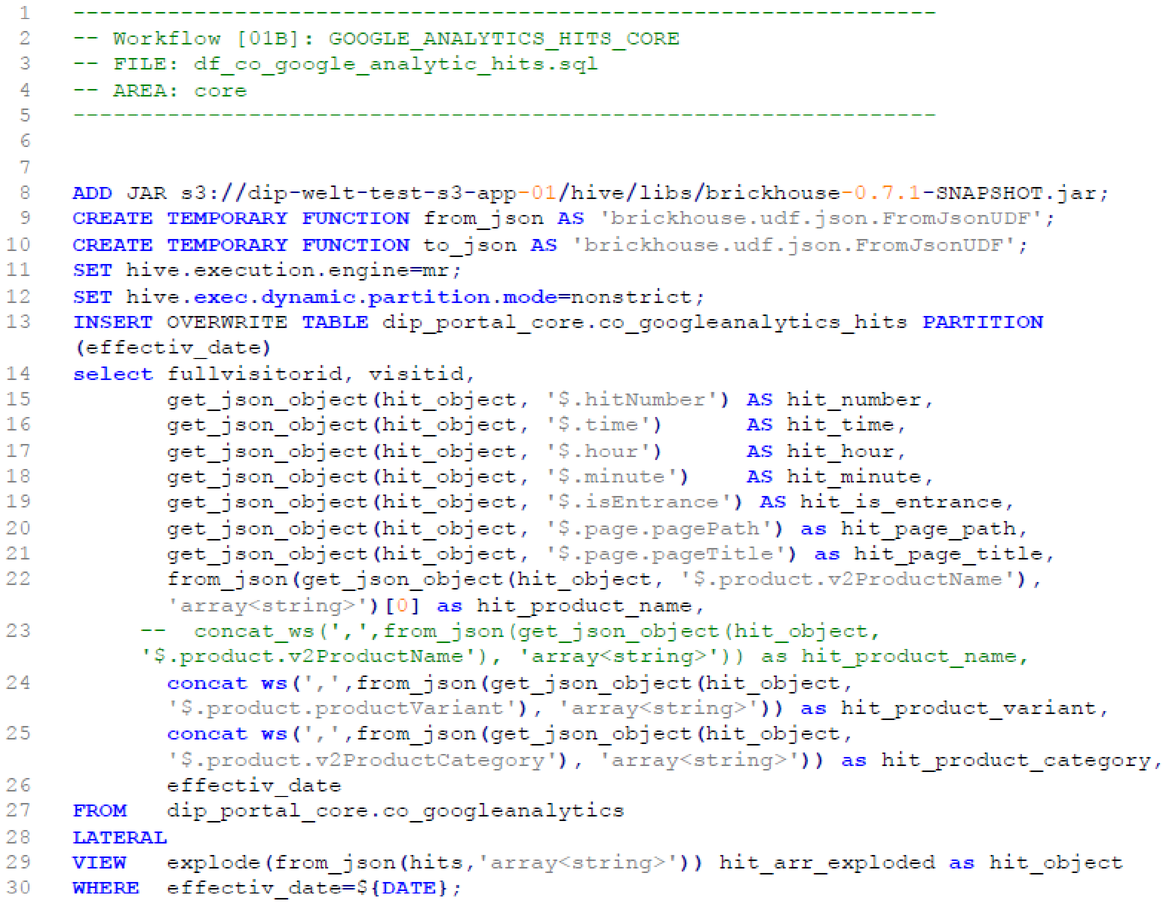
```
----------------------------------------------------------------
-- Workflow [01A]: GOOGLE_ANALYTICS_HITS_CUSTOMDIM_CORE
-- FILE: df_co_google_analytics_hits_customdim.sql
-- AREA: core
----------------------------------------------------------------

ADD JAR s3://dip-welt-test-s3-app-01/hive/libs/brickhouse-0.7.1-SNAPSHOT.jar;
CREATE TEMPORARY FUNCTION from_json AS 'brickhouse.udf.json.FromJsonUDF';
CREATE TEMPORARY FUNCTION to_json AS 'brickhouse.udf.json.FromJsonUDF';
SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;
WITH co_googleanalytics_hits_customdim_tmp
AS
(
SELECT fullvisitorid,
       sort_array(from_json(hits_customdimensions,'array<string>')) AS hits_customdimensions,
       collect_set(get_json_object(hits_customdimensions_object, '$.value'))[0] AS hits_customdim_68,
       collect_set(get_json_object(hits_customdimensions_object, '$.value'))[1] AS hits_customdim_69,
       rank() over (PARTITION BY collect_set(get_json_object(hits_customdimensions_object, '$.value'))[1] ORDER BY fullvisitorid DESC) AS rank_fullvisitorid,
       effectiv_date
FROM   dip_portal_core.co_googleanalytics
LATERAL
VIEW   explode(sort_array(from_json(hits_customdimensions,'array<string>'))) hits_customdimensions_arr_exploded AS hits_customdimensions_object
WHERE  effectiv_date=${DATE}
AND    get_json_object(hits_customdimensions_object, '$.index') IN (68,69)
GROUP
BY     fullvisitorid,
       hits_customdimensions,effectiv_date
),
co_googleanalytics_hits_customdim_multiple_fullvisitorid_tmp
AS
(
SELECT
       fullvisitorid,
       hits_customdim_68,
       hits_customdim_69,
       effectiv_date
FROM   co_googleanalytics_hits_customdim_tmp
WHERE  rank_fullvisitorid = 1 -- remove multiple fullvisitorid per device
AND    hits_customdim_69 IS NOT NULL
)

INSERT OVERWRITE TABLE dip_portal_core.co_googleanalytics_hits_customdim PARTITION (effectiv_date)
SELECT
       DISTINCT fullvisitorid, -- remove multiple fullvisitorids (these exist because of multiple sessions per day)
       hits_customdim_68,
       hits_customdim_69,
       CAST(effectiv_date AS string)
FROM   co_googleanalytics_hits_customdim_multiple_fullvisitorid_tmp
;

```

Appendix 12: WF[01B]_core_create



```sql
----------------------------------------------------------------
-- Workflow [01B]: GOOGLE_ANALYTICS_HITS_CORE
-- FILE: create_co_googleanalytics_hits.sql
-- AREA: core
----------------------------------------------------------------

CREATE TABLE `co_googleanalytics_hits`(
  `fullvisitorid` string,
  `visitid` int,
  `hit_number` string,
  `hit_time` string,
  `hit_hour` string,
  `hit_minute` string,
  `hit_is_entrance` string,
  `hit_page_path` string,
  `hit_page_title` string,
  `hit_product_name` string,
  `hit_product_variant` string,
  `hit_product_category` string)
PARTITIONED BY (
  `effectiv_date` string)
ROW FORMAT SERDE
  'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
STORED AS INPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'
LOCATION
  's3a://dip-portal-test-s3-data-01/warehouse/dip_portal_core.db/co_googleanalytics_hits'
TBLPROPERTIES (
  'PARQUET.COMPRESS'='SNAPPY',
  'transient_lastDdlTime'='1519984012')


```

Appendix 13: WF[01B]_core_run

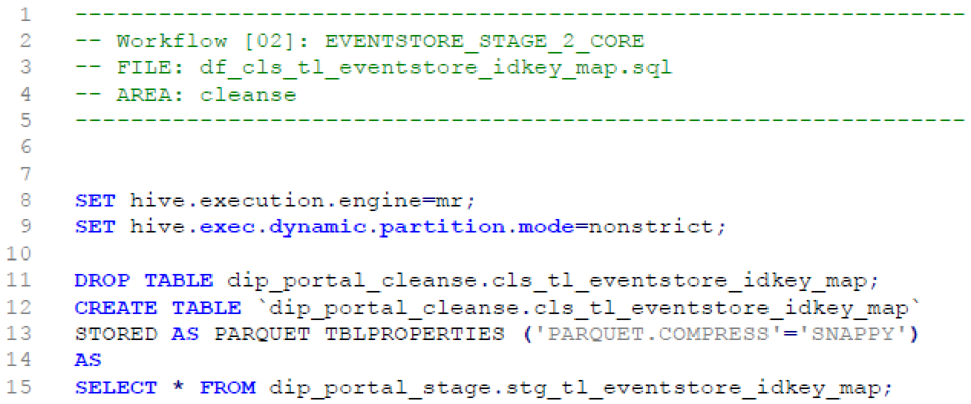


```
----------------------------------------------------------------
-- Workflow [01B]: GOOGLE_ANALYTICS_HITS_CORE
-- FILE: df_co_google_analytic_hits.sql
-- AREA: core
----------------------------------------------------------------


ADD JAR s3://dip-welt-test-s3-app-01/hive/libs/brickhouse-0.7.1-SNAPSHOT.jar;
CREATE TEMPORARY FUNCTION from_json AS 'brickhouse.udf.json.FromJsonUDF';
CREATE TEMPORARY FUNCTION to_json AS 'brickhouse.udf.json.FromJsonUDF';
SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;
INSERT OVERWRITE TABLE dip_portal_core.co_googleanalytics_hits PARTITION (effectiv_date)
select fullvisitorid, visitid,
       get_json_object(hit_object, '$.hitNumber') AS hit_number,
       get_json_object(hit_object, '$.time')      AS hit_time,
       get_json_object(hit_object, '$.hour')      AS hit_hour,
       get_json_object(hit_object, '$.minute')    AS hit_minute,
       get_json_object(hit_object, '$.isEntrance') AS hit_is_entrance,
       get_json_object(hit_object, '$.page.pagePath') as hit_page_path,
       get_json_object(hit_object, '$.page.pageTitle') as hit_page_title,
       from_json(get_json_object(hit_object, '$.product.v2ProductName'), 'array<string>')[0] as hit_product_name,
     --  concat_ws(',',from_json(get_json_object(hit_object, '$.product.v2ProductName'), 'array<string>')) as hit_product_name,
       concat ws(',',from_json(get_json_object(hit_object, '$.product.productVariant'), 'array<string>')) as hit_product_variant,
       concat ws(',',from_json(get_json_object(hit_object, '$.product.v2ProductCategory'), 'array<string>')) as hit_product_category,
       effectiv_date
FROM   dip_portal_core.co_googleanalytics
LATERAL
VIEW   explode(from_json(hits,'array<string>')) hit_arr_exploded as hit_object
WHERE  effectiv_date=${DATE};
```

Appendix 14: WF[02]_staging_run

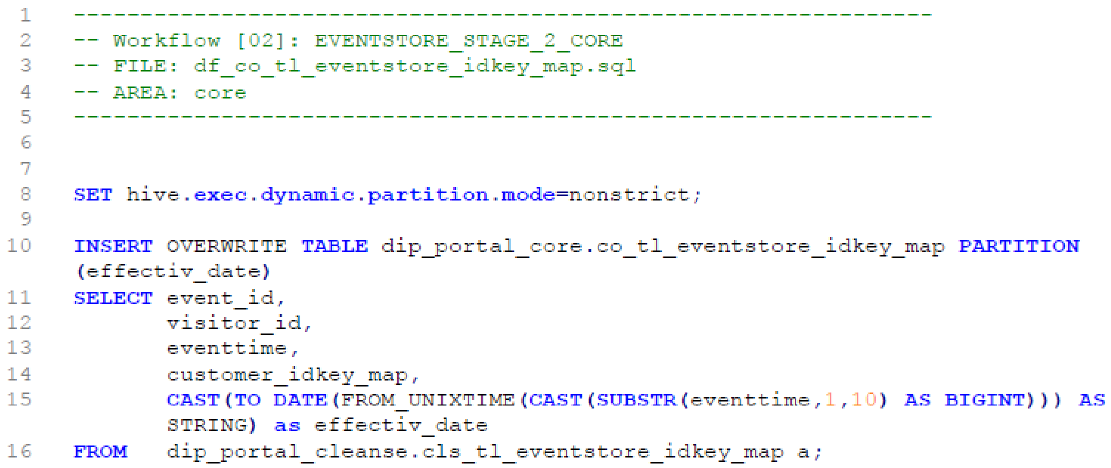
```
----------------------------------------------------------------
-- Workflow [02]: EVENTSTORE_STAGE_2_CORE
-- FILE: df_stg_tl_eventstore_idkey_map.sql
-- AREA: stage
----------------------------------------------------------------


SET hive.exec.dynamic.partition.mode=nonstrict;
MSCK REPAIR TABLE dip_portal_stage.stg_tl_eventstore_ext;
DROP TABLE dip_portal_stage.stg_tl_eventstore_idkey_map;
CREATE TABLE dip_portal_stage.stg_tl_eventstore_idkey_map
AS
SELECT get_json_object(json_string, '$.eventid')                as event_id,
       get_json_object(json_string, '$.visitorid')              as visitor_id,
       get_json_object(json_string, '$.eventtime')              as eventtime,
       get_json_object(json_string, '$.udo_customer_idkey_map')   as customer_idkey_map
FROM   dip_portal_stage.stg_tl_eventstore_ext a
WHERE  TO DATE(FROM_UNIXTIME(CAST(SUBSTR(get_json_object(json_string, '$.eventtime'),1,10) AS BIGINT))) = DATE SUB(${DATE}, 0)
AND    file_date BETWEEN CAST(DATE SUB(${DATE}, 1) AS STRING) AND CAST(DATE_SUB(${DATE}, 0) AS STRING);
```

Appendix 15: WF[02]_cleanse_run



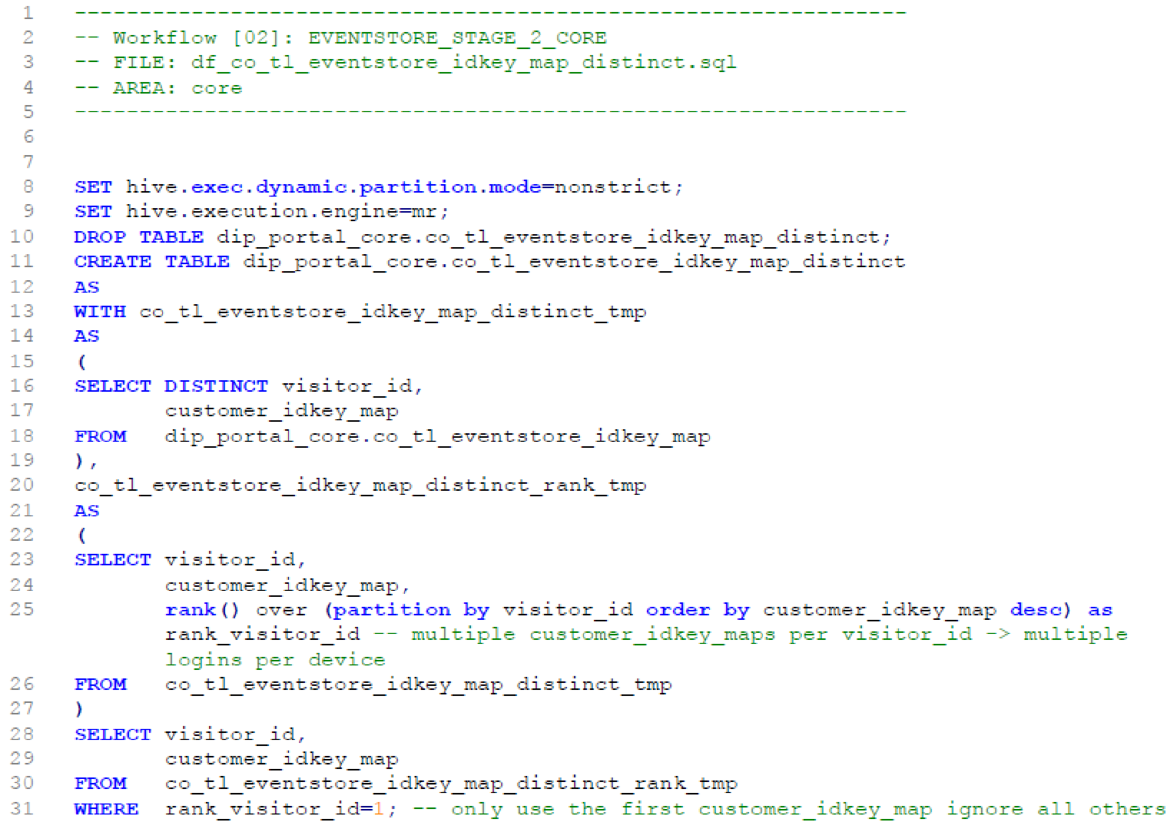
```
-----------------------------------------------------------------
-- Workflow [02]: EVENTSTORE_STAGE_2_CORE
-- FILE: df_cls_tl_eventstore_idkey_map.sql
-- AREA: cleanse
-----------------------------------------------------------------


SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;

DROP TABLE dip_portal_cleanse.cls_tl_eventstore_idkey_map;
CREATE TABLE `dip_portal_cleanse.cls_tl_eventstore_idkey_map`
STORED AS PARQUET TBLPROPERTIES ('PARQUET.COMPRESS'='SNAPPY')
AS
SELECT * FROM dip_portal_stage.stg_tl_eventstore_idkey_map;
```

Appendix 16: WF[02]_core_run01



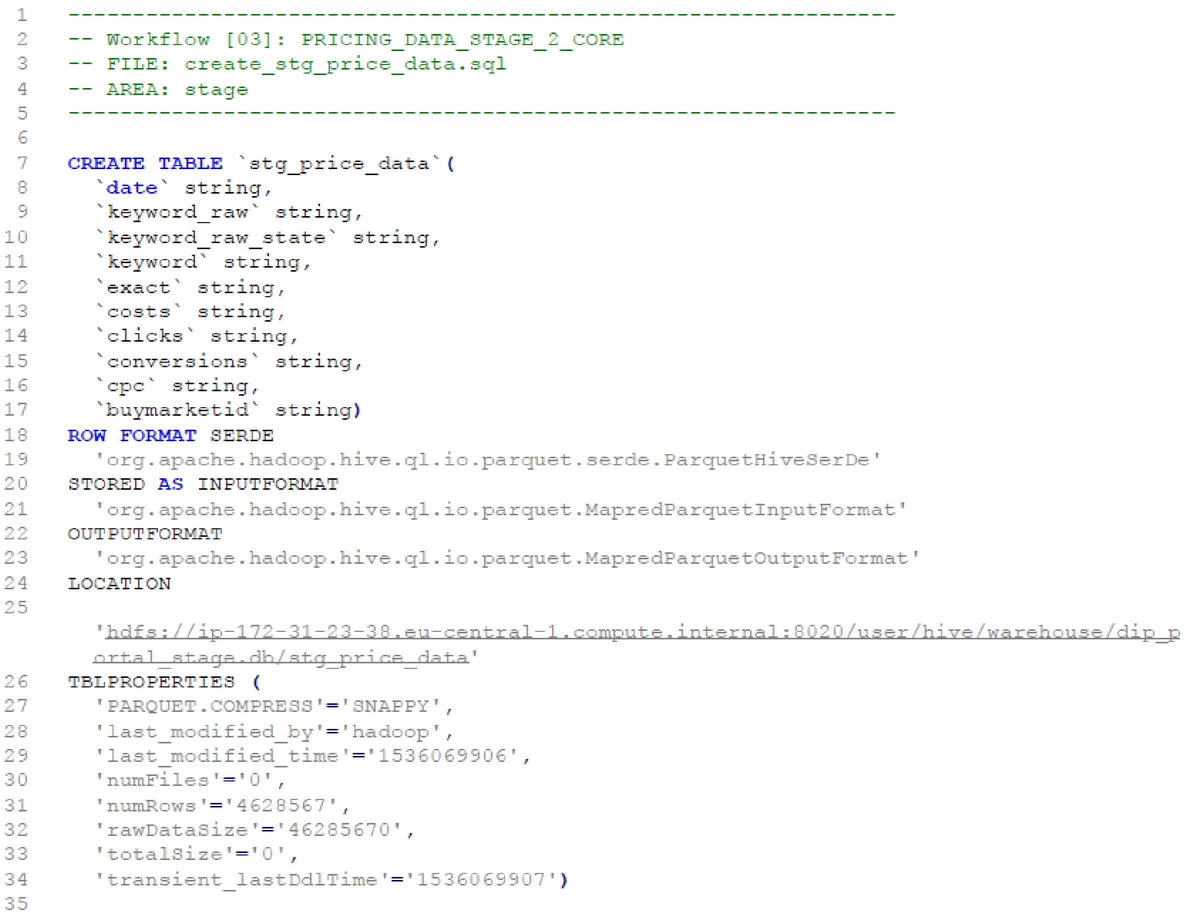
```
----------------------------------------------------------------
-- Workflow [02]: EVENTSTORE_STAGE_2_CORE
-- FILE: df_co_tl_eventstore_idkey_map.sql
-- AREA: core
----------------------------------------------------------------


SET hive.exec.dynamic.partition.mode=nonstrict;

INSERT OVERWRITE TABLE dip_portal_core.co_tl_eventstore_idkey_map PARTITION (effectiv_date)
SELECT event_id,
       visitor_id,
       eventtime,
       customer_idkey_map,
       CAST(TO DATE(FROM_UNIXTIME(CAST(SUBSTR(eventtime,1,10) AS BIGINT))) AS STRING) as effectiv_date
FROM   dip_portal_cleanse.cls_tl_eventstore_idkey_map a;
```

Appendix 17: WF[02]_core_run02



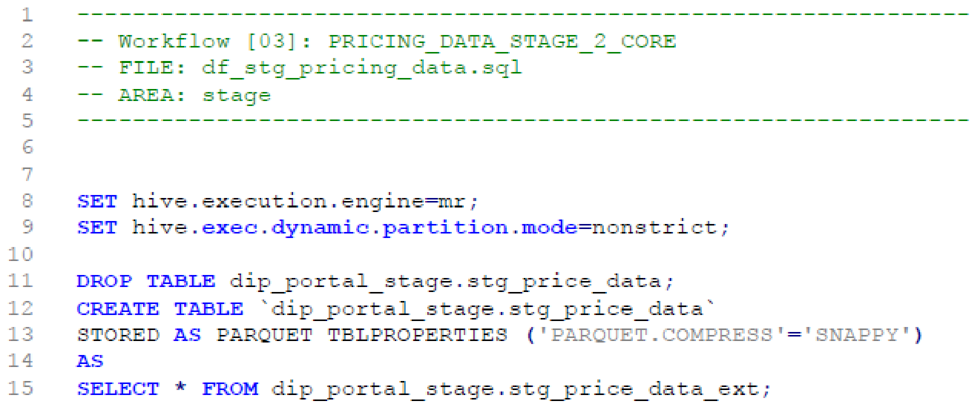
```sql
----------------------------------------------------------------
-- Workflow [02]: EVENTSTORE_STAGE_2_CORE
-- FILE: df_co_tl_eventstore_idkey_map_distinct.sql
-- AREA: core
----------------------------------------------------------------


SET hive.exec.dynamic.partition.mode=nonstrict;
SET hive.execution.engine=mr;
DROP TABLE dip_portal_core.co_tl_eventstore_idkey_map_distinct;
CREATE TABLE dip_portal_core.co_tl_eventstore_idkey_map_distinct
AS
WITH co_tl_eventstore_idkey_map_distinct_tmp
AS
(
SELECT DISTINCT visitor_id,
       customer_idkey_map
FROM   dip_portal_core.co_tl_eventstore_idkey_map
),
co_tl_eventstore_idkey_map_distinct_rank_tmp
AS
(
SELECT visitor_id,
       customer_idkey_map,
       rank() over (partition by visitor_id order by customer_idkey_map desc) as rank_visitor_id -- multiple customer_idkey_maps per visitor_id -> multiple logins per device
FROM   co_tl_eventstore_idkey_map_distinct_tmp
)
SELECT visitor_id,
       customer_idkey_map
FROM   co_tl_eventstore_idkey_map_distinct_rank_tmp
WHERE  rank_visitor_id=1; -- only use the first customer_idkey_map ignore all others
```

Appendix 18: WF[03]_stage_create



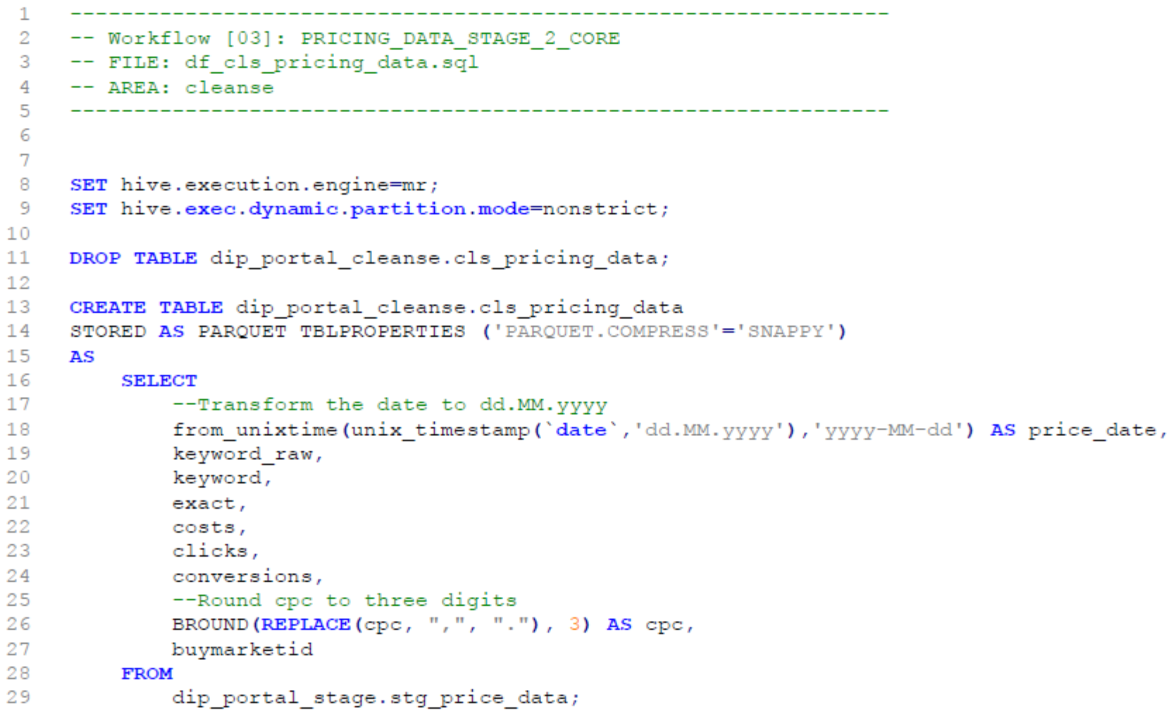
```
----------------------------------------------------------------
-- Workflow [03]: PRICING_DATA_STAGE_2_CORE
-- FILE: create_stg_price_data.sql
-- AREA: stage
----------------------------------------------------------------

CREATE TABLE `stg_price_data`(
  `date` string,
  `keyword_raw` string,
  `keyword_raw_state` string,
  `keyword` string,
  `exact` string,
  `costs` string,
  `clicks` string,
  `conversions` string,
  `cpc` string,
  `buymarketid` string)
ROW FORMAT SERDE
  'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
STORED AS INPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'
LOCATION
  'hdfs://ip-172-31-23-38.eu-central-1.compute.internal:8020/user/hive/warehouse/dip_portal_stage.db/stg_price_data'
TBLPROPERTIES (
  'PARQUET.COMPRESS'='SNAPPY',
  'last_modified_by'='hadoop',
  'last_modified_time'='1536069906',
  'numFiles'='0',
  'numRows'='4628567',
  'rawDataSize'='46285670',
  'totalSize'='0',
  'transient_lastDdlTime'='1536069907')

```

Appendix 19: WF[03]_stage_run



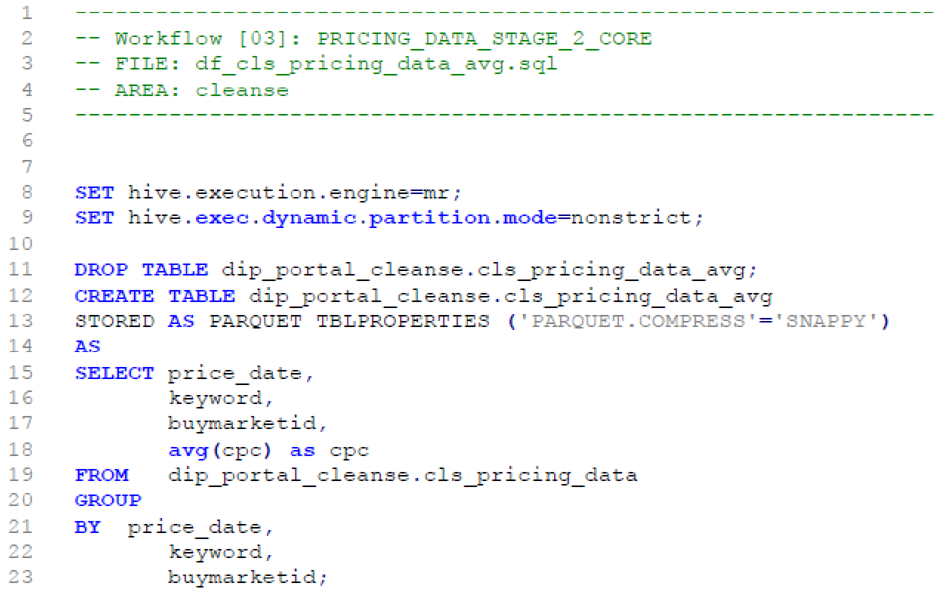
```sql
----------------------------------------------------------------
-- Workflow [03]: PRICING_DATA_STAGE_2_CORE
-- FILE: df_stg_pricing_data.sql
-- AREA: stage
----------------------------------------------------------------


SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;

DROP TABLE dip_portal_stage.stg_price_data;
CREATE TABLE `dip_portal_stage.stg_price_data`
STORED AS PARQUET TBLPROPERTIES ('PARQUET.COMPRESS'='SNAPPY')
AS
SELECT * FROM dip_portal_stage.stg_price_data_ext;
```

Appendix 20: WF[03]_cleanse_run01



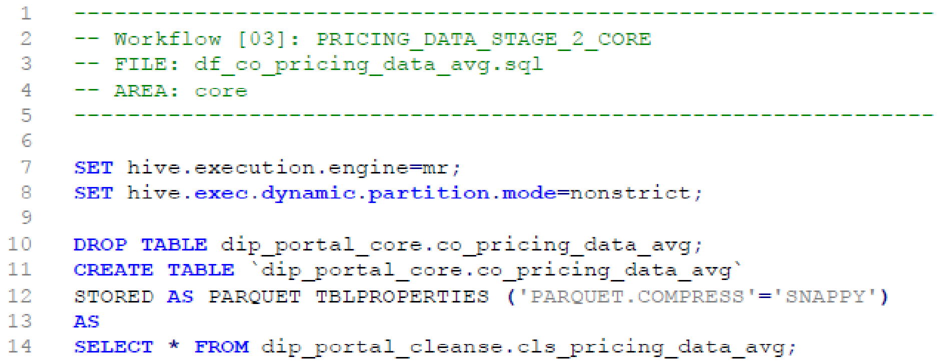
```sql
-----------------------------------------------------------------
-- Workflow [03]: PRICING_DATA_STAGE_2_CORE
-- FILE: df_cls_pricing_data.sql
-- AREA: cleanse
-----------------------------------------------------------------


SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;

DROP TABLE dip_portal_cleanse.cls_pricing_data;

CREATE TABLE dip_portal_cleanse.cls_pricing_data
STORED AS PARQUET TBLPROPERTIES ('PARQUET.COMPRESS'='SNAPPY')
AS
    SELECT
        --Transform the date to dd.MM.yyyy
        from_unixtime(unix_timestamp(`date`,'dd.MM.yyyy'),'yyyy-MM-dd') AS price_date,
        keyword_raw,
        keyword,
        exact,
        costs,
        clicks,
        conversions,
        --Round cpc to three digits
        BROUND(REPLACE(cpc, ",", "."), 3) AS cpc,
        buymarketid
    FROM
        dip_portal_stage.stg_price_data;
```

Appendix 21: WF[03]_cleanse_run02



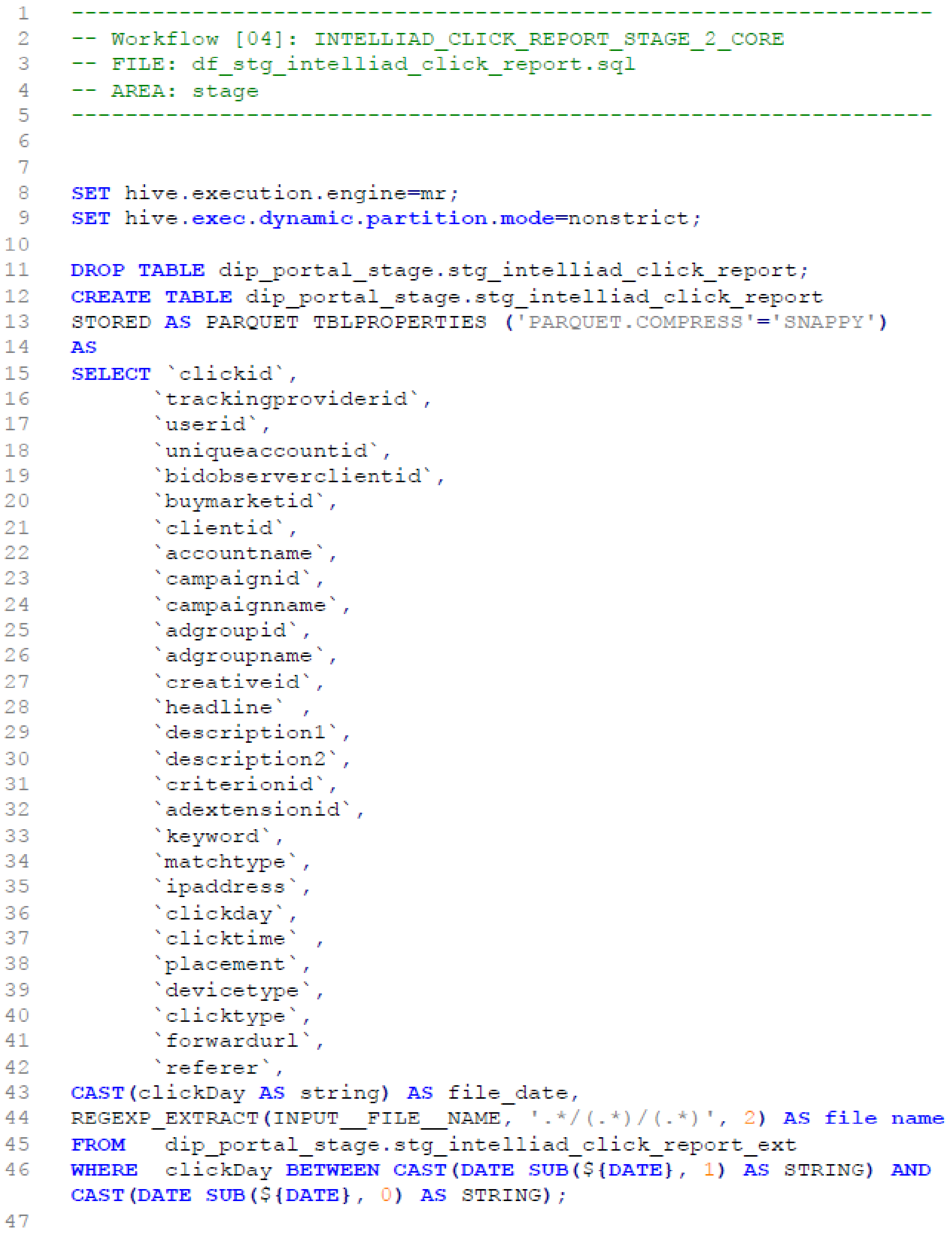
```sql
----------------------------------------------------------------
-- Workflow [03]: PRICING_DATA_STAGE_2_CORE
-- FILE: df_cls_pricing_data_avg.sql
-- AREA: cleanse
----------------------------------------------------------------


SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;

DROP TABLE dip_portal_cleanse.cls_pricing_data_avg;
CREATE TABLE dip_portal_cleanse.cls_pricing_data_avg
STORED AS PARQUET TBLPROPERTIES ('PARQUET.COMPRESS'='SNAPPY')
AS
SELECT price_date,
       keyword,
       buymarketid,
       avg(cpc) as cpc
FROM   dip_portal_cleanse.cls_pricing_data
GROUP
BY  price_date,
       keyword,
       buymarketid;
```

Appendix 22: WF[03]_core_run



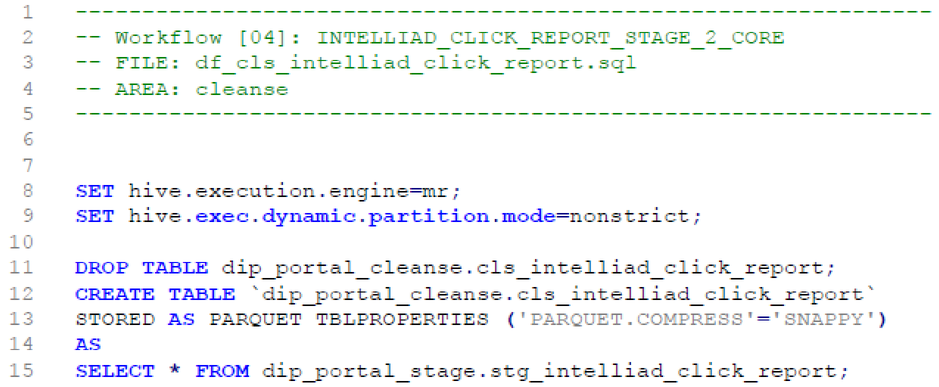
```sql
----------------------------------------------------------------
-- Workflow [03]: PRICING_DATA_STAGE_2_CORE
-- FILE: df_co_pricing_data_avg.sql
-- AREA: core
----------------------------------------------------------------

SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;

DROP TABLE dip_portal_core.co_pricing_data_avg;
CREATE TABLE `dip_portal_core.co_pricing_data_avg`
STORED AS PARQUET TBLPROPERTIES ('PARQUET.COMPRESS'='SNAPPY')
AS
SELECT * FROM dip_portal_cleanse.cls_pricing_data_avg;
```

Appendix 23: WF[04]_stage_run



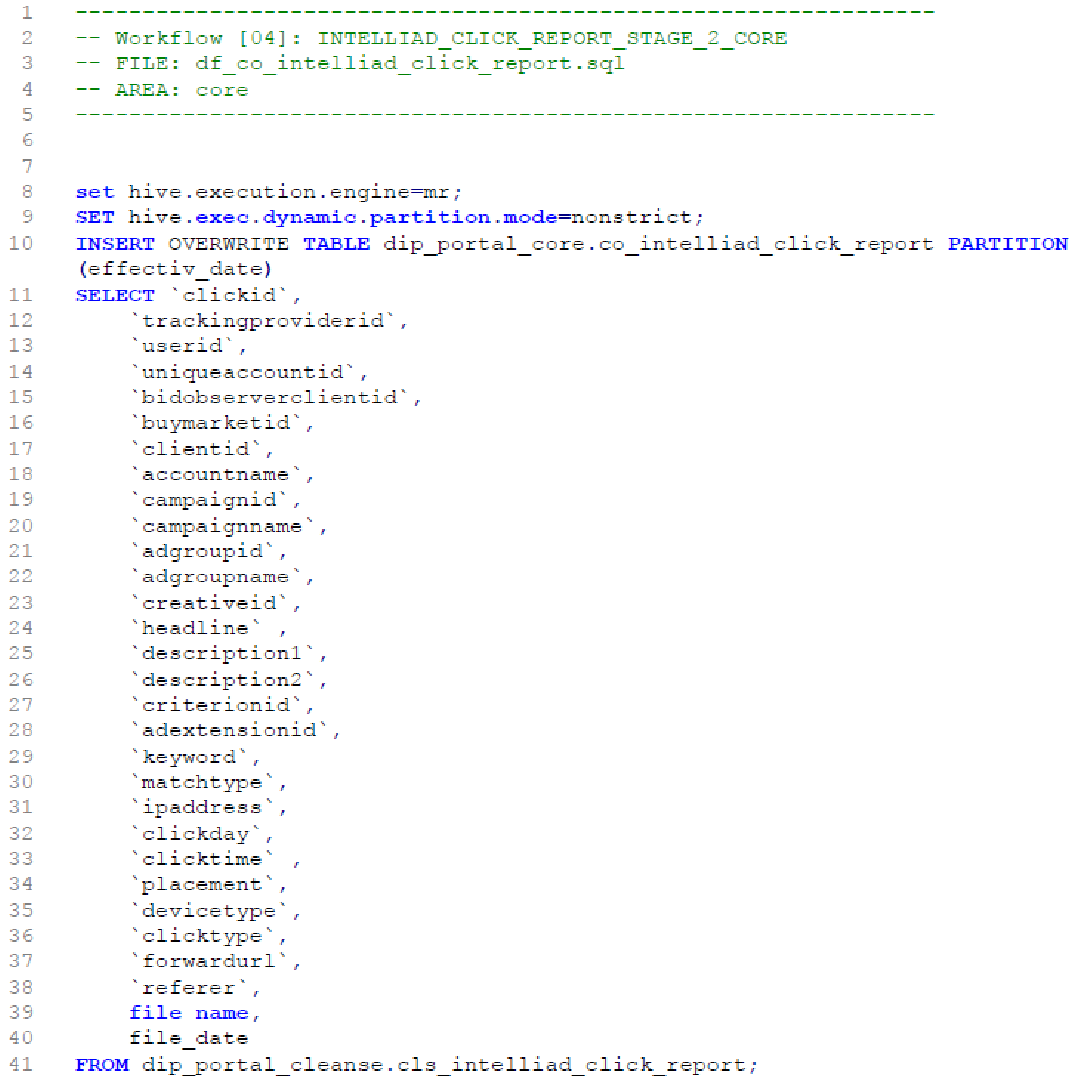
```
-----------------------------------------------------------------
-- Workflow [04]: INTELLIAD_CLICK_REPORT_STAGE_2_CORE
-- FILE: df_stg_intelliad_click_report.sql
-- AREA: stage
-----------------------------------------------------------------


SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;

DROP TABLE dip_portal_stage.stg_intelliad_click_report;
CREATE TABLE dip_portal_stage.stg_intelliad_click_report
STORED AS PARQUET TBLPROPERTIES ('PARQUET.COMPRESS'='SNAPPY')
AS
SELECT `clickid`,
       `trackingproviderid`,
       `userid`,
       `uniqueaccountid`,
       `bidobserverclientid`,
       `buymarketid`,
       `clientid`,
       `accountname`,
       `campaignid`,
       `campaignname`,
       `adgroupid`,
       `adgroupname`,
       `creativeid`,
       `headline` ,
       `description1`,
       `description2`,
       `criterionid`,
       `adextensionid`,
       `keyword`,
       `matchtype`,
       `ipaddress`,
       `clickday`,
       `clicktime` ,
       `placement`,
       `devicetype`,
       `clicktype`,
       `forwardurl`,
       `referer`,
CAST(clickDay AS string) AS file_date,
REGEXP_EXTRACT(INPUT__FILE__NAME, '.*/(.*)/(.*)', 2) AS file name
FROM   dip_portal_stage.stg_intelliad_click_report_ext
WHERE  clickDay BETWEEN CAST(DATE SUB(${DATE}, 1) AS STRING) AND
CAST(DATE SUB(${DATE}, 0) AS STRING);

```

Appendix 24: WF[04]_cleanse_run



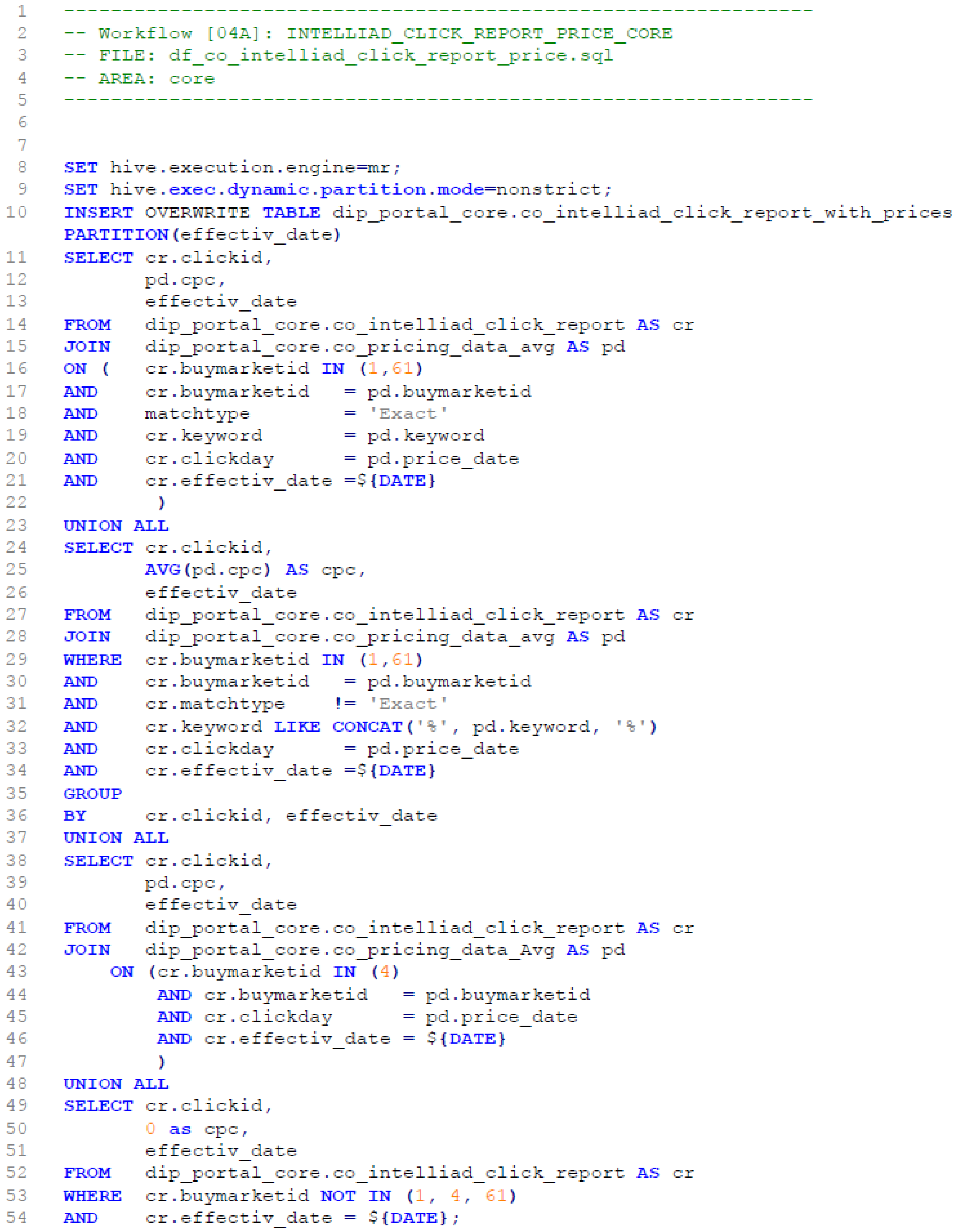
```
----------------------------------------------------------------
-- Workflow [04]: INTELLIAD_CLICK_REPORT_STAGE_2_CORE
-- FILE: df_cls_intelliad_click_report.sql
-- AREA: cleanse
----------------------------------------------------------------


SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;

DROP TABLE dip_portal_cleanse.cls_intelliad_click_report;
CREATE TABLE `dip_portal_cleanse.cls_intelliad_click_report`
STORED AS PARQUET TBLPROPERTIES ('PARQUET.COMPRESS'='SNAPPY')
AS
SELECT * FROM dip_portal_stage.stg_intelliad_click_report;
```

Appendix 25: WF[04]_core_run



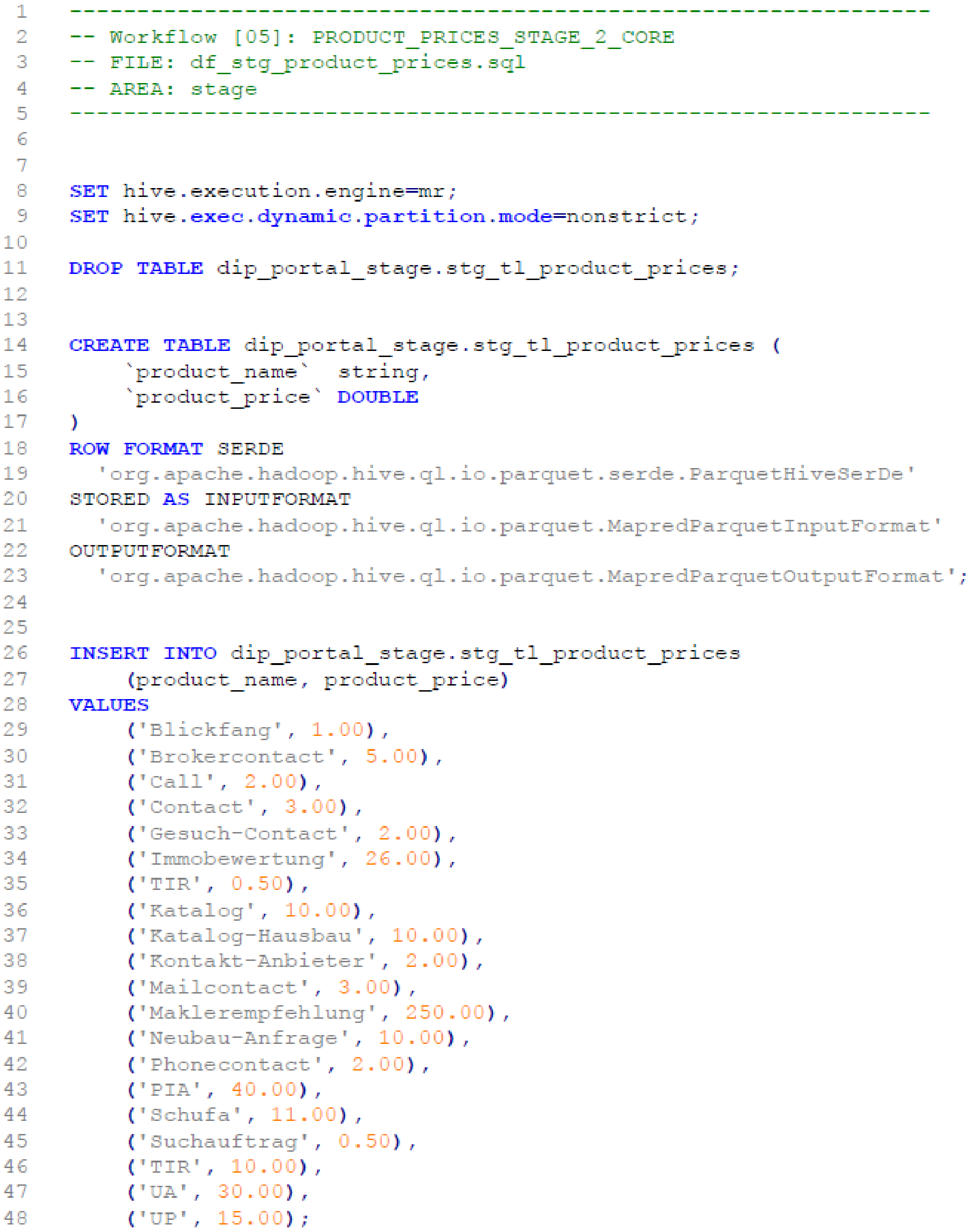
```
----------------------------------------------------------------
-- Workflow [04]: INTELLIAD_CLICK_REPORT_STAGE_2_CORE
-- FILE: df_co_intelliad_click_report.sql
-- AREA: core
----------------------------------------------------------------


set hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;
INSERT OVERWRITE TABLE dip_portal_core.co_intelliad_click_report PARTITION (effectiv_date)
SELECT `clickid`,
    `trackingproviderid`,
    `userid`,
    `uniqueaccountid`,
    `bidobserverclientid`,
    `buymarketid`,
    `clientid`,
    `accountname`,
    `campaignid`,
    `campaignname`,
    `adgroupid`,
    `adgroupname`,
    `creativeid`,
    `headline` ,
    `description1`,
    `description2`,
    `criterionid`,
    `adextensionid`,
    `keyword`,
    `matchtype`,
    `ipaddress`,
    `clickday`,
    `clicktime`  ,
    `placement`,
    `devicetype`,
    `clicktype`,
    `forwardurl`,
    `referer`,
    file name,
    file_date
FROM dip_portal_cleanse.cls_intelliad_click_report;
```

Appendix 26: WF[04A]_core_run



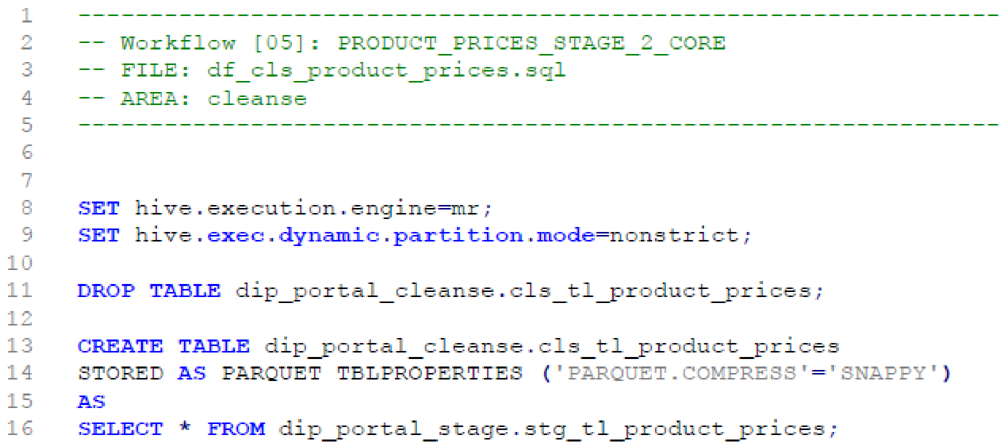
```
-----------------------------------------------------------------
-- Workflow [04A]: INTELLIAD_CLICK_REPORT_PRICE_CORE
-- FILE: df_co_intelliad_click_report_price.sql
-- AREA: core
-----------------------------------------------------------------


SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;
INSERT OVERWRITE TABLE dip_portal_core.co_intelliad_click_report_with_prices PARTITION(effectiv_date)
SELECT cr.clickid,
       pd.cpc,
       effectiv_date
FROM   dip_portal_core.co_intelliad_click_report AS cr
JOIN   dip_portal_core.co_pricing_data_avg AS pd
ON (   cr.buymarketid IN (1,61)
AND    cr.buymarketid   = pd.buymarketid
AND    matchtype        = 'Exact'
AND    cr.keyword       = pd.keyword
AND    cr.clickday      = pd.price_date
AND    cr.effectiv_date =${DATE}
        )
UNION ALL
SELECT cr.clickid,
       AVG(pd.cpc) AS cpc,
       effectiv_date
FROM   dip_portal_core.co_intelliad_click_report AS cr
JOIN   dip_portal_core.co_pricing_data_avg AS pd
WHERE  cr.buymarketid IN (1,61)
AND    cr.buymarketid   = pd.buymarketid
AND    cr.matchtype    != 'Exact'
AND    cr.keyword LIKE CONCAT('%', pd.keyword, '%')
AND    cr.clickday      = pd.price_date
AND    cr.effectiv_date =${DATE}
GROUP
BY     cr.clickid, effectiv_date
UNION ALL
SELECT cr.clickid,
       pd.cpc,
       effectiv_date
FROM   dip_portal_core.co_intelliad_click_report AS cr
JOIN   dip_portal_core.co_pricing_data_Avg AS pd
    ON (cr.buymarketid IN (4)
        AND cr.buymarketid   = pd.buymarketid
        AND cr.clickday      = pd.price_date
        AND cr.effectiv_date = ${DATE}
        )
UNION ALL
SELECT cr.clickid,
       0 as cpc,
       effectiv_date
FROM   dip_portal_core.co_intelliad_click_report AS cr
WHERE  cr.buymarketid NOT IN (1, 4, 61)
AND    cr.effectiv_date = ${DATE};
```

Appendix 27: WF[05]_stage_run



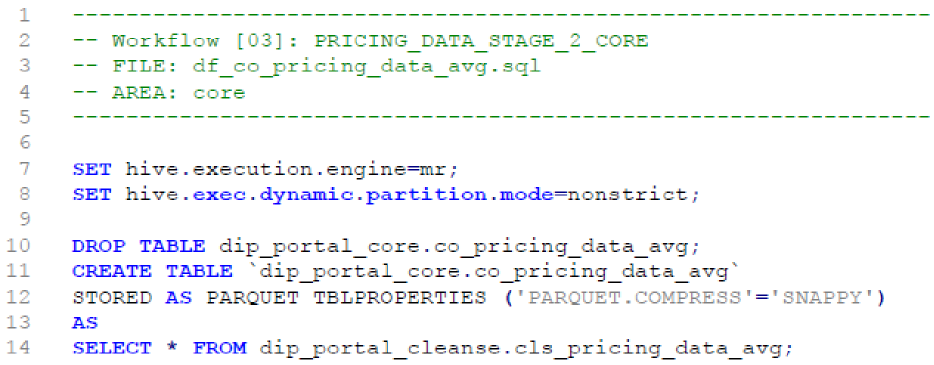
```sql
----------------------------------------------------------------
-- Workflow [05]: PRODUCT_PRICES_STAGE_2_CORE
-- FILE: df_stg_product_prices.sql
-- AREA: stage
----------------------------------------------------------------


SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;

DROP TABLE dip_portal_stage.stg_tl_product_prices;


CREATE TABLE dip_portal_stage.stg_tl_product_prices (
    `product_name`  string,
    `product_price` DOUBLE
)
ROW FORMAT SERDE
  'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
STORED AS INPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat';


INSERT INTO dip_portal_stage.stg_tl_product_prices
    (product_name, product_price)
VALUES
    ('Blickfang', 1.00),
    ('Brokercontact', 5.00),
    ('Call', 2.00),
    ('Contact', 3.00),
    ('Gesuch-Contact', 2.00),
    ('Immobewertung', 26.00),
    ('TIR', 0.50),
    ('Katalog', 10.00),
    ('Katalog-Hausbau', 10.00),
    ('Kontakt-Anbieter', 2.00),
    ('Mailcontact', 3.00),
    ('Maklerempfehlung', 250.00),
    ('Neubau-Anfrage', 10.00),
    ('Phonecontact', 2.00),
    ('PIA', 40.00),
    ('Schufa', 11.00),
    ('Suchauftrag', 0.50),
    ('TIR', 10.00),
    ('UA', 30.00),
    ('UP', 15.00);
```

Appendix 28: WF[05]_cleanse_run



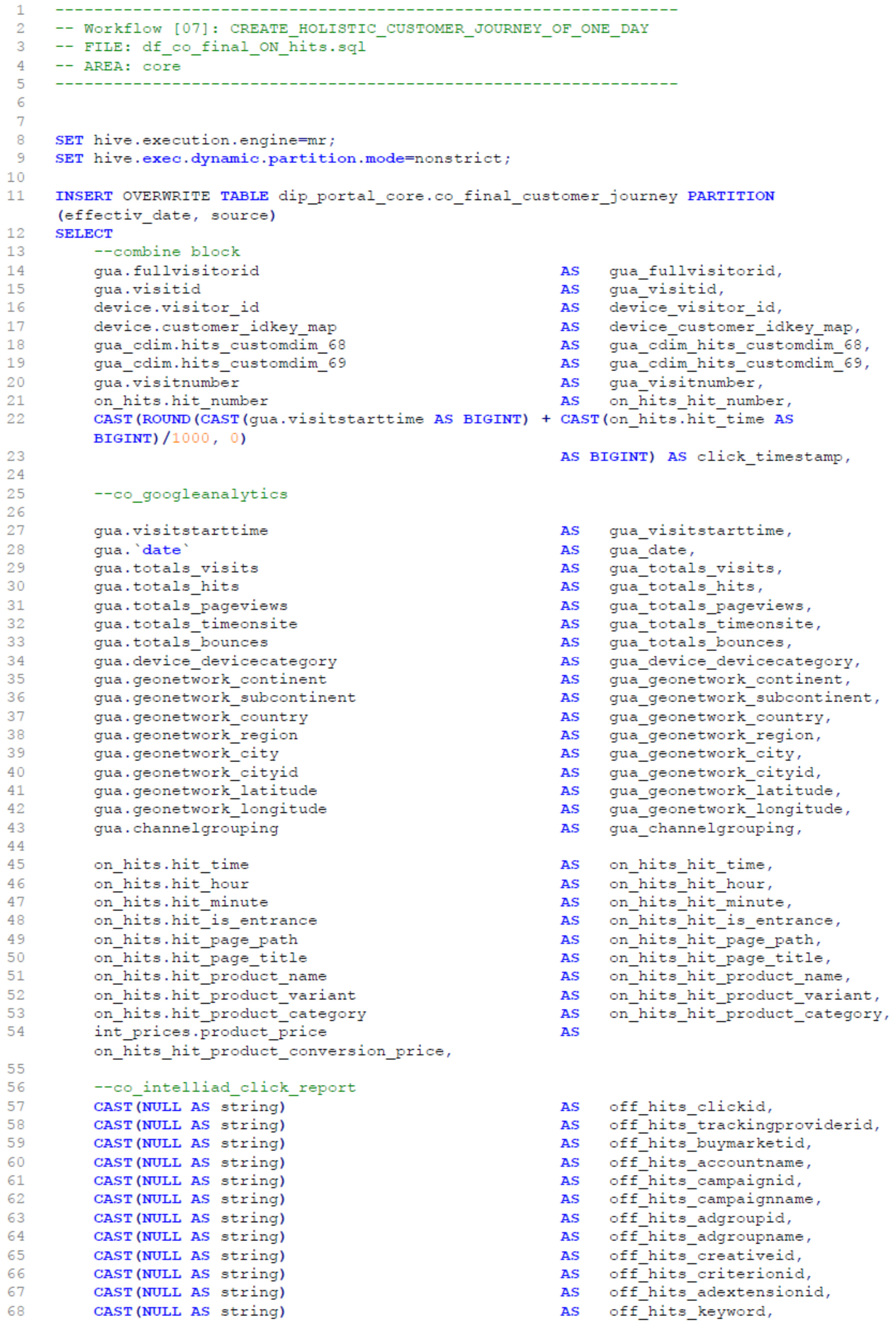
```sql
----------------------------------------------------------------
-- Workflow [05]: PRODUCT_PRICES_STAGE_2_CORE
-- FILE: df_cls_product_prices.sql
-- AREA: cleanse
----------------------------------------------------------------


SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;

DROP TABLE dip_portal_cleanse.cls_tl_product_prices;

CREATE TABLE dip_portal_cleanse.cls_tl_product_prices
STORED AS PARQUET TBLPROPERTIES ('PARQUET.COMPRESS'='SNAPPY')
AS
SELECT * FROM dip_portal_stage.stg_tl_product_prices;
```

Appendix 29: WF[05]_core_run

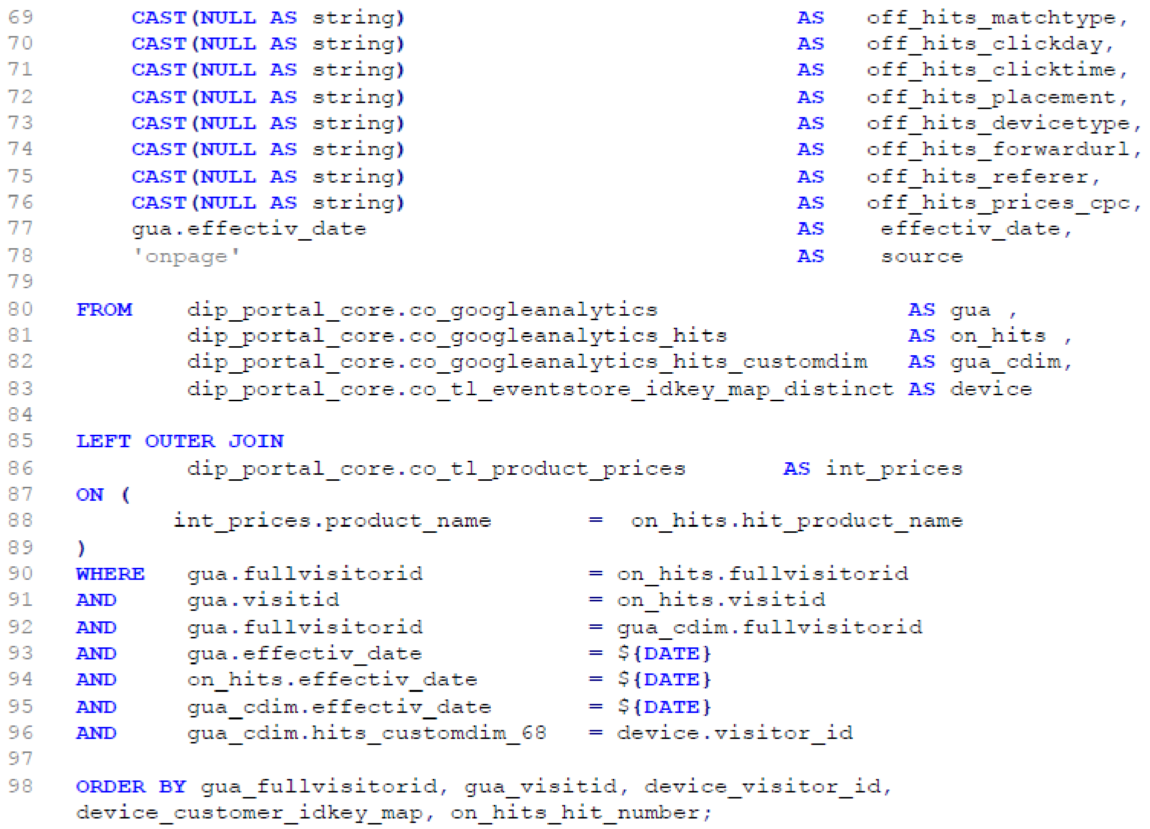


```
----------------------------------------------------------------
-- Workflow [03]: PRICING_DATA_STAGE_2_CORE
-- FILE: df_co_pricing_data_avg.sql
-- AREA: core
----------------------------------------------------------------

SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;

DROP TABLE dip_portal_core.co_pricing_data_avg;
CREATE TABLE `dip_portal_core.co_pricing_data_avg`
STORED AS PARQUET TBLPROPERTIES ('PARQUET.COMPRESS'='SNAPPY')
AS
SELECT * FROM dip_portal_cleanse.cls_pricing_data_avg;
```

Appendix 30: WF[07]_core_run01

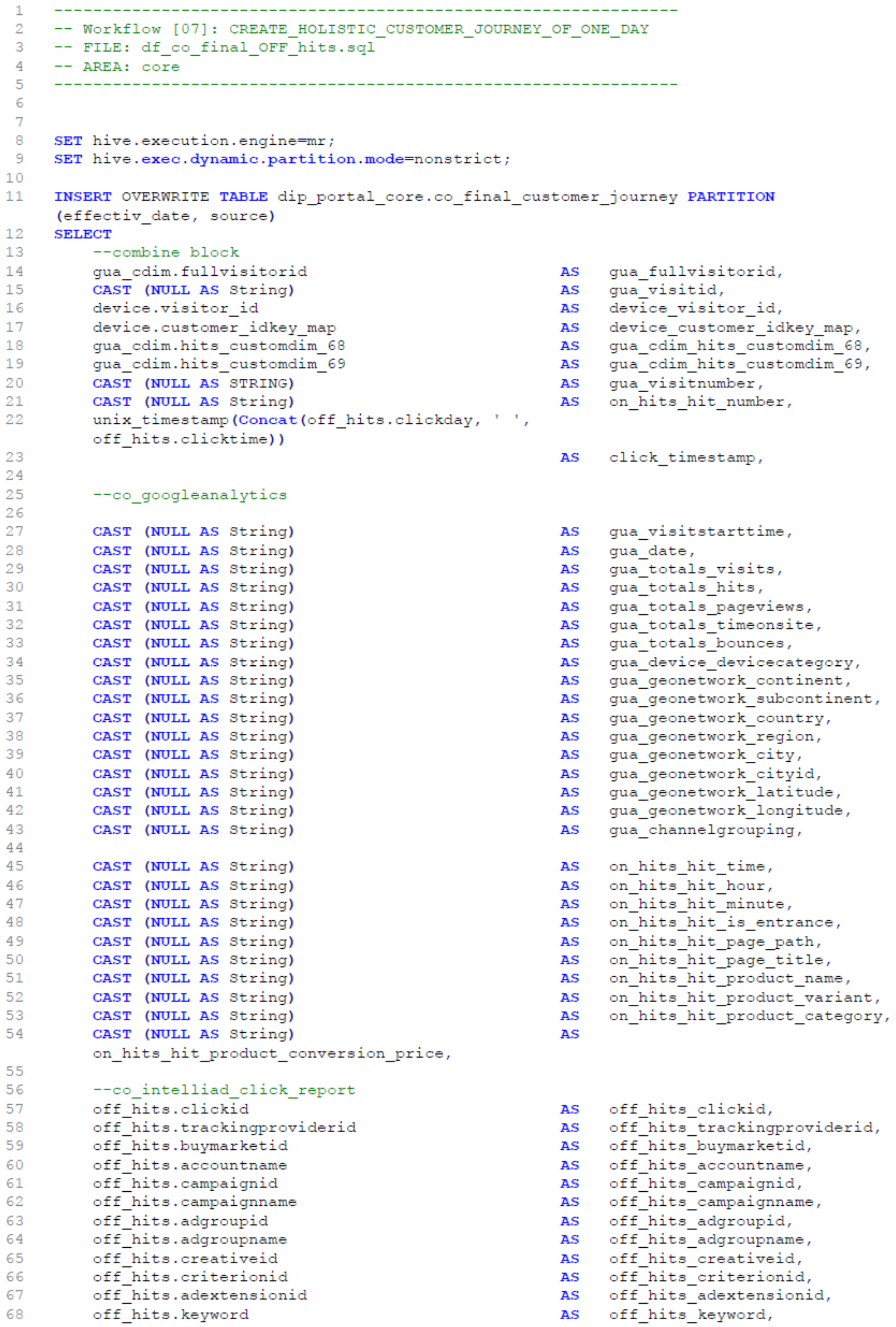


```
-------------------------------------------------------------------
-- Workflow [07]: CREATE_HOLISTIC_CUSTOMER_JOURNEY_OF_ONE_DAY
-- FILE: df_co_final_ON_hits.sql
-- AREA: core
-------------------------------------------------------------------


SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;

INSERT OVERWRITE TABLE dip_portal_core.co_final_customer_journey PARTITION (effectiv_date, source)
SELECT
    --combine block
    gua.fullvisitorid                                AS   gua_fullvisitorid,
    gua.visitid                                      AS   gua_visitid,
    device.visitor_id                                AS   device_visitor_id,
    device.customer_idkey_map                        AS   device_customer_idkey_map,
    gua_cdim.hits_customdim_68                       AS   gua_cdim_hits_customdim_68,
    gua_cdim.hits_customdim_69                       AS   gua_cdim_hits_customdim_69,
    gua.visitnumber                                  AS   gua_visitnumber,
    on_hits.hit_number                               AS   on_hits_hit_number,
    CAST(ROUND(CAST(gua.visitstarttime AS BIGINT) + CAST(on_hits.hit_time AS BIGINT)/1000, 0)
                                                     AS BIGINT) AS click_timestamp,

    --co_googleanalytics

    gua.visitstarttime                               AS   gua_visitstarttime,
    gua.`date`                                       AS   gua_date,
    gua.totals_visits                                AS   gua_totals_visits,
    gua.totals_hits                                  AS   gua_totals_hits,
    gua.totals_pageviews                             AS   gua_totals_pageviews,
    gua.totals_timeonsite                            AS   gua_totals_timeonsite,
    gua.totals_bounces                               AS   gua_totals_bounces,
    gua.device_devicecategory                        AS   gua_device_devicecategory,
    gua.geonetwork_continent                         AS   gua_geonetwork_continent,
    gua.geonetwork_subcontinent                      AS   gua_geonetwork_subcontinent,
    gua.geonetwork_country                           AS   gua_geonetwork_country,
    gua.geonetwork_region                            AS   gua_geonetwork_region,
    gua.geonetwork_city                              AS   gua_geonetwork_city,
    gua.geonetwork_cityid                            AS   gua_geonetwork_cityid,
    gua.geonetwork_latitude                          AS   gua_geonetwork_latitude,
    gua.geonetwork_longitude                         AS   gua_geonetwork_longitude,
    gua.channelgrouping                              AS   gua_channelgrouping,

    on_hits.hit_time                                 AS   on_hits_hit_time,
    on_hits.hit_hour                                 AS   on_hits_hit_hour,
    on_hits.hit_minute                               AS   on_hits_hit_minute,
    on_hits.hit_is_entrance                          AS   on_hits_hit_is_entrance,
    on_hits.hit_page_path                            AS   on_hits_hit_page_path,
    on_hits.hit_page_title                           AS   on_hits_hit_page_title,
    on_hits.hit_product_name                         AS   on_hits_hit_product_name,
    on_hits.hit_product_variant                      AS   on_hits_hit_product_variant,
    on_hits.hit_product_category                     AS   on_hits_hit_product_category,
    int_prices.product_price                         AS   on_hits_hit_product_conversion_price,

    --co_intelliad_click_report
    CAST(NULL AS string)                             AS   off_hits_clickid,
    CAST(NULL AS string)                             AS   off_hits_trackingproviderid,
    CAST(NULL AS string)                             AS   off_hits_buymarketid,
    CAST(NULL AS string)                             AS   off_hits_accountname,
    CAST(NULL AS string)                             AS   off_hits_campaignid,
    CAST(NULL AS string)                             AS   off_hits_campaignname,
    CAST(NULL AS string)                             AS   off_hits_adgroupid,
    CAST(NULL AS string)                             AS   off_hits_adgroupname,
    CAST(NULL AS string)                             AS   off_hits_creativeid,
    CAST(NULL AS string)                             AS   off_hits_criterionid,
    CAST(NULL AS string)                             AS   off_hits_adextensionid,
    CAST(NULL AS string)                             AS   off_hits_keyword,
```



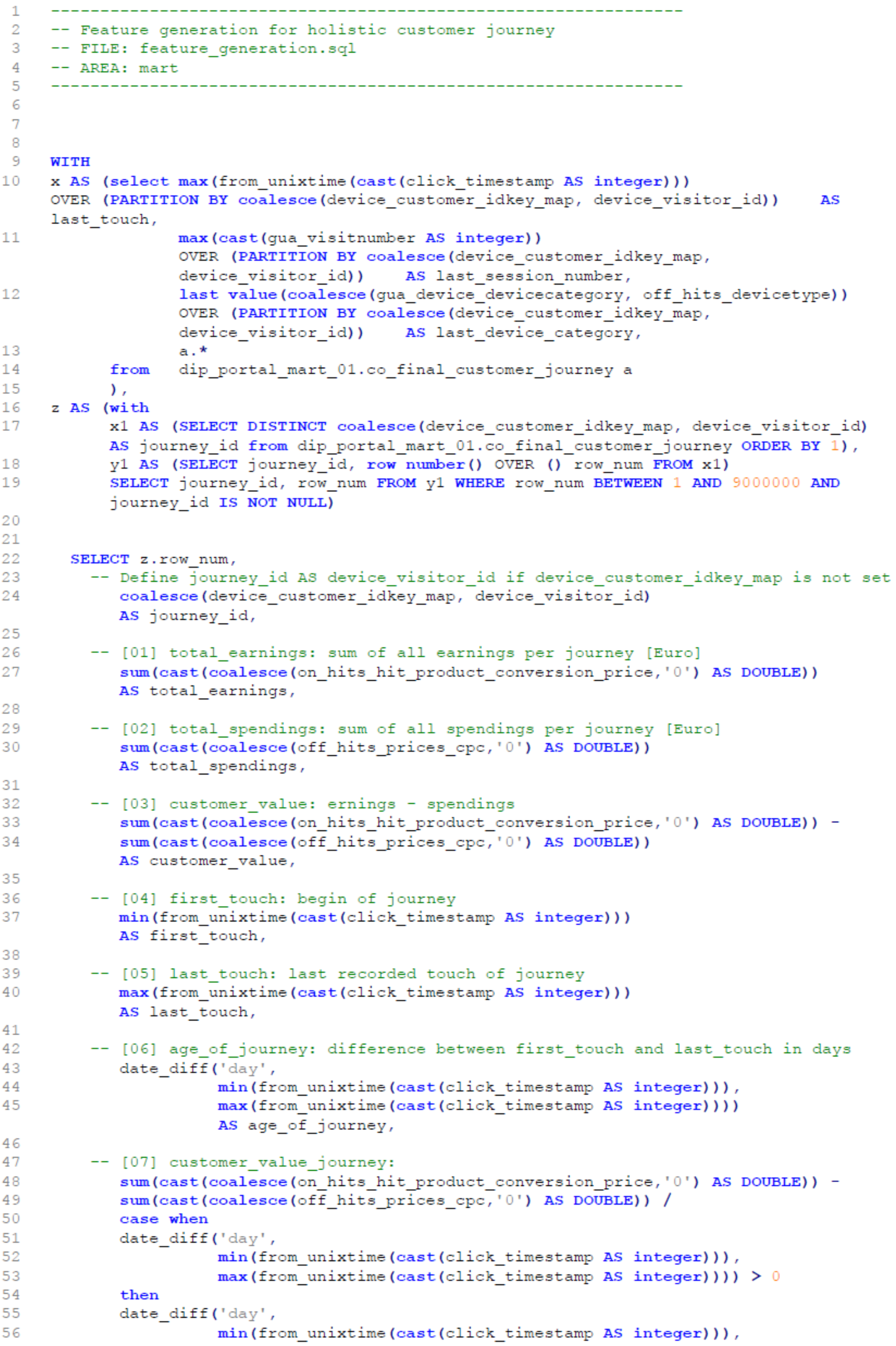
```
    CAST(NULL AS string)                            AS   off_hits_matchtype,
    CAST(NULL AS string)                            AS   off_hits_clickday,
    CAST(NULL AS string)                            AS   off_hits_clicktime,
    CAST(NULL AS string)                            AS   off_hits_placement,
    CAST(NULL AS string)                            AS   off_hits_devicetype,
    CAST(NULL AS string)                            AS   off_hits_forwardurl,
    CAST(NULL AS string)                            AS   off_hits_referer,
    CAST(NULL AS string)                            AS   off_hits_prices_cpc,
    gua.effectiv_date                               AS    effectiv_date,
    'onpage'                                        AS    source

FROM    dip_portal_core.co_googleanalytics                  AS gua ,
        dip_portal_core.co_googleanalytics_hits             AS on_hits ,
        dip_portal_core.co_googleanalytics_hits_customdim   AS gua_cdim,
        dip_portal_core.co_tl_eventstore_idkey_map_distinct AS device

LEFT OUTER JOIN
        dip_portal_core.co_tl_product_prices       AS int_prices
ON (
       int_prices.product_name       =  on_hits.hit_product_name
)
WHERE   gua.fullvisitorid            = on_hits.fullvisitorid
AND     gua.visitid                  = on_hits.visitid
AND     gua.fullvisitorid            = gua_cdim.fullvisitorid
AND     gua.effectiv_date            = ${DATE}
AND     on_hits.effectiv_date        = ${DATE}
AND     gua_cdim.effectiv_date       = ${DATE}
AND     gua_cdim.hits_customdim_68   = device.visitor_id

ORDER BY gua_fullvisitorid, gua_visitid, device_visitor_id,
device_customer_idkey_map, on_hits_hit_number;
```

Appendix 31: WF[07]_core_run02



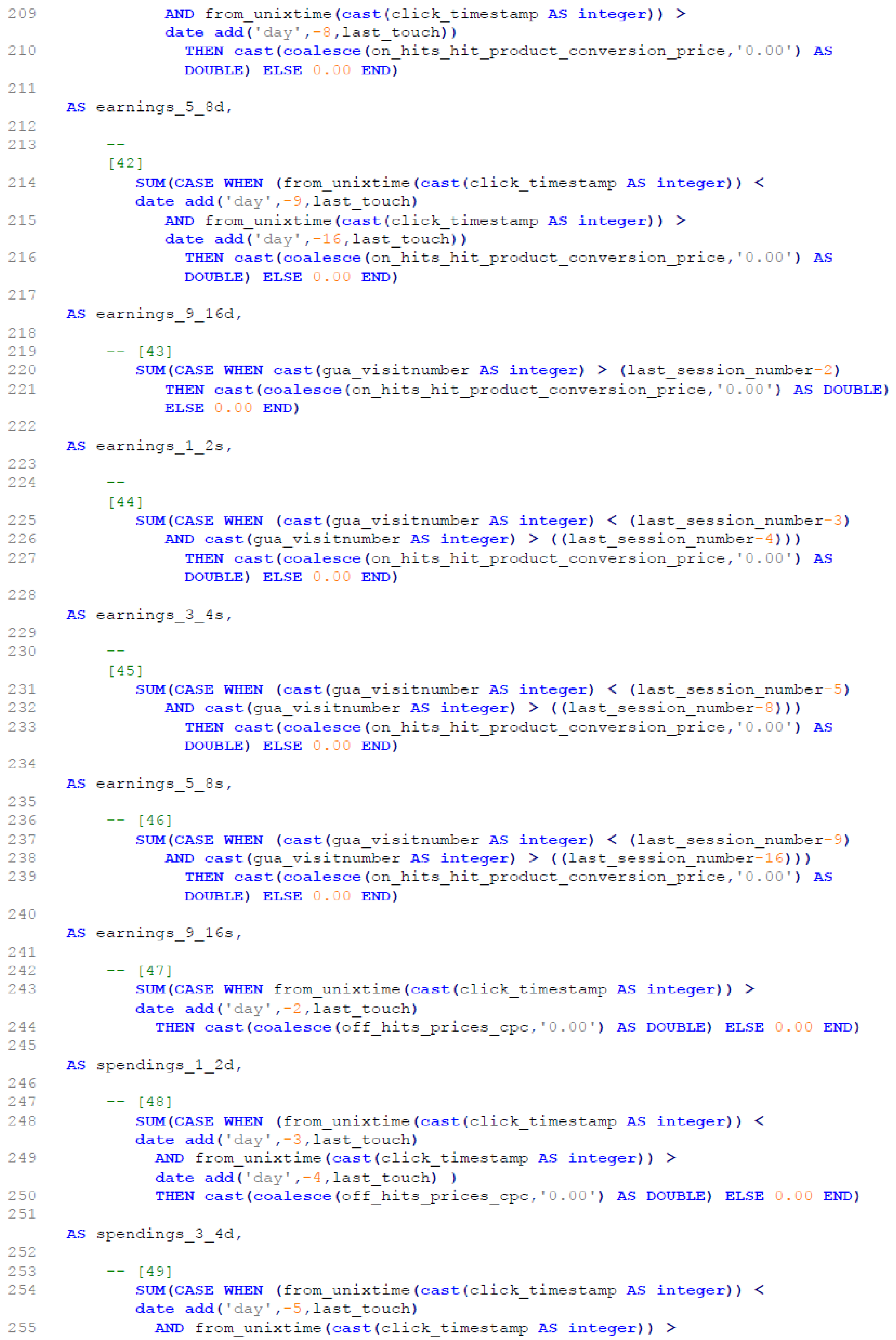
```
-----------------------------------------------------------------
-- Workflow [07]: CREATE_HOLISTIC_CUSTOMER_JOURNEY_OF_ONE_DAY
-- FILE: df_co_final_OFF_hits.sql
-- AREA: core
-----------------------------------------------------------------


SET hive.execution.engine=mr;
SET hive.exec.dynamic.partition.mode=nonstrict;

INSERT OVERWRITE TABLE dip_portal_core.co_final_customer_journey PARTITION
(effectiv_date, source)
SELECT
    --combine block
    gua_cdim.fullvisitorid                          AS   gua_fullvisitorid,
    CAST (NULL AS String)                           AS   gua_visitid,
    device.visitor_id                               AS   device_visitor_id,
    device.customer_idkey_map                       AS   device_customer_idkey_map,
    gua_cdim.hits_customdim_68                      AS   gua_cdim_hits_customdim_68,
    gua_cdim.hits_customdim_69                      AS   gua_cdim_hits_customdim_69,
    CAST (NULL AS STRING)                           AS   gua_visitnumber,
    CAST (NULL AS String)                           AS   on_hits_hit_number,
    unix_timestamp(Concat(off_hits.clickday, ' ',
    off_hits.clicktime))
                                                    AS   click_timestamp,

    --co_googleanalytics

    CAST (NULL AS String)                           AS   gua_visitstarttime,
    CAST (NULL AS String)                           AS   gua_date,
    CAST (NULL AS String)                           AS   gua_totals_visits,
    CAST (NULL AS String)                           AS   gua_totals_hits,
    CAST (NULL AS String)                           AS   gua_totals_pageviews,
    CAST (NULL AS String)                           AS   gua_totals_timeonsite,
    CAST (NULL AS String)                           AS   gua_totals_bounces,
    CAST (NULL AS String)                           AS   gua_device_devicecategory,
    CAST (NULL AS String)                           AS   gua_geonetwork_continent,
    CAST (NULL AS String)                           AS   gua_geonetwork_subcontinent,
    CAST (NULL AS String)                           AS   gua_geonetwork_country,
    CAST (NULL AS String)                           AS   gua_geonetwork_region,
    CAST (NULL AS String)                           AS   gua_geonetwork_city,
    CAST (NULL AS String)                           AS   gua_geonetwork_cityid,
    CAST (NULL AS String)                           AS   gua_geonetwork_latitude,
    CAST (NULL AS String)                           AS   gua_geonetwork_longitude,
    CAST (NULL AS String)                           AS   gua_channelgrouping,

    CAST (NULL AS String)                           AS   on_hits_hit_time,
    CAST (NULL AS String)                           AS   on_hits_hit_hour,
    CAST (NULL AS String)                           AS   on_hits_hit_minute,
    CAST (NULL AS String)                           AS   on_hits_hit_is_entrance,
    CAST (NULL AS String)                           AS   on_hits_hit_page_path,
    CAST (NULL AS String)                           AS   on_hits_hit_page_title,
    CAST (NULL AS String)                           AS   on_hits_hit_product_name,
    CAST (NULL AS String)                           AS   on_hits_hit_product_variant,
    CAST (NULL AS String)                           AS   on_hits_hit_product_category,
    CAST (NULL AS String)                           AS
    on_hits_hit_product_conversion_price,

    --co_intelliad_click_report
    off_hits.clickid                                AS   off_hits_clickid,
    off_hits.trackingproviderid                     AS   off_hits_trackingproviderid,
    off_hits.buymarketid                            AS   off_hits_buymarketid,
    off_hits.accountname                            AS   off_hits_accountname,
    off_hits.campaignid                             AS   off_hits_campaignid,
    off_hits.campaignname                           AS   off_hits_campaignname,
    off_hits.adgroupid                              AS   off_hits_adgroupid,
    off_hits.adgroupname                            AS   off_hits_adgroupname,
    off_hits.creativeid                             AS   off_hits_creativeid,
    off_hits.criterionid                            AS   off_hits_criterionid,
    off_hits.adextensionid                          AS   off_hits_adextensionid,
    off_hits.keyword                                AS   off_hits_keyword,
    off_hits.matchtype                                AS   off_hits_matchtype,
    off_hits.clickday                                 AS   off_hits_clickday,
    off_hits.clicktime                                AS   off_hits_clicktime,
    off_hits.placement                                AS   off_hits_placement,
    off_hits.devicetype                               AS   off_hits_devicetype,
    off_hits.forwardurl                               AS   off_hits_forwardurl,
    off_hits.referer                                  AS   off_hits_referer,
    off_hits_prices.cpc                               AS   off_hits_prices_cpc,
    off_hits.effectiv_date                            AS   effectiv_date,
    'offpage'                                         AS   source

FROM    dip_portal_core.co_googleanalytics_hits_customdim     AS gua_cdim,
        dip_portal_core.co_tl_eventstore_idkey_map_distinct   AS device,
        dip_portal_core.co_intelliad_click_report             AS off_hits,
        dip_portal_core.co_intelliad_click_report_with_prices AS off_hits_prices

WHERE  gua_cdim.effectiv_date           = ${DATE}
    AND off_hits.effectiv_date          = ${DATE}
    AND off_hits_prices.effectiv_date   = ${DATE}                -- map the custom dim
    AND gua_cdim.hits_customdim_68      = device.visitor_id      -- map tealium visitor_id
    AND  gua_cdim.hits_customdim_69     = off_hits.userid        -- map intelliad userid
    AND off_hits_prices.clickid         = off_hits.clickid       -- map intelliad actions with cpc prices

ORDER BY gua_fullvisitorid, off_hits_clickday, off_hits_clicktime;
```

Appendix 32: Feature generation sql script



```sql
----------------------------------------------------------------
-- Feature generation for holistic customer journey
-- FILE: feature_generation.sql
-- AREA: mart
----------------------------------------------------------------



WITH
x AS (select max(from_unixtime(cast(click_timestamp AS integer)))
OVER (PARTITION BY coalesce(device_customer_idkey_map, device_visitor_id))    AS
last_touch,
             max(cast(gua_visitnumber AS integer))
             OVER (PARTITION BY coalesce(device_customer_idkey_map,
             device_visitor_id))    AS last_session_number,
             last value(coalesce(gua_device_devicecategory, off_hits_devicetype))
             OVER (PARTITION BY coalesce(device_customer_idkey_map,
             device_visitor_id))    AS last_device_category,
             a.*
      from   dip_portal_mart_01.co_final_customer_journey a
      ),
z AS (with
      x1 AS (SELECT DISTINCT coalesce(device_customer_idkey_map, device_visitor_id)
      AS journey_id from dip_portal_mart_01.co_final_customer_journey ORDER BY 1),
      y1 AS (SELECT journey_id, row number() OVER () row_num FROM x1)
      SELECT journey_id, row_num FROM y1 WHERE row_num BETWEEN 1 AND 9000000 AND
      journey_id IS NOT NULL)


  SELECT z.row_num,
    -- Define journey_id AS device_visitor_id if device_customer_idkey_map is not set
       coalesce(device_customer_idkey_map, device_visitor_id)
       AS journey_id,

    -- [01] total_earnings: sum of all earnings per journey [Euro]
       sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
       AS total_earnings,

    -- [02] total_spendings: sum of all spendings per journey [Euro]
       sum(cast(coalesce(off_hits_prices_cpc,'0') AS DOUBLE))
       AS total_spendings,

    -- [03] customer_value: ernings - spendings
       sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE)) -
       sum(cast(coalesce(off_hits_prices_cpc,'0') AS DOUBLE))
       AS customer_value,

    -- [04] first_touch: begin of journey
       min(from_unixtime(cast(click_timestamp AS integer)))
       AS first_touch,

    -- [05] last_touch: last recorded touch of journey
       max(from_unixtime(cast(click_timestamp AS integer)))
       AS last_touch,

    -- [06] age_of_journey: difference between first_touch and last_touch in days
       date_diff('day',
                 min(from_unixtime(cast(click_timestamp AS integer))),
                 max(from_unixtime(cast(click_timestamp AS integer))))
                 AS age_of_journey,

    -- [07] customer_value_journey:
       sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE)) -
       sum(cast(coalesce(off_hits_prices_cpc,'0') AS DOUBLE)) /
       case when
       date_diff('day',
                 min(from_unixtime(cast(click_timestamp AS integer))),
                 max(from_unixtime(cast(click_timestamp AS integer)))) > 0
       then
       date_diff('day',
                 min(from_unixtime(cast(click_timestamp AS integer))),
                 max(from_unixtime(cast(click_timestamp AS integer))))
       else 1 end
       AS customer_value_journey,

    -- [08] session_cnt: count of onpage sessions
       count(distinct gua_visitnumber)
       AS session_cnt,


    -- [09] is_logged_in : is defined if user hAS entered its email address
       case when count(distinct device_customer_idkey_map) > 0 then 1 else 0 end
       AS is_logged_in,


    -- [10] is_cross_device_user: is using different device classes
       case
         when count(distinct coalesce(lower(gua_device_devicecategory),
                                      lower(off_hits_devicetype))) > 1
         then 1 else 0 end
         AS is_cross_device_user,

    -- [11] avg_events_per_session: all events (on- and offpage) /
    session_cnt
       COUNT(*) / count(distinct coalesce(gua_visitnumber,'0'))
       AS avg_events_per_session,

    -- [12] hit_cnt: all onpage and offpage hits
       COUNT(*)
       AS total_hit_cnt,

    -- [13] overall_journey_cnt: count all journeys
       COUNT(COUNT(*)) OVER (PARTITION BY 'ALL')
       AS overall_journey_cnt,

    -- [14]
       sum(sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE)))
         OVER (PARTITION BY 'ALL') /
       COUNT(COUNT(*)) OVER (PARTITION BY 'ALL')
       AS overall_avg_earning_per_journey,

    -- [15]
       (sum(sum(cast(coalesce(off_hits_prices_cpc,'0') AS DOUBLE)))
         OVER (PARTITION BY 'ALL') /
       COUNT(COUNT(*)) OVER (PARTITION BY 'ALL'))
       AS overall_avg_spendings_per_journey,

    -- [16]
       (COUNT(distinct device_visitor_id)*1.00) /
       COUNT(COUNT(distinct device_visitor_id)) OVER (PARTITION BY 'ALL')
       AS percentage_of_overall_mean_device_cnt_per_journey,

    -- [17]
       (count(distinct gua_visitnumber)*1.00) /
       COUNT(COUNT(distinct gua_visitnumber)) OVER (PARTITION BY 'ALL')
       AS percentage_of_overall_mean_session_cnt_per_journey,

    -- [18]
       array_union(array_distinct(array_agg(lower(gua_device_devicecategory))),
                   array_distinct(array_agg(lower(off_hits_devicetype))))
                   AS device_array,

    -- [19]
       case when contains(
         array_union(array_distinct(array_agg(lower(gua_device_devicecategory))),
         array_distinct(array_agg(lower(off_hits_devicetype)))),
         'desktop')
         then 1 else 0 end
         AS uses_desktop,

    -- [20]
       case when contains(
         array_union(array_distinct(array_agg(lower(gua_device_devicecategory))),
         array_distinct(array_agg(lower(off_hits_devicetype)))),
         'mobile')
         then 1 else 0 end
         AS uses_mobile,

    -- [21]
       case when contains(
         array_union(array_distinct(array_agg(lower(gua_device_devicecategory))),
         array_distinct(array_agg(lower(off_hits_devicetype)))),
         'tablet')
         then 1 else 0 end
         AS uses_tablet,

    -- [22]
        COUNT(DISTINCT (CASE WHEN lower(coalesce(gua_device_devicecategory,
        off_hits_devicetype)) = 'desktop'
          THEN  gua_visitnumber END))/(count(distinct gua_visitnumber)*1.00)
AS desktop_usage,

    --
    [23]
        COUNT(DISTINCT (CASE WHEN lower(coalesce(gua_device_devicecategory,
        off_hits_devicetype)) = 'mobile'
          THEN  gua_visitnumber END))/(count(distinct gua_visitnumber)*1.00)
AS mobile_usage,

    -- [24]
        COUNT(DISTINCT (CASE WHEN lower(coalesce(gua_device_devicecategory,
        off_hits_devicetype)) = 'tablet'
          THEN  gua_visitnumber END))/(count(distinct gua_visitnumber)*1.00)
AS tablet_usage,

    -- [25]
       array_distinct(array_agg(lower(gua_channelgrouping)))
       AS channel_array,

    -- [26]
       count(distinct gua_channelgrouping)
       AS cnt_channel,

    -- [27]
       count(CASE WHEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS
       DOUBLE) > 0 THEN 1 END)
AS cnt_earnings_events,

    -- [28]
       count(CASE WHEN cast(coalesce(off_hits_prices_cpc,'0') AS DOUBLE) > 0 THEN 1
       END)
AS cnt_spendings_events,

    -- [29]
       CAST(count(CASE WHEN source = 'onpage' THEN 1.00 END)/(COUNT(*)*1.00) AS
       DOUBLE)
AS total_ratio_touchpoint_onsite,

    -- [30]
       CAST(count(CASE WHEN source = 'offpage' THEN 1.00 END)/(COUNT(*)*1.00) AS
       DOUBLE)
AS total_ratio_touchpoint_offsite,
    -- [31]
       COUNT_IF(from_unixtime(cast(click_timestamp AS integer)) >
       date_add('day',-2,last_touch))
```

```
AS hits_1_2d,

    --
    [32]
       COUNT_IF(from_unixtime(cast(click_timestamp AS integer)) <
       date_add('day',-3,last_touch)
          AND from_unixtime(cast(click_timestamp AS integer)) >
          date_add('day',-4,last_touch))

AS hits_3_4d,

    --
    [33]
       COUNT_IF(from_unixtime(cast(click_timestamp AS integer)) <
       date_add('day',-5,last_touch)
          AND from_unixtime(cast(click_timestamp AS integer)) >
          date_add('day',-8,last_touch))

AS hits_5_8d,

    --
    [34]
       COUNT_IF(from_unixtime(cast(click_timestamp AS integer)) <
       date_add('day',-9,last_touch)
          AND from_unixtime(cast(click_timestamp AS integer)) >
          date_add('day',-16,last_touch))

AS hits_9_16d,

    -- [35]
       COUNT_IF(cast(gua_visitnumber AS integer) > (last_session_number-2))
       AS hits_1_2s,

    -- [36]
       COUNT_IF(cast(gua_visitnumber AS integer) < (last_session_number-3)
          AND cast(gua_visitnumber AS integer) > (last_session_number-4))
          AS hits_3_4s,

    -- [37]
       COUNT_IF(cast(gua_visitnumber AS integer) < (last_session_number-5)
          AND cast(gua_visitnumber AS integer) > (last_session_number-8))
          AS hits_5_8s,

    -- [38]
       COUNT_IF(cast(gua_visitnumber AS integer) < (last_session_number-9)
          AND cast(gua_visitnumber AS integer) > (last_session_number-16))
          AS hits_9_16s,

    -- [39]
       SUM(CASE WHEN (from_unixtime(cast(click_timestamp AS integer)) >
       date_add('day',-2,last_touch))
          THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00') AS DOUBLE)
          ELSE 0.00 END)

AS earnings_1_2d,

    --
    [40]
       SUM(CASE WHEN (from_unixtime(cast(click_timestamp AS integer)) <
       date_add('day',-3,last_touch)
          AND from_unixtime(cast(click_timestamp AS integer)) >
          date_add('day',-4,last_touch))
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00') AS
            DOUBLE) ELSE 0.00 END)

AS earnings_3_4d,

    --
    [41]
       SUM(CASE WHEN (from_unixtime(cast(click_timestamp AS integer)) <
       date_add('day',-5,last_touch)
```



```
          AND from_unixtime(cast(click_timestamp AS integer)) > date add('day',-8,last_touch))
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00') AS DOUBLE) ELSE 0.00 END)

AS earnings_5_8d,

    -- [42]
      SUM(CASE WHEN (from_unixtime(cast(click_timestamp AS integer)) < date add('day',-9,last_touch)
          AND from_unixtime(cast(click_timestamp AS integer)) > date add('day',-16,last_touch))
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00') AS DOUBLE) ELSE 0.00 END)

AS earnings_9_16d,

    -- [43]
      SUM(CASE WHEN cast(gua_visitnumber AS integer) > (last_session_number-2)
          THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00') AS DOUBLE)
          ELSE 0.00 END)

AS earnings_1_2s,

    -- [44]
      SUM(CASE WHEN (cast(gua_visitnumber AS integer) < (last_session_number-3)
          AND cast(gua_visitnumber AS integer) > ((last_session_number-4)))
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00') AS DOUBLE) ELSE 0.00 END)

AS earnings_3_4s,

    -- [45]
      SUM(CASE WHEN (cast(gua_visitnumber AS integer) < (last_session_number-5)
          AND cast(gua_visitnumber AS integer) > ((last_session_number-8)))
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00') AS DOUBLE) ELSE 0.00 END)

AS earnings_5_8s,

    -- [46]
      SUM(CASE WHEN (cast(gua_visitnumber AS integer) < (last_session_number-9)
          AND cast(gua_visitnumber AS integer) > ((last_session_number-16)))
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00') AS DOUBLE) ELSE 0.00 END)

AS earnings_9_16s,

    -- [47]
      SUM(CASE WHEN from_unixtime(cast(click_timestamp AS integer)) > date add('day',-2,last_touch)
        THEN cast(coalesce(off_hits_prices_cpc,'0.00') AS DOUBLE) ELSE 0.00 END)

AS spendings_1_2d,

    -- [48]
      SUM(CASE WHEN (from_unixtime(cast(click_timestamp AS integer)) < date add('day',-3,last_touch)
        AND from_unixtime(cast(click_timestamp AS integer)) > date add('day',-4,last_touch) )
        THEN cast(coalesce(off_hits_prices_cpc,'0.00') AS DOUBLE) ELSE 0.00 END)

AS spendings_3_4d,

    -- [49]
      SUM(CASE WHEN (from_unixtime(cast(click_timestamp AS integer)) < date add('day',-5,last_touch)
        AND from_unixtime(cast(click_timestamp AS integer)) >
```

```
           date_add('day',-8,last_touch) )
           THEN cast(coalesce(off_hits_prices_cpc,'0.00') AS DOUBLE) ELSE 0.00 END)

  AS spendings_5_8d,

      --
      [50]
         SUM(CASE WHEN (from_unixtime(cast(click_timestamp AS integer)) <
         date_add('day',-9,last_touch)
           AND from_unixtime(cast(click_timestamp AS integer)) >
           date_add('day',-16,last_touch) )
           THEN cast(coalesce(off_hits_prices_cpc,'0.00') AS DOUBLE) ELSE 0.00 END)

  AS spendings_9_16d,



      -- [51] 2 sessions customer_value (earnings_1_2s - spendings_1_2d
         SUM(CASE WHEN cast(gua_visitnumber AS integer) > (last_session_number-2)
          THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00') AS DOUBLE)
          ELSE 0.00 END) -
           SUM(CASE WHEN from_unixtime(cast(click_timestamp AS integer)) >
           date_add('day',-2,last_touch)
            THEN cast(coalesce(off_hits_prices_cpc,'0.00') AS DOUBLE) ELSE 0.00 END)
            AS customer_value_latest,

      -- [52]
         lower(last_device_category)
         AS last_device_category,

      -- [53]
         sum(CASE WHEN lower(on_hits_hit_product_name) = 'blickfang'
           THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE)
           ELSE 0 END )/
             sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
             AS product_percent_blickfang,

      -- [54]
         sum(CASE WHEN lower(on_hits_hit_product_name) = 'brokercontact'
           THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE)
           ELSE 0 END )/
             sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
             AS product_percent_brokercontact,

      -- [55]
         sum(CASE WHEN lower(on_hits_hit_product_name) = 'maklerempfehlung'
           THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE)
           ELSE 0 END )/
             sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
             AS product_percent_maklerempfehlung,

      -- [56]
         sum(CASE WHEN lower(on_hits_hit_product_name) = 'suchagent'
           THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE)
           ELSE 0 END )/
             sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
             AS product_percent_suchagent,

      -- [57]
         sum(CASE WHEN lower(on_hits_hit_product_name) = 'neubau-anfrage'
           THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE)
           ELSE 0 END )/
             sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
             AS product_percent_neubauanfrage,

      -- [58]
         sum(CASE WHEN lower(on_hits_hit_product_name) = 'phonecontact'
           THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE)
           ELSE 0 END )/
             sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
             AS product_percent_phonecontact,

        -- [59]
           sum(CASE WHEN lower(on_hits_hit_product_name) = 'contact'
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE) ELSE
            0 END )/
               sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
               AS product_percent_contact,

        -- [60]
           sum(CASE WHEN lower(on_hits_hit_product_name) = 'katalog-hausbau'
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE) ELSE
            0 END )/
               sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
               AS product_percent_kataloghausbau,

        -- [61]
           sum(CASE WHEN lower(on_hits_hit_product_name) = 'tir'
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE) ELSE
            0 END )/
               sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
               AS product_percent_tir,

        -- [62]
           sum(CASE WHEN lower(on_hits_hit_product_name) = 'gesuch-contact'
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE) ELSE
            0 END )/
               sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
               AS product_percent_gesuchcontact,

        -- [63]
           sum(CASE WHEN lower(on_hits_hit_product_name) = 'isa'
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE) ELSE
            0 END )/
               sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
               AS product_percent_isa,

        -- [64]
           sum(CASE WHEN lower(on_hits_hit_product_name) = 'call'
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE) ELSE
            0 END )/
               sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
               AS product_percent_call,

        -- [65]
           sum(CASE WHEN lower(on_hits_hit_product_name) = 'immobewertung'
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE) ELSE
            0 END )/
               sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
               AS product_percent_immobewertung,

        -- [66]
           sum(CASE WHEN lower(on_hits_hit_product_name) = 'pia'
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE) ELSE
            0 END )/
               sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
               AS product_percent_pia,

        -- [67]
           sum(CASE WHEN lower(on_hits_hit_product_name) = 'mailcontact'
            THEN cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE) ELSE
            0 END )/
               sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS DOUBLE))
               AS product_percent_mailcontact,

        -- [68]
           array_distinct(array_agg(off_hits_adgroupname))
           AS used_channels,

        -- [69]
           array_distinct(array_agg(concat(off_hits_campaignid, '-',
           off_hits_campaignid)))
```

```

AS used_channels_cleaned,

    -- [70]
       array_remove(array_distinct(array_agg(coalesce(off_hits_buymarketid,
       'REMOVE_ME') )), 'REMOVE_ME')

AS used_markets,

    -- [71] conversion propabiltiy amount = abs(customer_value_latest /
    customer_value)
       ABS((SUM(CASE WHEN cast(gua_visitnumber AS integer) > (last_session_number-2)
          THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00') AS DOUBLE)
          ELSE 0.00 END) -
            SUM(CASE WHEN from_unixtime(cast(click_timestamp AS integer)) >
            date add('day',-2,last_touch)
                THEN cast(coalesce(off_hits_prices_cpc,'0.00') AS DOUBLE) ELSE 0.00
                END) ) /
              (sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS
              DOUBLE)) -
                sum(cast(coalesce(off_hits_prices_cpc,'0') AS DOUBLE))))
                AS conversions_probability_amount,



    -- [72] CASE WHEN((customer_value_latest < 0 AND customer_value > 0)
    --         OR (customer_value_latest < 0 AND customer_value < 0))
    --         THEN conversion_probability_amount * -1
    --         ELSE conversion_probability_amount END
        CASE WHEN(
               --customer_value_latest < 0
               (SUM(CASE WHEN cast(gua_visitnumber AS integer) >
               (last_session_number-2)
                    THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00')
                    AS DOUBLE) ELSE 0.00 END) -
                    SUM(CASE WHEN from_unixtime(cast(click_timestamp AS integer)) >
                    date add('day',-2,last_touch)
                        THEN cast(coalesce(off_hits_prices_cpc,'0.00') AS DOUBLE)
                        ELSE 0.00 END))    < 0)
            THEN
              -- conversion_probability_amount * -1
              (ABS((SUM(CASE WHEN cast(gua_visitnumber AS integer) >
              (last_session_number-2)
                    THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00')
                    AS DOUBLE) ELSE 0.00 END) -
                SUM(CASE WHEN from_unixtime(cast(click_timestamp AS integer)) >
                date add('day',-2,last_touch)
                    THEN cast(coalesce(off_hits_prices_cpc,'0.00') AS DOUBLE) ELSE
                    0.00 END) ) /
                (sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS
                DOUBLE)) -
                    sum(cast(coalesce(off_hits_prices_cpc,'0') AS DOUBLE))))) * (-1)
            ELSE
              -- conversion_probability_amount
              ABS((SUM(CASE WHEN cast(gua_visitnumber AS integer) >
              (last_session_number-2)
                THEN cast(coalesce(on_hits_hit_product_conversion_price,'0.00') AS
                DOUBLE) ELSE 0.00 END) -
                SUM(CASE WHEN from_unixtime(cast(click_timestamp AS integer)) >
                date add('day',-2,last_touch)
                    THEN cast(coalesce(off_hits_prices_cpc,'0.00') AS DOUBLE) ELSE
                    0.00 END))   /
                (sum(cast(coalesce(on_hits_hit_product_conversion_price,'0') AS
                DOUBLE)) -
                    sum(cast(coalesce(off_hits_prices_cpc,'0') AS DOUBLE)))
                    END       AS conversions_probability


FROM   z,x
WHERE  coalesce(device_customer_idkey_map, device_visitor_id) = z.journey_id
AND    coalesce(device_customer_idkey_map, device_visitor_id) IS NOT NULL
```

```
GROUP
BY      coalesce(device_customer_idkey_map, device_visitor_id), last_touch, last_session_number,lower(last_device_category), row_num
ORDER
BY      row_num;

```

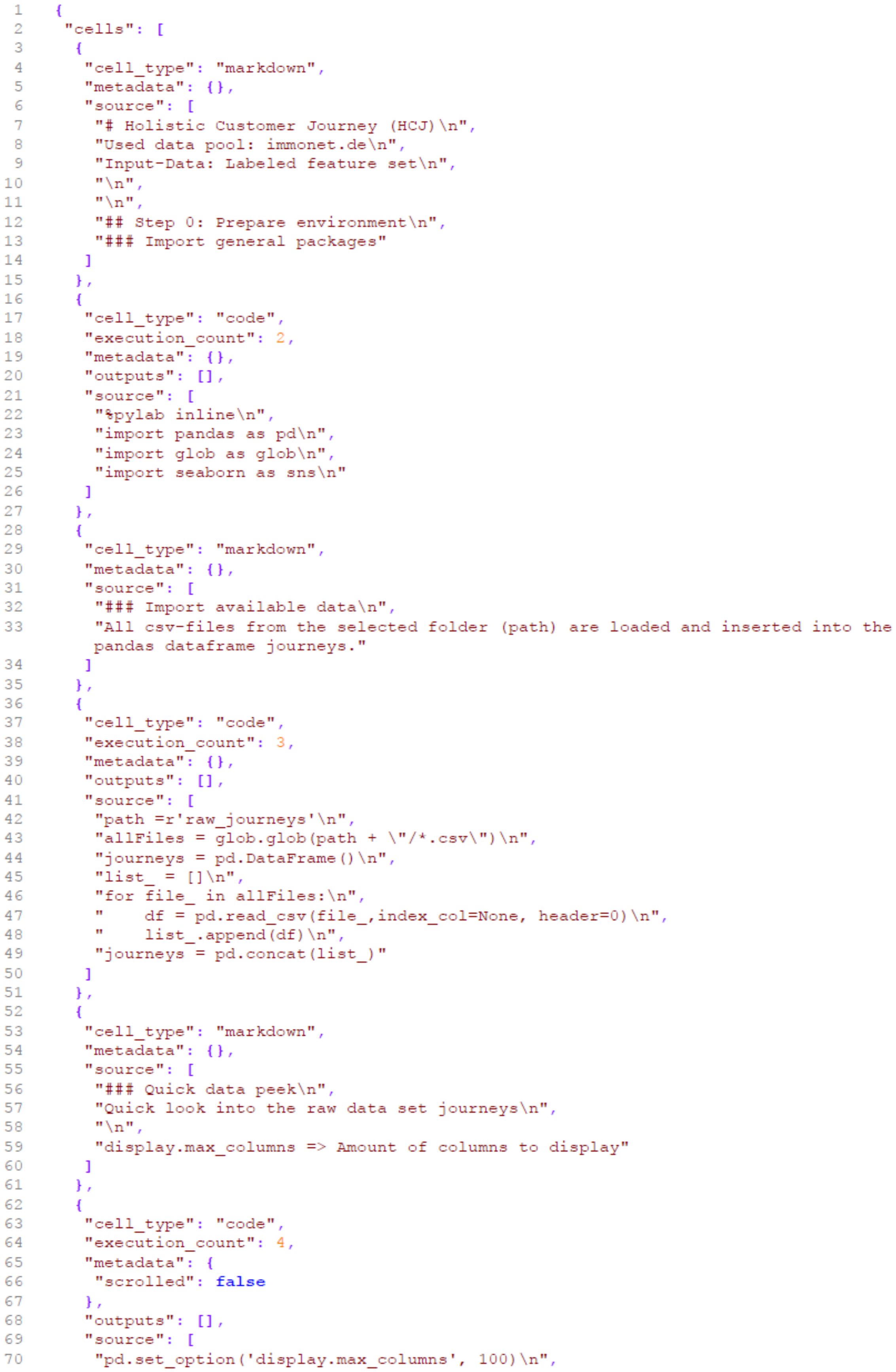
Appendix 33: Descriptive statistic values and distribution of features

feature name	mean	std	min	25%	50%	75%	max	percentage
total_earnings	25,54	78,781	0	3	9	22	44778	
total_spendings	0,083	0,317	0	0	0	0,082	81,509	
customer_value	25,456	78,684	-11,999	3	8,769	22	44778	
age_of_journey	11,366	21,644	0	0	0	11	91	
session_cnt	3,55	8,184	1	1	1	3	1115	
is_logged_in	0,063	0,243	0	0	0	0	1	
is_cross_device _user			0				1	6,46%
avg_events_per _session	5,629	6,804	1	2	4	7	1423	
total_hit_cnt	28,773	78,898	1	5	10	25	17156	
uses_desktop			0				1	48,51%
uses_mobile			0				1	44,96%
uses_tablet			0				1	13,13%
desktop_usage								41,95%
mobile_usage								45,34%
tablet_usage								12,71%
cnt_channel	1,275	0,571	1	1	1	1	8	
cnt_earnings _events	8,546	26,312	0	1	3	8	14926	
cnt_spendings _events	0,773	3,469	0	0	0	1	956	
total_ratio _touchpoint_onsite								81,58%
total_ratio _touchpoint_offsite								18,42%

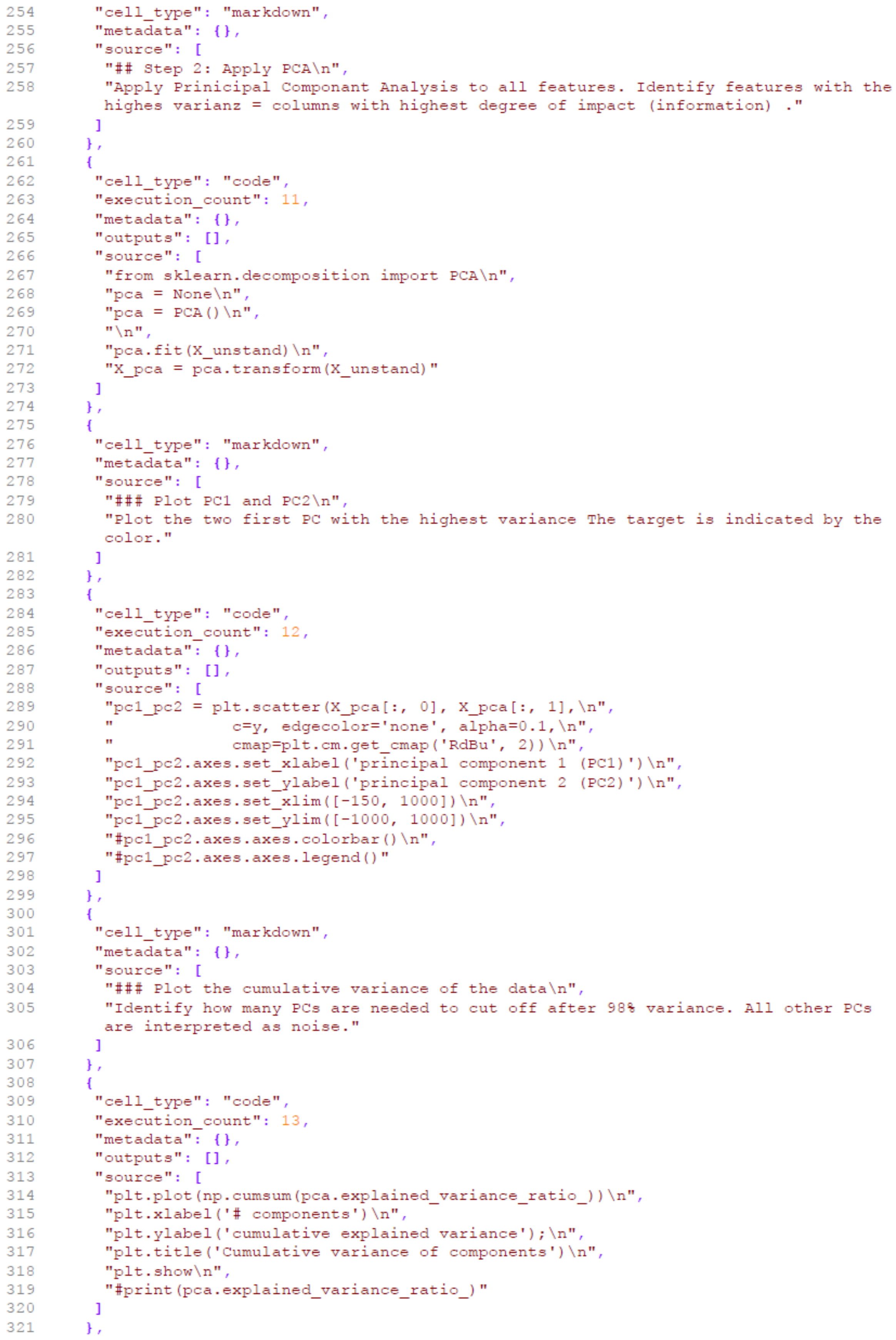
hits_5_8s	1,421	7,35118	0	0	0	0	1164	
hits_9_16s	2,015	12,55241	0	0	0	0	2193	
earnings_1_2d	10,068	20,60443	0	0	3	12	12066	
earnings_3_4d	0,445	4,975688	0	0	0	0	2817	
earnings_5_8d	1,215	10,14743	0	0	0	0	9618	
earnings_9_16d	2,211	16,48583	0	0	0	0	14700	
earnings_1_2s	10,486	18,13194	0	3	6	12	4329	
earnings_3_4s	0,444	4,97569	0	0	0	0	2817	
earnings_5_8s	1,492	8,75387	0	0	0	0	2814	
earnings_9_16s	2,099	14,61043	0	0	0	0	5028	
spendings_1_2d	0,038	0,09907	0	0	0	0	9,452	
spendings_3_4d	0,001	0,02098	0	0	0	0	5,248	
spendings_5_8d	0,004	0,03843	0	0	0	0	6,903	
spendings_9_16d	0,007	0,06051	0	0	0	0	12,660	
product_percent _blickfang	0,000	0,00022	0	0	0	0	0,25	
product_percent _brokercontact	0,000005	0,00548	0	0	0	0	1	
product_percent _maklerempfehlung	0,00001	0,00352	0	0	0	0	1	
product_percent _suchagent	0	0	0	0	0	0	1	
product_percent _neubauanfrage	0,00009	0,00678	0	0	0	0	0	
product_percent _phonecontact	0,00536	0,03275	0	0	1	1	1	
product_percent _contact	0,98207	0,07028	0	0	0	0	1	
product_percent _kataloghausbau	0,00004	0,00561	0	0	0	0	1	

product_percent_tir	0,00008	0,00233	0	0	0	0	1	
product_percent_gesuchcontact	0,00000	0,00172	0	0	0	0	1	
product_percent_isa	0	0	0	0	0	0	0	
product_percent_call	0,0008	0,01170	0	0	0	0	1	
product_percent_immobewertung	0,0000	0,00524	0	0	0	0	1	
product_percent_pia	0,0002	0,01268	0	0	0	0	1	
product_percent_mailcontact	0,0112	0,05337	0	0	0	0	1	

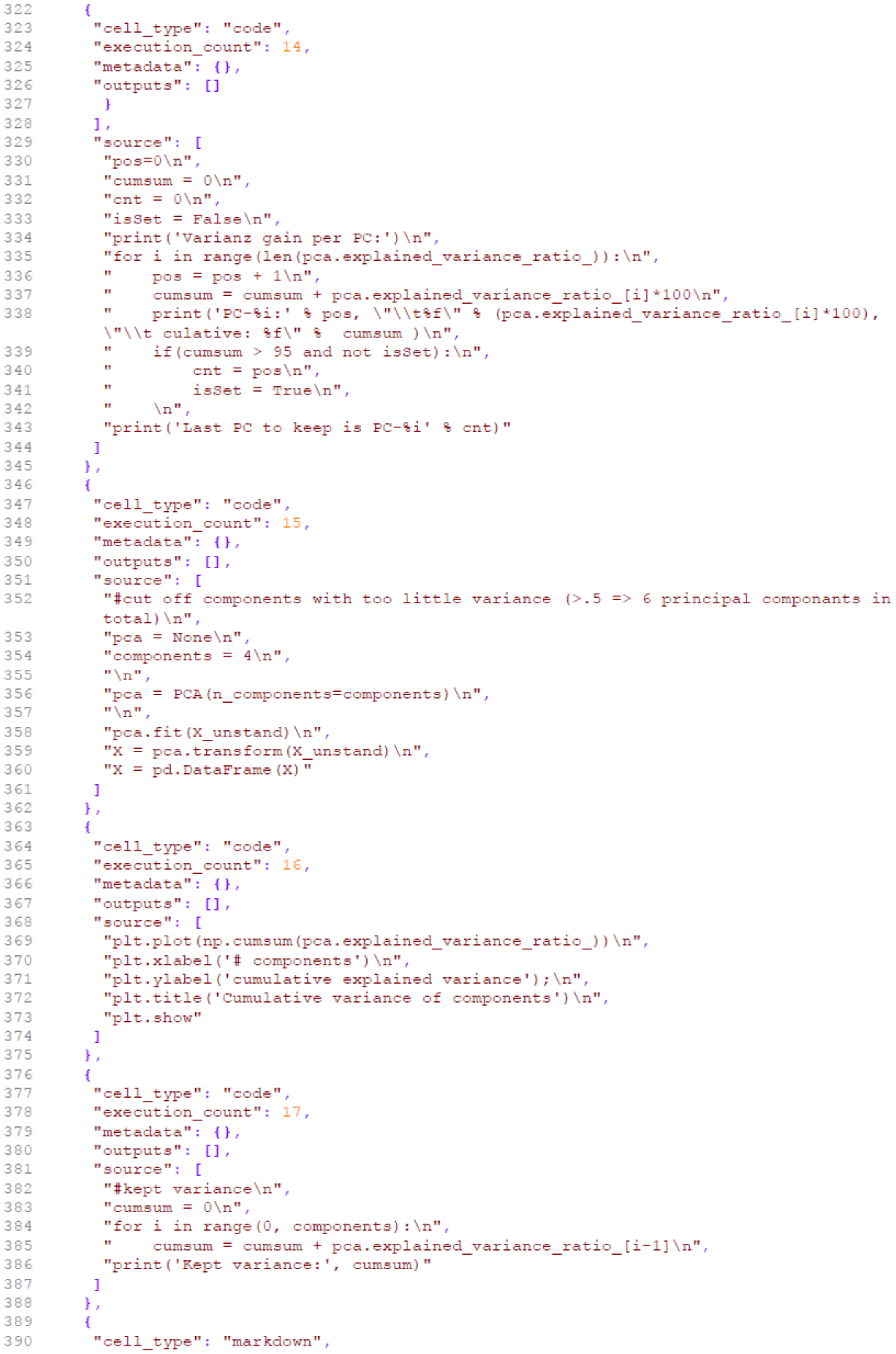
Appendix 34: Python: Jupyter notebook



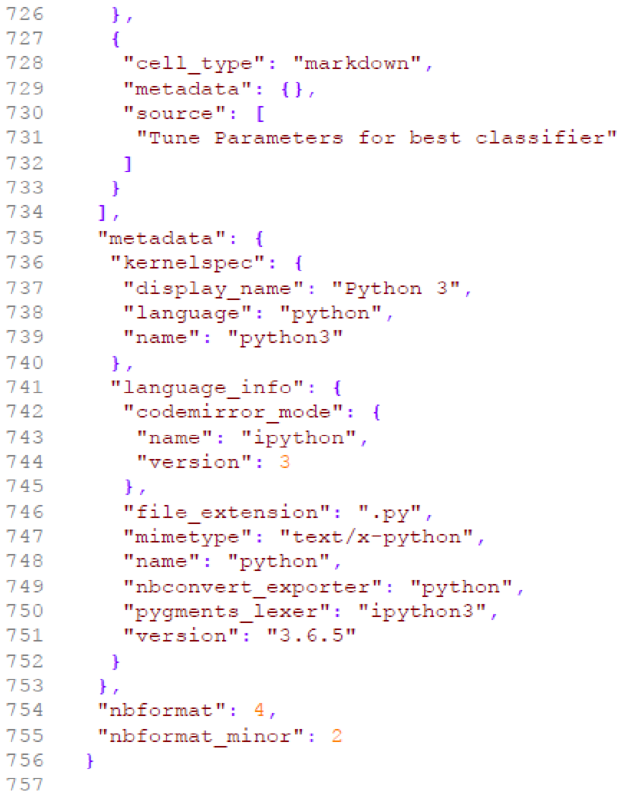
```
{
 "cells": [
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "# Holistic Customer Journey (HCJ)\n",
    "Used data pool: immonet.de\n",
    "Input-Data: Labeled feature set\n",
    "\n",
    "\n",
    "## Step 0: Prepare environment\n",
    "### Import general packages"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 2,
   "metadata": {},
   "outputs": [],
   "source": [
    "%pylab inline\n",
    "import pandas as pd\n",
    "import glob as glob\n",
    "import seaborn as sns\n"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "### Import available data\n",
    "All csv-files from the selected folder (path) are loaded and inserted into the pandas dataframe journeys."
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 3,
   "metadata": {},
   "outputs": [],
   "source": [
    "path =r'raw_journeys'\n",
    "allFiles = glob.glob(path + \"/*.csv\")\n",
    "journeys = pd.DataFrame()\n",
    "list_ = []\n",
    "for file_ in allFiles:\n",
    "    df = pd.read_csv(file_,index_col=None, header=0)\n",
    "    list_.append(df)\n",
    "journeys = pd.concat(list_)"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "### Quick data peek\n",
    "Quick look into the raw data set journeys\n",
    "\n",
    "display.max_columns => Amount of columns to display"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 4,
   "metadata": {
    "scrolled": false
   },
   "outputs": [],
   "source": [
    "pd.set_option('display.max_columns', 100)\n",
   "journeys.head()"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 5,
  "metadata": {},
  "outputs": [],
  "source": [
   "total_cnt = journeys.shape[0]\n",
   "total_cd_journeys = sum(journeys['is_cross_device_user'])\n",
   "ratio = total_cd_journeys / total_cnt\n",
   "\n",
   "print(\"Total count: \\t\\t\", total_cnt)\n",
   "print(\"Cross device journeys:\\t \", total_cd_journeys, \"\\tRatio:\\t\", ratio * 100, \"%\")\n"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "### Descriptive Statistics\n",
   "#### Dimensions of Data"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 6,
  "metadata": {
   "scrolled": true
  },
  "outputs": [],
  "source": [
   "journeys.shape"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "#### Datatype of attributes"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 7,
  "metadata": {
   "scrolled": true
  },
  "outputs": [],
  "source": [
   "journeys.dtypes"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "## Step 1: Standartize features\n",
   "* Select relevant features for further processing.\n",
   "* One-hot-encoding\n",
   "* Set the target\n",
   "\n",
   "mean = 0; varianz = 1"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 8,
  "metadata": {},
    "outputs": [],
    "source": [
     "#List with all available feature data\n",
     "all_features = list(journeys)\n",
     "\n",
     "#not relevant features\n",
     "irrelevant_features_list = {'row_num', 'journey_id', 'first_touch', 'last_touch', \n",
     "                            'customer_value_journey', 'overall_journey_cnt', \n",
     "                            'overall_avg_earning_per_journey', 'overall_avg_spendings_per_journey', \n",
     "                            'device_array', 'used_channels', 'used_channels_cleaned', \n",
     "                            'conversions_probability_amount', 'conversions_probability'}\n",
     "\n",
     "#get only relevant features all_features - irrelevant_features \n",
     "relevant_features = [e for e in all_features if e not in irrelevant_features_list]\n",
     "#print(relevant_features)\n",
     "\n",
     "# Get the features\n",
     "_X = journeys.loc[:, relevant_features]\n",
     "\n",
     "# \"One-hot encoding\" of features  'channel_array', 'last_device_category', 'used_markets'\n",
     "\n",
     "#Encoding for channel_array\n",
     "_X['organic search']       = _X['channel_array'].apply(lambda x: 1 if 'organic search' in x else 0)\n",
     "_X['display']              = _X['channel_array'].apply(lambda x: 1 if 'display' in x else 0)\n",
     "_X['referral']             = _X['channel_array'].apply(lambda x: 1 if 'referral' in x else 0)\n",
     "_X['direct']               = _X['channel_array'].apply(lambda x: 1 if 'direct' in x else 0)\n",
     "_X['paid search']          = _X['channel_array'].apply(lambda x: 1 if 'paid search' in x else 0)\n",
     "_X['social']               = _X['channel_array'].apply(lambda x: 1 if 'social' in x else 0)\n",
     "_X['other']                = _X['channel_array'].apply(lambda x: 1 if '(other)' in x else 0)\n",
     "_X['affiliates']           = _X['channel_array'].apply(lambda x: 1 if 'affiliates' in x else 0)\n",
     "_X['email']                = _X['channel_array'].apply(lambda x: 1 if 'email' in x else 0)\n",
     "# remove source column\n",
     "_X.drop(columns='channel_array', inplace=True)\n",
     "\n",
     "#Encoding for last_device_category\n",
     "_X['last_device_mobile']   = _X['last_device_category'].apply(lambda x: 1 if 'mobile' in x else 0)\n",
     "_X['last_device_tablet']   = _X['last_device_category'].apply(lambda x: 1 if 'tablet' in x else 0)\n",
     "_X['last_device_desktop']  = _X['last_device_category'].apply(lambda x: 1 if 'desktop' in x else 0)\n",
     "# remove source column\n",
     "_X.drop(columns='last_device_category', inplace=True)\n",
     "\n",
     "#Encoding for used_markets\n",
     "_X['buymarket_1']          = _X['used_markets'].apply(lambda x: 1 if '1' in x else 0)\n",
     "_X['buymarket_10']         = _X['used_markets'].apply(lambda x: 1 if '10' in x else 0)\n",
     "_X['buymarket_100']        = _X['used_markets'].apply(lambda x: 1 if '100' in x else 0)\n",
     "_X['buymarket_11']         = _X['used_markets'].apply(lambda x: 1 if '11' in x else 0)\n",
     "_X['buymarket_12']         = _X['used_markets'].apply(lambda x: 1 if '12' in x else 0)\n",
     "_X['buymarket_13']         = _X['used_markets'].apply(lambda x: 1 if '13' in x
    else 0)\n",
    "_X['buymarket_4']          = _X['used_markets'].apply(lambda x: 1 if '4' in x else 0)\n",
    "_X['buymarket_61']         = _X['used_markets'].apply(lambda x: 1 if '61' in x else 0)\n",
    "_X['buymarket_63']         = _X['used_markets'].apply(lambda x: 1 if '63' in x else 0)\n",
    "_X['buymarket_68']         = _X['used_markets'].apply(lambda x: 1 if '68' in x else 0)\n",
    "# remove source column\n",
    "_X.drop(columns='used_markets', inplace=True)\n",
    "\n",
    "#replace NaN values with 0 (those values occur because of a division by 0 in the prior step)\n",
    "_X.replace(np.nan, 0, inplace=True)\n",
    "X_unstand = _X\n",
    "X_unstand = pd.DataFrame(X_unstand)\n",
    "\n",
    "\n",
    "# Get the target\n",
    "\n",
    "target = ['conversions_probability'] \n",
    "_y = journeys.loc[:, target]\n",
    "y = _y['conversions_probability'].apply(lambda x: 1 if x>=0.5 else 0)\n"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "### Quick look into the raw features"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 9,
   "metadata": {},
   "outputs": [],
   "source": [
    "#print(X_unstand.head())\n",
    "total_y = y.shape[0]\n",
    "pos = sum(y)\n",
    "neg = total_y - pos\n",
    "\n",
    "print(\"Count of positive:\\t\", pos, '\\t', pos/total_y*100, '%')\n",
    "print('Count of negative:\\t', neg, '\\t', neg/total_y*100, '%')\n"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "### Distribution of the target\n",
    "\n",
    "Data selected from the intial journeys data frame"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 10,
   "metadata": {},
   "outputs": [],
   "source": [
    "plt_target = journeys['conversions_probability'].hist(bins=10000)\n",
    "plt_target.set_xlim([-1.5 , 1.5])\n",
    "plt_target.set_title('Occurence of conversion probability value')\n",
    "plt_target.set_xlabel('onversions probability')\n",
    "plt_target.set_ylabel('amount of journeys')\n"
   ]
  },
  {
```



```
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "## Step 2: Apply PCA\n",
    "Apply Prinicipal Componant Analysis to all features. Identify features with the highes varianz = columns with highest degree of impact (information) ."
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 11,
   "metadata": {},
   "outputs": [],
   "source": [
    "from sklearn.decomposition import PCA\n",
    "pca = None\n",
    "pca = PCA()\n",
    "\n",
    "pca.fit(X_unstand)\n",
    "X_pca = pca.transform(X_unstand)"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "### Plot PC1 and PC2\n",
    "Plot the two first PC with the highest variance The target is indicated by the color."
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 12,
   "metadata": {},
   "outputs": [],
   "source": [
    "pc1_pc2 = plt.scatter(X_pca[:, 0], X_pca[:, 1],\n",
    "            c=y, edgecolor='none', alpha=0.1,\n",
    "            cmap=plt.cm.get_cmap('RdBu', 2))\n",
    "pc1_pc2.axes.set_xlabel('principal component 1 (PC1)')\n",
    "pc1_pc2.axes.set_ylabel('principal component 2 (PC2)')\n",
    "pc1_pc2.axes.set_xlim([-150, 1000])\n",
    "pc1_pc2.axes.set_ylim([-1000, 1000])\n",
    "#pc1_pc2.axes.axes.colorbar()\n",
    "#pc1_pc2.axes.axes.legend()"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "### Plot the cumulative variance of the data\n",
    "Identify how many PCs are needed to cut off after 98% variance. All other PCs are interpreted as noise."
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 13,
   "metadata": {},
   "outputs": [],
   "source": [
    "plt.plot(np.cumsum(pca.explained_variance_ratio_))\n",
    "plt.xlabel('# components')\n",
    "plt.ylabel('cumulative explained variance');\n",
    "plt.title('Cumulative variance of components')\n",
    "plt.show\n",
    "#print(pca.explained_variance_ratio_)"
   ]
  },
```



```
  {
   "cell_type": "code",
   "execution_count": 14,
   "metadata": {},
   "outputs": []
    }
   ],
   "source": [
    "pos=0\n",
    "cumsum = 0\n",
    "cnt = 0\n",
    "isSet = False\n",
    "print('Varianz gain per PC:')\n",
    "for i in range(len(pca.explained_variance_ratio_)):\n",
    "    pos = pos + 1\n",
    "    cumsum = cumsum + pca.explained_variance_ratio_[i]*100\n",
    "    print('PC-%i:' % pos, \"\\t%f\" % (pca.explained_variance_ratio_[i]*100), \"\\t culative: %f\" %  cumsum )\n",
    "    if(cumsum > 95 and not isSet):\n",
    "        cnt = pos\n",
    "        isSet = True\n",
    "    \n",
    "print('Last PC to keep is PC-%i' % cnt)"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 15,
   "metadata": {},
   "outputs": [],
   "source": [
    "#cut off components with too little variance (>.5 => 6 principal componants in total)\n",
    "pca = None\n",
    "components = 4\n",
    "\n",
    "pca = PCA(n_components=components)\n",
    "\n",
    "pca.fit(X_unstand)\n",
    "X = pca.transform(X_unstand)\n",
    "X = pd.DataFrame(X)"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 16,
   "metadata": {},
   "outputs": [],
   "source": [
    "plt.plot(np.cumsum(pca.explained_variance_ratio_))\n",
    "plt.xlabel('# components')\n",
    "plt.ylabel('cumulative explained variance');\n",
    "plt.title('Cumulative variance of components')\n",
    "plt.show"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 17,
   "metadata": {},
   "outputs": [],
   "source": [
    "#kept variance\n",
    "cumsum = 0\n",
    "for i in range(0, components):\n",
    "    cumsum = cumsum + pca.explained_variance_ratio_[i-1]\n",
    "print('Kept variance:', cumsum)"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "## Step 3: Identify best ML algorithm"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": [
    "#reduce data\n",
    "X = X.loc[:99999,:]\n",
    "y = y[:100000]"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": [
    "import xgboost as xgb\n"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 18,
   "metadata": {},
   "outputs": [],
   "source": [
    "from sklearn.model_selection import train_test_split\n",
    "from sklearn.metrics import make_scorer, f1_score\n",
    "from sklearn.metrics import classification_report\n",
    "from datetime import datetime\n",
    "\n",
    "# Import classifier\n",
    "from sklearn.ensemble import RandomForestClassifier\n",
    "# n_estimators\n",
    "from sklearn.ensemble import ExtraTreesClassifier\n",
    "# n_estimators, complexity of learners, e.g. through max_depth\n",
    "from sklearn.ensemble import AdaBoostClassifier\n",
    "# n_estimators, complexity of learners, e.g. through max_depth\n",
    "from sklearn.ensemble import GradientBoostingClassifier\n",
    "# n_estimators, complexity of learners, e.g. through max_depth\n",
    "#from xgboost import XGBClassifier\n",
    "\n",
    "\n",
    "\n",
    "# make scorer\n",
    "# prec_scorer   = make_scorer(precision_score, pos_label = 1)
    #greater_is_better = True by default\n",
    "# recall_scorer = make_scorer(recall_score, pos_label
    ='conversion_probability')    #greater_is_better = True by default\n",
    "f1_scorer      = make_scorer(f1_score, pos_label = 1) #greater_is_better = True
    by default, pos_label = 1 (positive: invest!)"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 19,
   "metadata": {},
   "outputs": [],
   "source": [
    "# dictionary of classifiers to test\n",
    "all_classifier = {\n",
    "   # '_xgBoost': XGBClassifier,\n",
    "    '_randomForest': RandomForestClassifier,\n",
    "    '_extraTree': ExtraTreesClassifier,  \n",
    "    '_AdaBoost': AdaBoostClassifier,\n",
    "    '_GradientBoost': GradientBoostingClassifier \n",
    "}"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": [
    "# n_estimators: from 10 to 250 in steps of 10 (#amount of trees)\n",
    "param = np.arange(10,250,90)\n",
    "\n",
    "# tables to store results of training\n",
    "f1 = {}\n",
    "\n",
    "# build \"empty\" object with 0 values for each classifier\n",
    "for key in all_classifier:\n",
    "    f1[key]            = np.zeros(len(param))"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 20,
   "metadata": {},
   "outputs": [],
   "source": [
    "from sklearn.model_selection import cross_val_score\n",
    "\n",
    "# split all data randomly in two sections: train (train and validation) and test (overall test) set. \n",
    "# setting random_state ensures the same splitting each call\n",
    "X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)\n"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": [
    "f1['_randomForest'] = np.zeros(len(param))\n",
    "f1['_randomForest'][0] = mean([0.98560824, 0.98558985])\n",
    "f1['_randomForest'][1] = mean([0.98664676, 0.98654891])\n",
    "f1['_randomForest'][2] = mean([0.98671273, 0.98657042])\n",
    "f1['_extraTree'] = np.zeros(len(param))\n",
    "f1['_extraTree'][0] = mean([0.98545404, 0.9852567 ])\n",
    "f1['_extraTree'][1] = mean([0.98657799, 0.98649169])\n",
    "f1['_extraTree'][2] = mean([0.9866782,  0.98655321])"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": [
    "print(f1)"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": [
    "import datetime\n",
    "_cur = []\n",
    "\n",
    "# do training\n",
    "# over all classifier\n",
   "for clf_key in all_classifier:\n",
   "    \n",
   "    #current classifier\n",
   "    print(clf_key)\n",
   "    #current time\n",
   "    print(datetime.datetime.now().time())\n",
   "    # over all configuration parameter   \n",
   "    for param_index in range(len(param)):\n",
   "        \n",
   "            \n",
   "        # initialize classifier with parameter\n",
   "        clf = all_classifier[clf_key](n_estimators = param[param_index]) #, n_jobs = -1)\n",
   " \n",
   "        # train model, store f1-score from average of cross-validation runs\n",
   "        _cur = cross_val_score(clf, X_train, y_train, cv=2, scoring=f1_scorer)\n",
   "        print(_cur)\n",
   "        f1[clf_key][param_index] = mean(_cur)\n",
   "                   \n",
   "print(datetime.datetime.now().time())"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "Identify best performing classifier"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": null,
  "metadata": {},
  "outputs": [],
  "source": [
   "for key, dat in f1.items():\n",
   "    plt.plot(param,dat)\n",
   "    plt.xlabel('n_estimator: Amount of Trees')\n",
   "    plt.ylabel('mean of f1 scores')\n",
   "\n",
   "plt.legend(framealpha=0.5, labels=[ \n",
   "                                   'RandomForest',\n",
   "                                   'ExtraTree',\n",
   "                                   'AdaBoost',\n",
   "                                   'GradientBoost'\n",
   "                                   ])\n",
   "    #legende bauen matplotlib.legend\n",
   "print('Size of trainngs set:', len(X_train))\n",
   "plt.show()"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": null,
  "metadata": {},
  "outputs": [],
  "source": [
   "X_train.shape()"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "## Step 4: Identify best model configuration"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 26,
"metadata": {
 "scrolled": false
},
"outputs": [],
"source": [
 "from sklearn.metrics import accuracy_score\n",
 "import datetime\n",
 "\n",
 "# training data : X_train, y_train\n",
 "# overall test data : X_test, y_test\n",
 "\n",
 "# split train data into training and validation set.\n",
 "X_train_train, X_train_test, y_train_train, y_train_test =
 train_test_split(X_train, y_train, test_size=0.2,\n",
 "
 random_state=42)\n",
 "min_samples_split = [1.0, 2, 4, 8, 10]\n",
 "max_depth = [10, 12, 14, 16, 20, 25, 50]\n",
 "n_estimators = np.arange(10,331,40)\n",
 "\n",
 "\n",
 "result_set_length = len(min_samples_split) * len(max_depth) *
 len(n_estimators)\n",
 "\n",
 "results = {}\n",
 "results['config'] = [''] * result_set_length\n",
 "results['min_samples_split'] = np.zeros(result_set_length) \n",
 "results['max_depth'] = np.zeros(result_set_length)\n",
 "results['n_estimators'] = np.zeros(result_set_length) \n",
 "results['train_error'] = np.zeros(result_set_length)    \n",
 "results['test_error'] = np.zeros(result_set_length)\n",
 "results['whole_run_error'] = np.zeros(result_set_length)\n",
 "\n",
 "\n",
 "i = -1\n",
 "for cur_min_samples_split in min_samples_split:\n",
 "    for cur_max_depth in max_depth:\n",
 "        for cur_n_estimators in n_estimators:\n",
 "            i = i + 1\n",
 "            cur_results = {}\n",
 "            #current time\n",
 "            print(datetime.datetime.now().time())\n",
 "            print('Current hyperparameters: min_sample_split:',
 cur_min_samples_split, \n",
 "                  ' max_depth:', cur_max_depth, ' n_estimators:',
 cur_n_estimators)           \n",
 "            results['config'][i] =   \"min_samples_split: {0},  max_depth: {1},
 n_estimators: {2}\".format(cur_min_samples_split, cur_max_depth,
 cur_n_estimators)           \n",
 "            results['min_samples_split'][i] = cur_min_samples_split\n",
 "            results['max_depth'][i] = cur_max_depth\n",
 "            results['n_estimators'][i] = cur_n_estimators\n",
 "            \n",
 "            #building the classifier\n",
 "            #eval_metric is set to error by default for a classification\n",
 "            xg_clf = all_classifier['_randomForest'](\n",
 "                               n_jobs = -1, \n",
 "                               #objective = 'binary:logistic', \n",
 "                               min_samples_split = cur_min_samples_split, \n",
 "                               max_depth = cur_max_depth,\n",
 "                               n_estimators = cur_n_estimators\n",
 "                                )\n",
 "            #Train the model with training data\n",
 "            xg_clf.fit(X_train_train, y_train_train)\n",
 "             \n",
 "            #predict on the trainings data set. How well can the model predict
 the trainng data (pattern learning)\n",
 "            train_train_pred = xg_clf.predict(X_train_train)  \n",
 "            #compare the prediction results with the \"real\" results        \n",
 "            results['train_error'][i] = 1 - accuracy_score(y_train_train,
 train_train_pred)  \n",
    "            \n",
    "       \n",
    "            #apply model to the validation data (X_train_test)\n",
    "            train_test_pred = xg_clf.predict(X_train_test)\n",
    "            #compare the prediction results with the \"real\" results \n",
    "            results['test_error'][i] = 1 - accuracy_score(y_train_test, train_test_pred)\n",
    "            \n",
    "            \n",
    "            results['whole_run_error'][i] = (results['train_error'][i] + results['test_error'][i]) / 2\n",
    "            print('Error on test set: ', results['test_error'][i])\n",
    "            print()\n",
    "           "
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": [
    "# get index of min error from test set. \n",
    "index = results['test_error'].argmin()\n",
    "# print best configuration with minimal over- / undefitting\n",
    "print(\"Best model should use: \", results['config'][index])\n",
    "print()\n",
    "print()\n",
    "print(results)\n"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "## Step 5: Create optimal model for prediction\n",
    "\n",
    "Create the final model with best classifier and best hyperparameters"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 27,
   "metadata": {},
   "outputs": [],
   "source": [
    "# Best configuration: Best model should use:  min_smples_split: 8,  max_depth: 20, n_estimators: 290\n",
    "\n",
    "xg_clf = all_classifier['_randomForest'](n_jobs = -1, \n",
    "                               min_samples_split = 8,  \n",
    "                               max_depth = 20,\n",
    "                               n_estimators = 290)\n",
    "            \n",
    "# train the model with all available training data\n",
    "xg_clf.fit(X_train, y_train)\n",
    "             \n",
    "#predict on the test data set.\n",
    "test_train_pred = xg_clf.predict(X_test)\n",
    "\n",
    "\n",
    "#apply model to the validation data (X_train_test)\n",
    "train_test_pred = xg_clf.predict(X_train_test)\n",
    "#compare the prediction results with the \"real\" results \n",
    "performance = accuracy_score(y_test, test_train_pred)\n",
    "error_in_reality = 1 - performance\n",
    "\n",
    "print('Prediction performance of the model on test set:', performance, ' Error:', error_in_reality)\n",
    "            "
   ]
```



```
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "Tune Parameters for best classifier"
   ]
  }
 ],
 "metadata": {
  "kernelspec": {
   "display_name": "Python 3",
   "language": "python",
   "name": "python3"
  },
  "language_info": {
   "codemirror_mode": {
    "name": "ipython",
    "version": 3
   },
   "file_extension": ".py",
   "mimetype": "text/x-python",
   "name": "python",
   "nbconvert_exporter": "python",
   "pygments_lexer": "ipython3",
   "version": "3.6.5"
  }
 },
 "nbformat": 4,
 "nbformat_minor": 2
}
```

Appendix 35: Performance of different hyperparameter combinations

The optimal configuration (smallest error on test set) is highlighted in bold.

Start	min_samples_split	max_depth	n_estimators	error on test set
09:29:39.851638	1	10	10	0,39182681371816200
09:29:50.766571	1	10	50	0,39182681371816200
09:30:01.232893	1	10	90	0,39182681371816200
09:30:16.924951	1	10	130	0,39182681371816200
09:30:38.496898	1	10	170	0,39182681371816200
09:31:05.367370	1	10	210	0,39182681371816200
09:31:37.829993	1	10	250	0,39182681371816200
09:32:16.244237	1	10	290	0,39182681371816200
09:33:00.963100	1	10	330	0,39182681371816200
09:33:51.261729	1	12	10	0,39182681371816200
09:33:54.663091	1	12	50	0,39182681371816200
09:34:04.565907	1	12	90	0,39182681371816200
09:34:20.381732	1	12	130	0,39182681371816200
09:34:42.012561	1	12	170	0,39182681371816200
09:35:09.005466	1	12	210	0,39182681371816200
09:35:41.885559	1	12	250	0,39182681371816200
09:36:20.873519	1	12	290	0,39182681371816200
09:37:04.992143	1	12	330	0,39182681371816200
09:37:54.856198	1	14	10	0,39182681371816200
09:37:58.331823	1	14	50	0,39182681371816200
09:38:08.181756	1	14	90	0,39182681371816200
09:38:23.788867	1	14	130	0,39182681371816200
09:38:44.991038	1	14	170	0,39182681371816200
09:39:11.902381	1	14	210	0,39182681371816200
09:39:44.753787	1	14	250	0,39182681371816200
09:40:23.385436	1	14	290	0,39182681371816200
09:41:07.989313	1	14	330	0,39182681371816200
09:41:57.778330	1	16	10	0,39182681371816200
09:42:01.233380	1	16	50	0,39182681371816200
09:42:10.928848	1	16	90	0,39182681371816200
09:42:26.388296	1	16	130	0,39182681371816200
09:42:47.605895	1	16	170	0,39182681371816200
09:43:14.620442	1	16	210	0,39182681371816200
09:43:47.460736	1	16	250	0,39182681371816200
09:44:26.022008	1	16	290	0,39182681371816200
09:45:10.357026	1	16	330	0,39182681371816200
09:48:52.149842	1	20	10	0,39182681371816200
09:48:56.035277	1	20	50	0,39182681371816200
09:49:05.730287	1	20	90	0,39182681371816200
09:49:21.151738	1	20	130	0,39182681371816200

09:49:42.408073	1	20	170	0,39182681371816200
09:50:09.581253	1	20	210	0,39182681371816200
09:50:42.439845	1	20	250	0,39182681371816200
09:51:21.104181	1	20	290	0,39182681371816200
09:52:05.515395	1	20	330	0,39182681371816200
09:52:55.318793	1	25	10	0,39182681371816200
09:52:58.838207	1	25	50	0,39182681371816200
09:53:08.552638	1	25	90	0,39182681371816200
09:53:24.077837	1	25	130	0,39182681371816200
09:53:45.184120	1	25	170	0,39182681371816200
09:54:12.433363	1	25	210	0,39182681371816200
09:54:45.234617	1	25	250	0,39182681371816200
09:55:23.577251	1	25	290	0,39182681371816200
09:56:07.683005	1	25	330	0,39182681371816200
09:56:57.691242	1	50	10	0,39182681371816200
09:57:01.138436	1	50	50	0,39182681371816200
09:57:10.859879	1	50	90	0,39182681371816200
09:57:26.356596	1	50	130	0,39182681371816200
09:57:47.709425	1	50	170	0,39182681371816200
09:58:14.609016	1	50	210	0,39182681371816200
09:58:46.877784	1	50	250	0,39182681371816200
09:59:25.690098	1	50	290	0,39182681371816200
10:00:10.025160	1	50	330	0,39182681371816200
17:07:18.237835	2	10	10	0,02459791863765370
17:07:46.350413	2	10	50	0,02349122583867720
17:08:41.108508	2	10	90	0,02353300534870570
17:10:09.267994	2	10	130	0,02374375976596040
17:12:05.270274	2	10	170	0,02310128374507810
17:14:31.769818	2	10	210	0,02321826637315780
17:17:29.429501	2	10	250	0,02361377906809410
17:20:55.428157	2	10	290	0,02323126444294450
17:24:50.998356	2	10	330	0,02323497817716920
17:29:12.024050	2	12	10	0,01959551863691100
17:29:41.135915	2	12	50	0,01909787825079400
17:30:46.154392	2	12	90	0,01933741410829050
17:32:24.120211	2	12	130	0,01912201752325490
17:34:36.683912	2	12	170	0,01918607943863190
17:37:21.638497	2	12	210	0,01911737535547390
17:40:41.983715	2	12	250	0,01915636956483390
17:44:35.493969	2	12	290	0,01916286859972720
17:49:02.662891	2	12	330	0,01916379703328340
17:54:01.157811	2	14	10	0,01777578886678180
17:54:33.429159	2	14	50	0,01739513110874460

17:55:43.672195	2	14	90	0,01732828389269900
17:57:30.237693	2	14	130	0,01724472487264210
17:59:54.915982	2	14	170	0,01722058560018120
18:02:54.781894	2	14	210	0,01721408656528780
18:06:32.191645	2	14	250	0,01714538248212990
18:10:52.604948	2	14	290	0,01724472487264210
18:15:46.245252	2	14	330	0,01711010200699480
18:21:13.164506	2	16	10	0,01676379629053650
18:21:47.477453	2	16	50	0,01641377683985350
18:23:01.732391	2	16	90	0,01628843830976810
18:24:55.496699	2	16	130	0,01626801277153200
18:27:29.845943	2	16	170	0,01625780000241390
18:30:45.398912	2	16	210	0,01628658144265570
18:34:39.118176	2	16	250	0,01628658144265570
18:39:13.531724	2	16	290	0,01622716169505960
18:44:22.650045	2	16	330	0,01628658144265570
10:00:59.996637	2	20	10	0,01660317728531600
10:01:36.458407	2	20	50	0,01594306102686600
10:02:56.727481	2	20	90	0,01581772249678070
10:04:59.761814	2	20	130	0,01581865093033680
10:07:47.312799	2	20	170	0,01578337045520170
10:11:16.305908	2	20	210	0,01580286755988160
10:15:27.688432	2	20	250	0,01579358322431970
10:20:22.791492	2	20	290	0,01580936659477490
10:25:58.135831	2	20	330	0,01574716154651030
10:32:18.097066	2	25	10	0,01703489888894360
10:32:54.775124	2	25	50	0,01599133957178780
10:34:16.943015	2	25	90	0,01589663934905670
10:36:19.926314	2	25	130	0,01585114610480340
10:39:07.754903	2	25	170	0,01576573021763400
10:42:40.951771	2	25	210	0,01574901841362270
10:46:56.835826	2	25	250	0,01581493719611200
10:51:52.521458	2	25	290	0,01581122346188730
10:57:31.915334	2	25	330	0,01582050779744920
11:03:51.198664	2	50	10	0,01709710393720820
11:04:28.899299	2	50	50	0,01600340920801830
11:05:49.808961	2	50	90	0,01595048849531550
11:07:54.384152	2	50	130	0,01589942464972520
11:10:40.212572	2	50	170	0,01592727765641080
11:14:05.426226	2	50	210	0,01591335115306800
11:18:19.919042	2	50	250	0,01587528537726430
11:23:13.549421	2	50	290	0,01586878634237090
18:50:15.446595	4	10	10	0,02416155486624520

18:50:42.126408	4	10	50	0,02361192220098170
18:51:38.809827	4	10	90	0,02284132234934540
18:53:05.007884	4	10	130	0,02318391433157890
18:55:02.475223	4	10	170	0,02359521039697030
18:57:27.156334	4	10	210	0,02331110972877670
19:00:23.594267	4	10	250	0,02348194150311530
19:03:52.150921	4	10	290	0,02316070349267410
19:07:46.265731	4	10	330	0,02346151596487920
19:12:14.201041	4	12	10	0,01933091507339730
19:12:43.536214	4	12	50	0,01915451269772150
19:13:50.454314	4	12	90	0,01917029606817670
19:15:27.356922	4	12	130	0,01918515100507570
19:17:39.501781	4	12	170	0,01915915486550240
19:20:25.303438	4	12	210	0,01911459005480540
19:23:44.084420	4	12	250	0,01913965776082240
19:27:34.603981	4	12	290	0,01904031537031030
19:32:03.414533	4	12	330	0,01909323608301300
19:37:06.599713	4	14	10	0,01776928983188850
19:37:38.413928	4	14	50	0,01715095308346710
19:38:48.948355	4	14	90	0,01716487958680990
19:40:39.030393	4	14	130	0,01716302271969760
19:43:05.748468	4	14	170	0,01716952175459080
19:46:10.210126	4	14	210	0,01722058560018120
19:49:52.750200	4	14	250	0,01716673645392230
19:54:08.432237	4	14	290	0,01717880609015270
19:59:00.894866	4	14	330	0,01721594343240020
20:04:33.532227	4	16	10	0,01675079822074990
20:05:08.925368	4	16	50	0,01637756793116210
20:06:24.022500	4	16	90	0,01636921202915650
20:08:16.536320	4	16	130	0,01633021781979660
20:10:54.102003	4	16	170	0,01628565300909950
20:14:06.299999	4	16	210	0,01626151373663860
20:17:58.992927	4	16	250	0,01631072071511660
20:22:30.877422	4	16	290	0,01627079807220050
20:27:41.492233	4	16	330	0,01628193927487480
11:35:15.784535	4	20	10	0,01655675560750650
11:35:53.581746	4	20	50	0,01587621381082040
11:37:12.008633	4	20	90	0,01574623311295420
11:39:15.257858	4	20	130	0,01581865093033680
11:41:59.873757	4	20	170	0,01577130081897120
11:45:27.672397	4	20	210	0,01579451165787590
11:49:37.245572	4	20	250	0,01578151358808930
11:54:32.108698	4	20	290	0,01574716154651030

12:00:06.391171	4	20	330	0,01576201648340930
12:06:20.231366	4	25	10	0,01673687171740710
12:06:59.830974	4	25	50	0,01591056585239940
12:08:19.826586	4	25	90	0,01586135887392140
12:10:25.556485	4	25	130	0,01579544009143210
12:13:12.064436	4	25	170	0,01580843816121870
12:16:44.835202	4	25	210	0,01575644588207220
12:20:58.698669	4	25	250	0,01571930853982470
12:25:53.383245	4	25	290	0,01577965672097690
12:31:31.061517	4	25	330	0,01575830274918450
12:37:50.432482	4	50	10	0,01693277119776280
12:38:27.909963	4	50	50	0,01598948270467540
12:39:46.808888	4	50	90	0,01591706488729280
12:41:49.850300	4	50	130	0,01586785790881480
12:44:37.298520	4	50	170	0,01582793526589870
12:48:10.156539	4	50	210	0,01578894105653880
12:52:27.743891	4	50	250	0,01582979213301100
12:57:22.757777	4	50	290	0,01580472442699390
13:03:01.620475	4	50	330	0,01580658129410630
20:33:30.492213	8	10	10	0,02388673853361340
20:33:56.026884	8	10	50	0,02355250245338560
20:34:54.573206	8	10	90	0,02328697045631570
20:36:24.586024	8	10	130	0,02309571314374100
20:38:21.863258	8	10	170	0,02365555857812260
20:40:46.561369	8	10	210	0,02338445597971560
20:43:42.006087	8	10	250	0,02348008463600290
20:47:10.334949	8	10	290	0,02359799569763880
20:51:06.552910	8	10	330	0,02342530705618780
20:55:33.690663	8	12	10	0,01983691136151990
20:56:02.316729	8	12	50	0,01932255917139150
20:57:06.779947	8	12	90	0,01945718203703890
20:58:46.509470	8	12	130	0,01919629220774990
21:00:56.770294	8	12	170	0,01918422257151950
21:03:42.038921	8	12	210	0,01910994788702440
21:07:01.736389	8	12	250	0,01910066355146250
21:10:54.371639	8	12	290	0,01920464810975570
21:15:27.618114	8	12	330	0,01917958040373850
21:20:27.877201	8	14	10	0,01782313897814740
21:21:00.811743	8	14	50	0,01729021811689530
21:22:11.787321	8	14	90	0,01733292606047990
21:23:59.346987	8	14	130	0,01724193957197350
21:26:23.514122	8	14	170	0,01724936704042300
21:29:26.198781	8	14	210	0,01719923162838880

21:33:07.352432	8	14	250	0,01716673645392230
21:37:21.849491	8	14	290	0,01721222969817540
21:42:16.704382	8	14	330	0,01719923162838880
21:47:46.390314	8	16	10	0,01685292591193060
21:48:21.181472	8	16	50	0,01635528552581360
21:49:37.767759	8	16	90	0,01636364142781940
21:51:32.688038	8	16	130	0,01634228745602700
21:54:08.702588	8	16	170	0,01626986963864440
21:57:25.787050	8	16	210	0,01629029517688050
22:01:20.700476	8	16	250	0,01629772264533000
22:05:52.404141	8	16	290	0,01629122361043660
22:11:08.976179	8	16	330	0,01628565300909950
13:48:58.886995	8	20	10	0,01646576911900010
13:49:44.150587	8	20	50	0,01590963741884330
13:51:03.626846	8	20	90	0,01585578827258430
13:53:06.624567	8	20	130	0,01583164900012340
13:55:53.086351	8	20	170	0,01579079792365120
13:59:19.515402	8	20	210	0,01572302227404940
14:03:30.982670	8	20	250	0,01579079792365120
14:08:26.194997	**8**	**20**	**290**	**0,01561280950493130**
14:14:03.488661	8	20	330	0,01580286755988160
14:20:22.581982	8	25	10	0,01655582717395040
14:21:01.577095	8	25	50	0,01591613645373650
14:22:20.672312	8	25	90	0,01580379599343780
14:24:28.651236	8	25	130	0,01577037238541500
14:27:14.640434	8	25	170	0,01582886369945490
14:30:44.377173	8	25	210	0,01572859287538650
14:34:57.599998	8	25	250	0,01572209384049320
14:39:57.488549	8	25	290	0,01572487914116180
14:45:33.327270	8	25	330	0,01571745167271230
14:51:56.316465	8	50	10	0,01657253897796170
14:52:33.932619	8	50	50	0,01591613645373650
14:53:54.002201	8	50	90	0,01582886369945490
14:55:57.663020	8	50	130	0,01578708418942640
14:58:46.456620	8	50	170	0,01574623311295420
15:02:18.329605	8	50	210	0,01575366058140370
15:06:31.427371	8	50	250	0,01573973407806080
15:11:26.456033	8	50	290	0,01570816733715040
15:17:09.323731	8	50	330	0,01571652323915610
22:17:05.683968	10	10	10	0,02400650646236180
22:17:32.677876	10	10	50	0,02324240564561870
22:18:30.428720	10	10	90	0,02307157387128010
22:19:55.647436	10	10	130	0,02348936897156480

22:21:52.660323	10	10	170	0,02336867260926030
22:24:20.164342	10	10	210	0,02343459139174970
22:27:18.614046	10	10	250	0,02312356615042660
22:30:43.910721	10	10	290	0,02341323741995730
22:34:37.233339	10	10	330	0,02340023935017070
22:39:02.141094	10	12	10	0,02000217253452150
22:39:31.720975	10	12	50	0,01913780089371000
22:40:36.831278	10	12	90	0,01910437728568730
22:42:14.918316	10	12	130	0,01903381633541700
22:44:26.337334	10	12	170	0,01913965776082240
22:47:12.326520	10	12	210	0,01908209488033880
22:50:30.423770	10	12	250	0,01915822643194630
22:54:23.026453	10	12	290	0,01918422257151950
22:58:47.907590	10	12	330	0,01914244306149100
23:03:48.472129	10	14	10	0,01775722019565800
23:04:20.160085	10	14	50	0,01729300341756390
23:05:31.840771	10	14	90	0,01725865137598490
23:07:21.273418	10	14	130	0,01723729740419260
23:09:45.672932	10	14	170	0,01716116585258520
23:12:51.620818	10	14	210	0,01723729740419260
23:16:30.829038	10	14	250	0,01714166874790520
23:20:49.693344	10	14	290	0,01719830319483270
23:25:40.578356	10	14	330	0,01718344825793370
23:31:18.412130	10	16	10	0,01688449265284100
23:31:53.236943	10	16	50	0,01641377683985350
23:33:08.360704	10	16	90	0,01636178456070700
23:35:06.742628	10	16	130	0,01631629131645380
23:37:40.197480	10	16	170	0,01632371878490330
23:40:58.901373	10	16	210	0,01637106889626890
23:44:50.313583	10	16	250	0,01632464721845940
23:49:23.812165	10	16	290	0,01628658144265570
23:54:36.614972	10	16	330	0,01622809012861590
15:23:28.674236	10	20	10	0,01634321588958320
15:24:06.834019	10	20	50	0,01586692947525850
15:25:25.593794	10	20	90	0,01579822539210070
15:27:30.618284	10	20	130	0,01581029502833120
15:30:17.808623	10	20	170	0,01579544009143210
15:33:45.183367	10	20	210	0,01579544009143210
15:37:53.708537	10	20	250	0,01578894105653880
15:42:43.769724	10	20	290	0,01578151358808930
15:48:20.946572	10	20	330	0,01579172635720730
15:54:37.803938	10	25	10	0,01650569176191610
15:55:14.790531	10	25	50	0,01585393140547190

15:56:36.024412	10	25	90	0,01576851551830270
15:58:39.877176	10	25	130	0,01576944395185880
16:01:26.659829	10	25	170	0,01579915382565680
16:04:57.033805	10	25	210	0,01576201648340930
16:09:11.099833	10	25	250	0,01574066251161690
16:14:02.961585	10	25	290	0,01576480178407790
16:19:37.434070	10	25	330	0,01572766444183030
16:26:05.625820	10	50	10	0,01647876718878670
16:26:45.120869	10	50	50	0,01591799332084890
16:28:05.413658	10	50	90	0,01583072056656730
16:30:08.090512	10	50	130	0,01573694877739220
16:32:54.513027	10	50	170	0,01572209384049320
16:36:24.371650	10	50	210	0,01575273214784740
16:40:38.008026	10	50	250	0,01577872828742070
16:45:40.070801	10	50	290	0,01578337045520170
16:51:18.760010	10	50	330	0,01572395070760560