



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



Escola Tècnica
Superior d'Enginyeria
Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica

Universitat Politècnica de València

Diseño e implementación de un Almacén de Datos para el análisis de los datos del COVID19

Trabajo Fin de Grado

Grado en Ingeniería Informática

Autor: Santiago Joel Elizalde Carrión

Tutora: Laura Mota Herranz

Curso 2021 - 2022

Resumen

Con el aumento desmesurado de datos informáticos que se han generado en los últimos años, han surgido diversos problemas en las organizaciones para conseguir el cumplimiento de sus fines. Uno de los problemas principales es que las herramientas convencionales que se usaban hasta el momento no estaban a la altura para procesar, almacenar y analizar grandes volúmenes de datos de forma eficiente. Es por ello por lo que surgen los Almacenes de Datos y los Sistemas de Información Estratégicos para cubrir estas necesidades mejorando la obtención de conocimientos a partir de los datos históricos albergados.

En este proyecto final de carrera, se presenta una solución económica, desarrollada desde cero, en el cual se diseñará un Almacén de Datos enfocado a la crisis del COVID-19. Además, se indicarán las operaciones realizadas de extracción, transformación y carga (conocidas comúnmente como ETL) sobre el banco de datos original. Posteriormente, se pondrá a prueba la información transportada al esquema final mediante una serie de informes de prueba empleados por un software *Business Intelligence*.

Cabe destacar que esta solución ha sido desarrollada tanto para el ámbito personal como el ámbito académico.

Palabras clave: Almacén de Datos, *Business Intelligence*, ETL, datos históricos.

Abstract

With the disproportionate increase in computer data that has been generated in recent years, several problems have arisen in organizations to achieve the fulfilment of their purposes. One of the main problems is that the conventional tools, which were used until now, were not up to the task to process, store and analyse large volumes of data efficiently. It is for this reason that Data Warehouses and Strategic Information Systems arise to meet these needs by improving the obtaining of knowledge from the historical data stored.

In this final degree project, an economic solution is presented, developed from scratch, in which a Data Warehouse will be designed focused on the COVID-19 crisis. In addition, the extraction, transformation and loading operations performed (commonly known as ETL) on the original database will be indicated. Subsequently, the information transported to the final schema will be tested by means of a series of test reports used by a Business Intelligence software.

It should be noted that this solution has been developed for both the personal and academic fields.

Keywords: Data Warehouse, Business Intelligence, ETL, historical data.

Resum

Amb l'augment desmesurat de dades informàtiques que s'han generat els últims anys, han sorgit diversos problemes a les organitzacions per aconseguir el compliment dels seus fins, Un dels problemes principals és que les eines convencionals, que s'utilitzaven fins ara no estaven a l'alçada per processar, emmagatzemar i analitzar grans volums de dades eficientment. És per això que sorgeixen els Magatzems de Dades y els Sistemes d'Informació Estratègics per cobrir aquestes necessitats millorant l'obtenció de coneixements a partir de les dades històriques albergades.

En aquest projecte final de carrera, es presenta una solució econòmica, desenvolupada des de zero, en el qual es dissenyarà un Magatzem de Dades enfocat a la crisi del COVID-19. A més, s'han d'indicar les operacions realitzades d'extracció, transformació i càrrega (conegudes comunament com a ETL) sobre el banc de dades original. Posteriorment, es posarà a prova la informació transportada a l' esquema final mitjançant una sèrie d'informes de prova emprats pel software *Business Intelligence*.

Cal destacar que aquesta solució ha estat desenvolupada tant per a l'àmbit personal com a l'àmbit acadèmic.

Paraules clau: Magatzem de Dades, *Business Intelligence*, ETL, dades històriques.

Índice

1	Introducción.....	9
1.1	Objetivos.....	9
1.2	Motivación.....	9
1.3	Metodologías.....	10
1.4	Estructura de la memoria.....	11
2	Contexto tecnológico.....	13
2.1	Significado de SIE.....	13
2.2	Introducción a los Almacenes de Datos.....	14
2.2.1	Propiedades de un Almacén de Datos.....	14
2.2.2	OLTP.....	16
2.2.3	OLAP.....	16
2.3	Herramientas OLAP.....	17
2.4	Metodología de diseño de un Almacén de Datos.....	20
2.4.1	Fase de Análisis.....	21
2.4.2	Fase de Diseño.....	21
2.4.2.1	Esquema en estrella.....	22
2.4.2.2	Esquema en copo de nieve.....	23
2.4.3	Fase de Implementación.....	25
2.5	Herramientas ETL.....	25
2.5.1	Fases de un proceso ETL.....	26
2.5.2	Funciones de un sistema ETL.....	27
2.5.3	Selección de herramienta ETL.....	28
3	Plan de trabajo.....	31
3.1	Planificación de las fases.....	31
3.2	Presupuesto.....	32
4	Análisis del problema.....	34
4.1	Temática seleccionada.....	34
4.2	Origen del banco de datos.....	34
5	Diseño y desarrollo de la solución.....	37
5.1	Diseño conceptual.....	37
5.2	Diseño lógico.....	45
5.3	Diseño físico.....	48
6	Implementación de la solución.....	49
6.1	Fuentes.....	50
6.2	Dimensiones.....	51
6.3	Hechos.....	55
6.4	Ampliación mediante DAX.....	57
7	Explotación del Almacén de Datos.....	60
8	Conclusiones.....	64
9	Bibliografía.....	65
10	Anexo.....	66
10.1	Objetivos de desarrollo sostenible.....	66

Índice de Ilustraciones

Ilustración 1 Características de un Almacén de Datos	15
Ilustración 2 Sistemas operacionales vs Sistemas analíticos.....	17
Ilustración 3 Arquitectura de un Sistema de Almacén de Datos	18
Ilustración 4 Esquema conceptual en estrella.....	23
Ilustración 5 Esquema relacional en estrella.....	23
Ilustración 6 Esquema conceptual en copo de nieve.....	24
Ilustración 7 Esquema relacional en copo de nieve.....	25
Ilustración 8 Proceso ETL: Extracción, Transformación y Transporte de datos.....	26
Ilustración 9 Portal web ourworldindata.org. Recuperado de https://ourworldindata.org/coronavirus	35
Ilustración 10 Portal web github.com, origen del banco de datos COVID-19 Dataset. Recuperado de https://github.com/owid/covid-19-data	35
Ilustración 11 Esquema conceptual representado mediante un diagrama UML	38
<i>Ilustración 12 Esquema relacional</i>	45
Ilustración 13 Consultas Power Query del modelo multidimensional.....	50
Ilustración 14 Configuración de la extracción de datos del origen.....	51
Ilustración 15 Función del Lenguaje M Date.StartOfWeek para obtener el primer día de la semana	52
Ilustración 16 Operaciones para cumplimentar las restricciones de clave primaria de la consulta TIEMPO_DIARIO.....	52
Ilustración 17 Operación de combinación de consultas para traducir los valores de ZONA_GEOGRAFICA	54
Ilustración 18 Filas de datos de carácter sumatorio que se han suprimido mediante un filtro en la consulta ZONA_GEOGRAFICA.....	55
Ilustración 19 Funciones DAX para sacar el nombre, el mes, el año y el número de la semana del año de un campo DATE.....	57
Ilustración 20 Jerarquía de la tabla TIEMPO_SEMANAL	58
Ilustración 21 Jerarquía de la tabla TIEMPO_DIARIO	58
Ilustración 22 Jerarquía de la tabla ZONA_GEOGRAFICA.....	58
Ilustración 23 Representación gráfica del modelo final del Almacén de Datos en Power BI.....	59
Ilustración 24 Gráfico de columnas agrupadas del número total de casos COVID-19 por cada zona geográfica junto a dos paneles de segmentación.....	60

Ilustración 25 Gráfico de columnas agrupadas del número total de casos COVID-19 acotado por el mes de enero y el continente de Europa61

Ilustración 26 Mapa coroplético del número total de muertes que hay en cada uno de los países almacenados en el modelo junto a un panel de segmentación por continentes.62

Ilustración 27 Gráfico circular del total de la tasa de camas que hay en cada continente y el respectivo porcentaje de la población que vive en extrema pobreza63

Índice de tablas

Tabla 1 Diagrama de Gantt: Planificación de las fases y estimación temporal	32
Tabla 2 Costes económicos del proyecto. Recuperado de https://es.indeed.com/career/salaries/sql?from=whatwhere	33
Tabla 3 Descripción de cada una de las clases del modelo	39
Tabla 4 Dimensión TIEMPO_DIARIO con las descripciones de sus atributos.	39
Tabla 5 Dimensión TIEMPO_SEMANAL con las descripciones de sus atributos	39
Tabla 7 Hecho principal IMPACTO_DIARIO con las descripciones de sus atributos	44
Tabla 8 Hecho principal IMPACTO_SEMANAL con las descripciones de sus atributos	45

1 Introducción

1.1 Objetivos

El objetivo de este proyecto es realizar el diseño y la implementación de un Almacén de Datos basado en la información sobre el impacto que tuvo a nivel internacional la aparición del virus COVID-19.

Posteriormente, con la explotación del Almacén de Datos creado, se podrían elaborar informes por parte del usuario con el fin de llevar a cabo análisis estadísticos de gran profundidad sobre la pandemia para gran parte de los países del mundo. Se trata de una cuestión sumamente delicada ya que se ha conseguido recopilar los datos de los avances de cada territorio geográfico de la lucha contra el virus de forma que, finalmente, se pueda hacer una evaluación integral.

Teniendo en cuenta las ventajas que ofrecen los Sistemas de Información Estratégicos actuales, será necesario optimizar el Almacén de Datos a desarrollar para brindar a los usuarios una experiencia ágil y sencilla en la explotación de datos de dicho almacén. Para ello, es crucial pensar con detalle cada uno de los pasos a seguir para lograr dicha meta.

Para la consecución de los objetivos, será necesario establecer un punto de partida y seguir un camino que, a su vez, cumpla los siguientes subobjetivos:

- Buscar un banco de datos relacionado con la temática propuesta que se adapte a las necesidades del proyecto.
- Comparar las diferentes soluciones software existentes en el mercado y seleccionar la que mejor se ajuste al desarrollo del Almacén de Datos.
- Analizar el carácter del banco de datos para que, posteriormente, se pueda modelar la arquitectura del almacén resultante.
- Comprobar la correlación de los datos albergados en el Almacén de Datos para descartar posibles presentaciones de resultados incorrectos.
- Conseguir que el Banco de Datos pueda retroalimentarse de la fuente de datos automáticamente.

1.2 Motivación

Uno de los motivos por los cuales decidí estudiar la carrera de Ingeniería Informática, aparte de las salidas laborales que conlleva cursarla y terminarla, ha sido la curiosidad de poder entender y observar el proceso que ha tenido que seguir el ser humano para conseguir los avances tecnológicos comprendidos en las TIC que hoy en día nos resultan conceptos de lo más normales y cotidianos cuando estas concepciones, hace menos de tres décadas, eran tan sólo ideas inconcebibles fruto de una mente ingeniosa que se dirige en camino a la innovación.

A lo largo de los primeros años de la carrera, fui entendiendo cómo se pudo llegar a estas metas y logros, pero no fue hasta que me impartieron, en el tercer año de carrera, la asignatura de Base de Datos y Sistemas de Información (BDA) cuando de verdad me fascinó el grado, concretamente, las interacciones (consultar, actualizar y reorganizar datos) que se pueden realizar con una base de datos relacional en SQL. Quedaba asombrado sólo de pensar que hace tiempo, en la antigüedad, lo que conocemos como un sistema formado por un conjunto de datos almacenados en disco, tan sólo era una simple biblioteca en donde cualquier persona que quisiera consultar o recopilar información, tenía que hacerlo de forma manual, por ejemplo, las búsquedas del histórico de las cosechas que se consiguieron años anteriores. Realizar esta labor era de poca eficacia ya que no se contaba con la ayuda de ninguna máquina y mucho menos, con ningún software de Sistema de Gestión de Base de Datos.

Esta fascinación se reforzó de forma abrumadora, cuando me impartieron por primera vez, en el primer cuatrimestre del cuarto año de carrera, la asignatura de Sistemas de Información Estratégicos (SIE). El uso que se le ha dado a los Sistemas de Información, a lo largo de los años en el mundo empresarial para optimizar las operaciones empresariales con el fin de conseguir una mejor posición en el mercado, es algo que da para hablar. Especialmente, me llamó la temática de los Almacenes de Datos o *Data Warehouse* y la herramienta *Business Intelligence* para poder realizar análisis de datos, denominada Power BI de Microsoft. Con todo lo que pude aprender en aquel entonces, me bastó para saber que mi carrera profesional se iba a enfocar por esa rama.

Cuando comencé a trabajar por segunda vez en mis prácticas extracurriculares en la empresa CesumIn S.L, puse en práctica con entusiasmo y con ganas de aprender más todo lo aprendido. Eso conllevó a que uno de mis compañeros de trabajo, que resultó ser el responsable del departamento de TIC, me propusiera realizar un almacén de datos desde cero a partir de un banco de datos extenso, una herramienta ETL y un software *Business Intelligence*. Esta idea se reforzó más cuando mi tutora de prácticas me propuso el mismo concepto cuando le comenté que quería presentar un TFG relacionado con bases de datos. Por este motivo, decidí desarrollar y presentar este proyecto.

1.3 Metodologías

A la hora de iniciar el desarrollo de un proyecto es indispensable plantearse desde un principio qué metodología usar en función de los requisitos de lo que se van a desarrollar en el trabajo. Es por esto por lo que hay que analizar y posteriormente seleccionar una metodología de las más usadas de los últimos momentos. Tenemos dos en concreto: *Waterfull* y *Agile*.

Por un lado, tenemos el proceso *Waterfull* también conocido como “cascada” en lengua castellana. Es uno de los métodos que más se ha usado tradicionalmente. Consiste en desarrollar un proyecto de forma secuencial. Para poder avanzar a la siguiente fase del proyecto, es necesario validar cada una de las fases antecesoras y darlas por concluidas. De tal forma que, si surge algún tipo de inconveniente en alguna de las fases del proyecto, se realiza un *backtracking* (vuelta atrás) desde el punto en donde se sospecha que ha surgido el fallo hasta la fase en donde se pausó el proceso. Posteriormente, se sigue el flujo del proyecto.

Por otro lado, tenemos el proceso *Agile* también conocido como “metodología ágil” en español. Consiste en realizar el desarrollo de un proyecto de forma simultánea, es decir, cada una de las fases del proyecto se puede realizar al mismo tiempo. Para poder desempeñar esta forma de trabajo, es necesario contar con un equipo con capacidades autoorganizativas y multifuncionales.

Con respecto a las ventajas de *Waterfall* tenemos:

1. Tiene una planificación y un diseño más sencillos y directos ya que cuenta con la participación temprana con el cliente.
2. El progreso del proyecto es más fácil de medir.
3. No se requiere la presencia estricta del cliente en todo momento.
4. Es idónea para el desarrollo de múltiples componentes software.
5. Se puede realizar un diseño completo y con más cautela del software.

Por otro lado, tenemos las ventajas de Agile:

1. El cliente se compromete a permanecer en el desarrollo.
2. El cliente tiene mayor participación, por lo tanto, se puede tener un mayor número de oportunidades para realizar modificaciones.

Una vez mencionado ambos procesos y sus ventajas y dadas las circunstancias de tiempos en las reuniones, tanto mi tutora de prácticas como yo hemos optado por elegir la metodología Waterfall. Es por ello por lo que, a lo largo del proceso de desarrollo del TFG, se han hecho reuniones con mi tutora, antes de avanzar al siguiente punto de control, para comentar los avances que se han realizado y evaluar las soluciones propuestas.

1.4 Estructura de la memoria

La memoria, que se ha desarrollado de forma paralela junto al proyecto de investigación, está constituida por los siguientes apartados que se comentan a continuación.

En el primer apartado, denominado Introducción, se establecen los objetivos del proyecto. Seguidamente, en el subapartado Motivación, se comentan cuáles han sido las circunstancias por las cuales se ha escogido el tema propuesto. Finalmente, se justifica qué metodología se ha optado para la organización del proyecto, en acorde a las necesidades tanto del estudiante como de la tutora del proyecto.

En el segundo apartado, Contexto tecnológico, se presenta un desglose de todos los conceptos teóricos relacionados con el posible desarrollo del Almacén de Datos. Para la organización de cada subapartado, se han utilizado como referencias las unidades temáticas correspondientes a la asignatura de SIE de la ETSIF (Universidad Politécnica de Valencia). A su vez, en el subapartado de herramientas OLAP, se harán unos breves estudios comparativos para seleccionar los posibles softwares a utilizar en este proyecto.

En el tercer apartado se presenta el plan de trabajo para la realización del trabajo, así como el presupuesto que ha conllevado.

El cuarto apartado, denominado Análisis del problema, se centra en contextualizar la temática que se ha seleccionado para el desarrollo del Almacén de Datos. A su vez, se indica el origen de la fuente de datos con la cual se retroalimentará dicho almacén.

En el quinto apartado, Diseño y desarrollo de la solución, se detalla cada uno de los pasos realizados para el diseño y desarrollo del Almacén de Datos. Es por ello, que se hace hincapié en todas las fases que constituyen al diseño de la solución, desde la fase del diseño conceptual hasta la fase del diseño físico.

En el sexto apartado, Implementación de la solución, se presentan cada una las tareas que hay que realizar dentro de la herramienta Business Intelligence para poder construir la solución planteada.

En el séptimo apartado, Explotación del Almacén de Datos, se centra en poner a prueba la solución implementada. Esta consiste en generar informes de prueba a partir de los datos albergados en el Almacén de Datos.

En el octavo apartado, Conclusiones, se analizan los puntos fuertes y débiles del proyecto a partir de los objetivos que se plantearon inicialmente, también se incluye una breve valoración.

Finalmente, en el noveno y último apartado, Bibliografía, se listan cada una de las referencias que se han usado para la elaboración del proyecto.

2 Contexto tecnológico

En este apartado se realiza una exposición detallada de los conceptos relevantes en la temática de los Almacenes de Datos y Sistemas de Información Estratégicos.

2.1 Significado de SIE

A lo largo de los años, se han planteado diversas definiciones para un Sistema de Información Estratégico (SIE), una de las cuales, de las más acertadas desde mi punto de vista, dice que son *“sistemas de cómputo a cualquier nivel que cambian metas, operaciones, productos, servicios o relaciones con el entorno que ayudan a la empresa a obtener una ventaja competitiva.”* (Laudon y Laudon, 1996)

Analizando la definición de SIE que se ha mencionado anteriormente, se entiende que se trata de una herramienta fundamental dentro del mundo empresarial para las compañías que tienen como objetivo el ánimo de lucro.

Una de las razones por la cual, la mayoría de las organizaciones, cuentan hoy en día con un Sistema de Información Estratégico es porque, en gran parte, realizan operaciones que abarcan y afectan a todos los departamentos de una empresa, como, por ejemplo: reducir costes de un proyecto, comprender y atender la opinión y el pensamiento de los clientes, optimizar los plazos de tiempo de un proceso empresarial, aumentar las ventas, ... Todos estos ejemplos son objetivos primordiales que una empresa intenta optimizar y solventar tras el día a día. Es por ello por lo que es importante que una empresa pueda contar con la ayuda de un instrumento capaz de proporcionarle, de forma fácil, asequible e inmediata, toda esta información para que pueda realizar cálculos de los rendimientos obtenidos. En consecuencia, se podría considerar que un Sistema de Información Estratégico es un aliado más dentro del barco empresarial que navega viento en popa para atravesar las adversidades del mercado en el que se sitúan. En contraparte, aquellas empresas que se nieguen a considerar y a usar estas herramientas, están destinadas a realizar esfuerzos innecesarios debido a una mala gestión de la información y, por consiguiente, comenzarán a hundirse junto a toda la organización en alta mar a causa de los futuros problemas con los clientes, los competidores, principalmente, los gastos económicos.

El mercado en el que se encuentran todas las empresas de un sector sufre constantes cambios y devenires que afectan para bien o para mal a la organización. A todo esto, hay que sumar la cuestión de que cada una de estas compañías, hacen uso de todos los recursos necesarios para obtener mejor posicionamiento en el terreno competitivo. Es por esta misma razón que, usar un SIE, se traduce en tener una gran ventaja competitiva. Por otro lado, hay que entender que tener una ventaja competitiva en una organización es poseer una característica cualificada que permite a una organización diferenciarse del resto de competidores, consiguiendo una situación muchísimo más favorable en el posicionamiento del mercado actual.

A lo largo de las últimas décadas, la funcionalidad de un SIE ha tenido diversos cambios sobre todo a nivel administrativo ya que pueden dirigir cualquier tipo de negocio para poder instaurar

mejoras en todos los niveles. Es por esta razón que, se puede afirmar que se trata de una poderosa herramienta para la gestión de datos en el órgano administrativo.

El desempeño de los SIE ha hecho que muchos eslabones de una empresa sufran las modificaciones necesarias para así dejar la puerta abierta a la introducción de las mejoras que se puedan instaurar (como se ha mencionado antes) en función de las necesidades que un negocio tiene. Este movimiento estratégico se puede realizar siempre y cuando se tenga de algún tipo de historial con el cual la empresa, que disponga del SIE, pueda apoyarse para iniciar la denominada toma de decisiones. Es aquí donde entran en el terreno de juego los activos estratégicos, procedentes de los SIE, que se usan en las organizaciones para poder recopilar un histórico de toda la información implicada en una empresa a lo largo del tiempo. Esta información histórica, organizada y depurada, se guarda en un Almacén de Datos (AD) (en inglés *Data Wharehouse*).

2.2 Introducción a los Almacenes de Datos

Una vez llegados a este punto, hay que adentrarse en la explicación de una de las ramas software de los SIE, que también se trata de una de las herramientas más usadas y sublimes por parte de las organizaciones, los Sistemas de Gestión de Almacenes de Datos.

Este tipo de sistema de gestión de datos se trata de nada más y nada menos que de un repositorio unificado (puede ser tanto físico como lógico) que almacena todos los datos recogidos por los diversos sistemas de una empresa, es decir, de diferentes orígenes de recolección de información, cuyo fin es la de permitir a los ejecutivos de la organización organizar, comprender, realizar análisis y utilizar los datos para mejorar la toma de decisiones estratégicas.

Esta excelente arquitectura para recapitular datos de un organismo es ya conocida en muchas empresas modernas ya que puede estructurar, centralizar y consolidar grandes cantidades de datos de múltiples fuentes. Cada una de estas entradas de datos abarcan tanto procesos externos como internos del propio negocio, dando como resultado la capacidad para que ningún gerente de un órgano administrativo tenga que tirar de corazonadas, datos incompletos o de muy mala calidad. De esta forma, se pueden evitar la obtención de resultados lentos, incoherentes o inexactos.

2.2.1 Propiedades de un Almacén de Datos

Una vez que se ha explicado el concepto de los Almacén de Datos. Hay que tener en cuenta estas cuatro características (Naeem, 2020):

1. **Orientado al tema:** Un Almacén de Datos se construye con el propósito de interactuar en torno a un tema en particular en lugar de para realiza operaciones en base al interés de una empresa. En otras palabras, el proceso de almacenamiento de datos está

destinado a enfocarse en un tema en específico. Se pueden encontrar ejemplos como: el control de las ventas, el *repricing*¹ elaborado en el departamento de marketing, etc.

2. **Integrado:** Esta característica se define como “el proceso empleado para consolidar información desde múltiples fuentes (como redes sociales, datos de sensores del Internet de las cosas, almacenes de datos, transacciones de clientes, etc.) y compartir una versión limpia y actualizada en toda una organización” (Talend, 2021). Esta cualidad es considerada como una de las más importantes, desde un principio por todas las organizaciones que invierten en mejorar sus herramientas de gestión, debido a que se trata de uno de los procesos más costosos tanto en gastos temporales como gastos económicos ya que, los datos que están unificados y albergados en el AD, provienen de diferentes fuentes de datos usados dentro y fuera de la empresa, por lo que es necesario optimizar este apartado para garantizar: la exclusión de datos duplicados, datos de poco interés y mejorar los resultados para ofrecer una mayor calidad a los consultores.
3. **No volátil:** La información que va entrando por las fuentes de datos del AD, se va incrementando de forma periódica. Esto quiere decir que, los nuevos datos que entran en el repositorio no eliminan ni tampoco sobrescriben los datos ya almacenados, como hacen las Tecnologías de Bases de Datos Relacionales al realizar una operación de escritura. En pocas palabras, los datos son permanentes.
4. **Variante de tiempo:** Esta característica es mucho más extensa que el resto de los sistemas operativos. La información almacenada en el almacén es relativa a un periodo de tiempo. Esto permite ofrecer información a través de un punto de vista histórico.

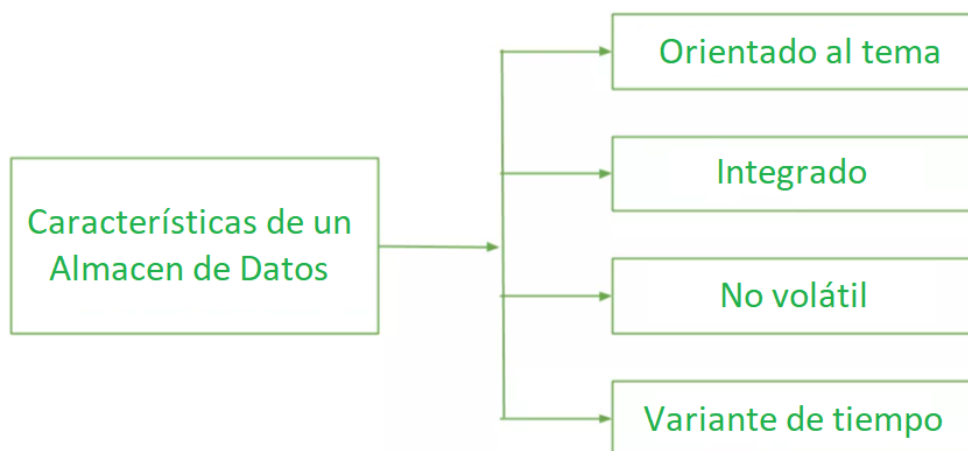


Ilustración 1 Características de un Almacén de Datos

Una vez comentado los aspectos más relevantes de un Almacén Datos, es necesario, tener la capacidad de poder distinguir dos tipos de tecnologías involucradas en la explicación de diferentes modelos representativos en la gestión de datos, de esta forma, se evitarán posibles confusiones por parte del usuario que se disponga a utilizarlas. Se tratan de los sistemas de procesamiento operacionales o transaccionales en línea (OLTP) y los sistemas analíticos en línea (OLAP).

¹ Proceso empresarial en el cual, se minimizan los precios de venta de los productos o servicios en función de los precios de la competencia. Su cometido es mejorar las ventas en el mercado situado.

2.2.2 OLTP

Los sistemas OLTP (*OnLine Transaction Processing*) se caracterizan por tener bases de datos relacionales como repositorio de información con una gran cantidad de operaciones cortas que utiliza el usuario para realizar modificaciones en los datos albergados en disco. Estas operaciones se utilizan para llevar a cabo acciones de cómputo de lectura, actualización, inserción y eliminación de datos, correspondientes a las instrucciones SELECT, UPDATE, INSERT y DELETE respectivamente. Además, cada una de las transacciones de escritura va acompañada por técnicas de validación (COMMIT) o de invalidación (ROLLBACK). La información de los sistemas transacciones tiende a ser volátil por lo que no se trata de información con carácter histórico. Esto es debido a que este tipo de infraestructura está diseñada para mantener únicamente información actual. Gracias a esta característica, el volumen de datos, con el cual trabaja, es relativamente bajo (en comparación de los sistemas OLAP) y esto conlleva a que sean muchísimo más veloces al realizar una transacción. También hay que tener en cuenta, las cuatro propiedades ACID para que este tipo de sistemas cumpla con una base de datos correcta y coherente:

1. **Atomicidad:** Todas las transacciones tienen un carácter atómico, es decir, o se ejecutan todas sus operaciones o no se ejecuta ninguna.
2. **Consistencia:** Para mantener una base de datos consistente, es necesario aceptar únicamente las transacciones que cumplan con las restricciones de integridad.
3. **Aislamiento:** Entre transacciones no debe haber interferencias. Por lo que se deben ejecutar como si se trataran de transacciones aisladas.
4. **Persistencia:** Las modificaciones generadas por todas las transacciones que hayan sido confirmadas por el sistema, se guardan en la base de datos en disco.

2.2.3 OLAP

Los sistemas OLAP (*On Line Analytical Processing*) que se caracterizan por tener un repositorio de información de gran volumen, de tal forma que sufren un volumen de transacciones bajo. En concreto, la acción más común es la de realizar operaciones de lectura, es decir, llevar a cabo consultas, correspondientes a la instrucción SELECT del lenguaje SQL, también se llevan a cabo, con muy poca frecuencia, operaciones de escritura. Las grandes cantidades de datos, con las que opera este tipo de sistema, necesitan cumplir la propiedad de no ser volátiles en el tiempo, con el fin de mantener la información con un carácter histórico de cara a realizar análisis para la toma de decisiones.

Sistema Operacional (BD-OLTP)	Sistema de almacén de datos (AD-OLAP)
- Almacena datos actuales.	- Almacena datos históricos.
- Almacena datos de detalle.	- Almacena datos de detalle y agregados a distintos niveles.
- Bases de datos medianas.	- Bases de datos grandes.
- Los datos son dinámicos (actualizables).	- Los datos son estáticos.
- Los procesos (transacciones) son repetitivos.	- Los procesos no son previsibles.
- El número de transacciones es elevado.	- El número de transacciones es bajo o medio.
- Tiempo de respuesta pequeño (segundos).	- Tiempo de respuesta variable (segundos-horas).
- Dedicado al procesamiento de transacciones.	- Dedicado al análisis de datos.
- Orientado a los procesos de la organización.	- Orientado a la información relevante.
- Soporta decisiones diarias.	- Soporta decisiones estratégicas.
- Sirve a muchos usuarios (administrativos).	- Sirve a técnicos de dirección.

comparación

Ilustración 2 Sistemas operacionales vs Sistemas analíticos

Dentro de los sistemas OLAP, se puede apreciar que utiliza dos tipos de servidores en los Almacenes de Datos:

- **Tecnología Relacional**, denominadas comúnmente como ROLAP. Se trata de Sistemas de gestión que optan a realizar los análisis de un AD sobre bases de datos relacionales. Sobre estos sistemas se acoplan herramientas OLAP como Pentaho o Power BI.
- **Tecnología Multidimensional**, denominadas comúnmente como MOLAP. Se trata de sistemas de almacenamiento que han sido optimizados para el análisis de datos mediante una visión multidimensional del objeto a estudiar.

En el siguiente apartado se analizará con más detenimiento las herramientas OLAP necesarias para la explotación de datos de un Almacén de Datos y las características de una Tecnología de datos con carácter multidimensional.

2.3 Herramientas OLAP

Una vez se han comentado los aspectos fundamentales sobre los Almacenes de Datos, es necesario pararse a analizar y a estudiar cuáles son las herramientas con las que se pueden sacar uso de la información albergada en estos tipos de almacenamientos estratégicos. Para ello, se realizará un análisis de la arquitectura de un sistema de Almacén de Datos.

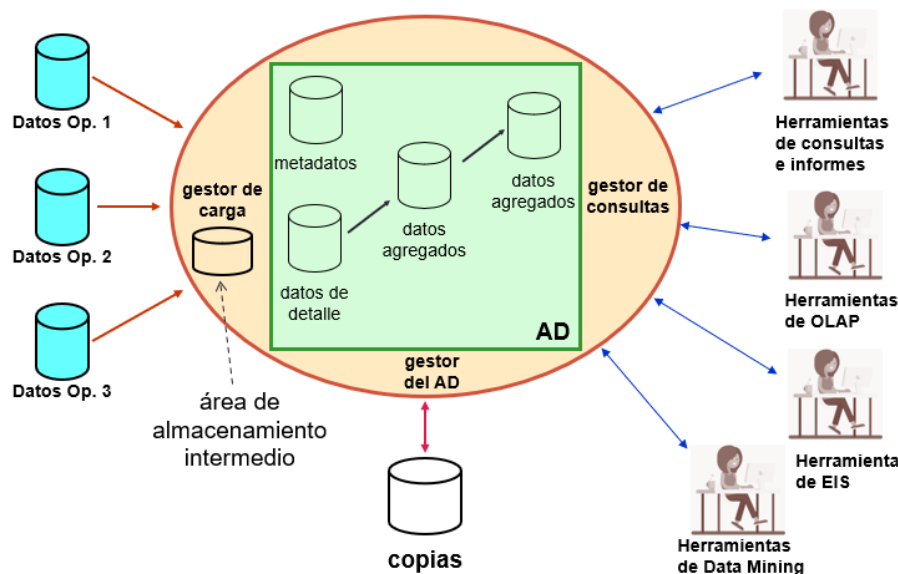


Ilustración 3 Arquitectura de un Sistema de Almacén de Datos

Este modelo ha tomado gran importancia a lo largo de estas últimas décadas ya que ha sido uno de los pilares más destacados e imprescindibles dentro del ecosistema de datos corporativos. A pesar de todos los cambios que han sufrido en el campo de la computación en la nube, en las tecnologías de la información y sobre todo, en el mundo de *Big Data*, no ha hecho más que tomar gran relevancia y subir peldaños en los Sistemas de Gestión de Datos.

Centrándose en esta arquitectura, dentro de los gestores de consultas de los Almacenes de Datos, tenemos las siguientes herramientas, externas al almacén, que hacen uso de los datos almacenados:

- **Herramientas de consultas e informes:** Se tratan de sistemas que permiten a los usuarios realizar la exploración de datos del Almacén de Datos desempeñando la función de nexo entre el depósito de la información y los usuarios. Mediante interfaces gráficas y una serie de pasos, entre ellas el uso de consultas estáticas, acceden a los metadatos almacenados para obtener resultados a través de consultas SQL. Por ello, se tratan de sistemas que hacen uso de esquemas relacionales para ofrecer resultados, que luego son expuestos a los usuarios en formatos que les sea familiar.
- **Herramientas de Data Mining (Minería de Datos):** Se trata de un poderoso instrumento, con la tecnología necesaria para permitir brindar al usuario soporte, para realizar tareas como el análisis y la extracción de conocimientos ocultos a partir de la información almacenada en un Almacén de Datos. Realizar Minería de Datos a partir de un Almacén de Datos, permite el análisis (mediante un proceso inductivo, en el que el sistema usa patrones, pautas o reglas predecibles) de factores de influencia en determinados casos de una empresa. Con ello, se consigue realizar predicciones a partir de estos casos de la organización, como pueden ser: estimar variables, comportamientos venideros, estimaciones de valores, etc.
- **Herramientas OLAP:** Se trata de una de las herramientas más reconocidas dentro de los Almacenes de Datos ya que es el motor de consultas de tipo dinámico especialista en el depósito de datos. Las herramientas OLAP son la tecnología software para la ejecución de análisis en línea, administración de datos y ejecución de consultas que permite extraer y estudiar información del negocio que las

dispone. Es mediante estas herramientas con las que las organizaciones pueden extraer información de interés, a partir de los datos almacenados en el Almacén de Datos, que refleja la situación del negocio.

Teniendo en cuenta todas las posibles utilidades que puede tener un Almacén de Datos y los requisitos del proyecto a desarrollar que implican la elaboración de informes estadísticos, se optará por el uso de una herramienta OLAP.

Se ha optado por usar una herramienta de este tipo ya que son las que más se adecuan a la hora de llevar a cabo la explotación de datos con el fin de realizar análisis a partir de informes elaborados mediante información segura y confiable. A su vez, las soluciones se apoyan con interfaces gráficas e intuitivas para que los usuarios o gestores administrativos, que se dedican a realizar los estudios a partir de los datos, puedan realizarlo de forma óptima evitando cualquier tipo de confusión o redundancia ya que, desde un principio, las herramientas OLAP dan la libertad de seleccionar y manipular únicamente aquellos datos que nos interesen.

Dentro de este tipo de herramientas hay un gran abanico de opciones para escoger. Es por ello, que será necesario llevar a cabo un análisis previo, para ello, se han tenido en cuenta y se han seleccionado dos de las herramientas más reconocidas dentro del mercado. Hay que tener presente que la herramienta que salga seleccionada cumpla con los requisitos mínimos, como es el caso de que se encuentre a nuestro alcance y que, sobre todo, cumpla con las características necesarias para que, pueda ser una solución viable al proyecto en cuanto al cometido de la explotación de datos del AD. Las posibles opciones son las siguientes:

- **Tableau:** se trata de una de las plataformas más conocidas de análisis de datos, para la visualización de datos interactivos.
 - **Ventajas:**
 - Tiene mayor madurez ya que lleva más tiempo en el mercado (fundada a principios de 2003).
 - **Desventajas:**
 - Se adecua mejor a trabajar únicamente con grandes volúmenes de datos. Enfocada a grandes compañías.
 - Para un usuario que está teniendo su primera toma de contacto con esta herramienta le resultará más tedioso dominar la interfaz y aprender a usarla.
 - Exige que los datos de entrada estén limpios y estructurados.
- **Power BI:** se trata de uno de los servicios de Microsoft más interesantes a la hora de llevar a cabo análisis de datos a partir de inteligencia empresarial.
 - **Ventajas:**
 - Tiene mayor versatilidad al gestionar cualquier cantidad de datos. Enfocada tanto a grandes como a pequeñas empresas.
 - Para los usuarios que estén familiarizados con las herramientas de ofimática de Microsoft (Power Point, Excel, etc.) la interfaz le resultará familiar.
 - Flexibilidad en la carga de datos de entrada. Se pueden reprocesar los datos de los ficheros o bases de datos de alimentación para limpiar y organizar la información consultada.

- Desventajas:
 - Tiene menos tiempo en el mercado (Fundada a mediados de 2010)

En cuanto a las **similitudes** de ambas herramientas. Tenemos que:

- En ambas herramientas es fácil gestionar la información suministrada siempre y cuando el modelo de datos esté bien diseñado y construido con datos limpios. Con ello, se conseguirán representaciones graficas mucho más organizadas y de mayor calidad.
- Tanto en una herramienta como en la otra, son buenas candidatas para sumergirse en el mundo de *Big Data*.
- Ambas plataformas admiten múltiples entradas de datos de diferentes fuentes (tanto externas como internas de la organización) a través de importaciones de datos al propio programa o simplemente accediendo a través de una conexión en vivo para conseguir resultados dinámicos (resultados en caliente)
- Ambas herramientas tienen un coste de 0€ siempre y cuando se acepte que estaremos ante servicios limitados en los cuales únicamente se caparán funcionalidades como la compartición de informes y datos en línea.

En conclusión, ambas herramientas están situadas en la cúspide como líderes en el sector de la visualización gráfica de datos, proporcionándonos funcionalidades muy similares, como las que se han comentado anteriormente. Tanto una como otra nos pueden de ser de gran ayuda a la hora de sacar el máximo partido a nuestros datos. Sin embargo, optaremos por seleccionar la herramienta Power BI ya que:

- Se trata de uno de los programas que se usan en las prácticas de la materia SIE del grado en Ingeniería Informática.
- Trabajaremos con un banco de datos cuya estructura original habrá que reprocesar (uso de Power Query).

Por consiguiente, la opción seleccionada es la más rentable en cuanto a viabilidad y experiencia personal.

2.4 Metodología de diseño de un Almacén de Datos

La visión **multidimensional** seguida por las herramientas de explotación de almacenes de datos (OLAP) ha inspirado los modelos y metodologías de diseño de este tipo de sistemas, por ello en la literatura se habla de Diseño o Modelado Multidimensional que es la fase en la que se genera un esquema multidimensional del Almacén de Datos.

En un **esquema multidimensional**:

- Se representa una actividad que es objeto de análisis (**hecho**) y las propiedades que caracterizan el proceso (**dimensiones**).
- La información relevante sobre el **hecho** se representa por un conjunto de indicadores (**medidas o atributos de hecho**).

- La información descriptiva de cada **dimensión** se representa por un conjunto de atributos (**atributos de dimensión**).

El modelado multidimensional se puede aplicar utilizando distintos modelos de datos conceptuales o lógicos, en el caso conceptual, la representación gráfica del esquema dependerá del modelo de datos utilizado (relacional, ER, UML, OO, ...).

La metodología de diseño de un Almacén de Datos sigue las fases clásicas del diseño de una base de datos:

1. Análisis.
2. Diseño.
3. Implementación.

2.4.1 Fase de Análisis

Los pasos que hay que seguir en esta primera fase para construir un buen Almacén de Datos son los siguientes:

- **Establecer los requisitos iniciales:** es necesario definir el alcance que va a tener el Almacén de Datos a partir de los datos obtenidos en las entrevistas con los usuarios finales del sistema a diseñar. Estos datos se pueden conseguir mediante informes elaborados con anterioridad o mediante la recopilación de los requisitos iniciales de los usuarios mediante la elaboración de dichas entrevistas. Gracias a estos requisitos se pueden concluir qué herramientas se necesitan para el desarrollo del AD o si es necesario implementar algún tipo de herramienta de extracción además de recopilar documentación de datos o procesos faltantes.
- **Definir las reglas de negocio:** Una de las tendencias actuales durante el desarrollo de un Sistema de Información Estratégico es el camino que seguirá a partir del enfoque de las reglas de negocio. Por ello desde un primer momento es de vital importancia tener en cuenta dichas reglas a lo largo de la construcción del Almacén de Datos. Estas reglas, mejor conocidas como reglas de restricción, nos permiten establecer restricciones de integridad sobre los datos pertenecientes al modelo a desarrollar.
- **Identificar las fuentes de datos:** a partir de las necesidades expresadas por los usuarios mediante los datos de las entrevistas, es posible hacerse una idea de qué tipo de información se utilizará para alimentar el Almacén de Datos a construir. Esta información puede pertenecer a uno de los sistemas OLTP de la organización o puede ser una fuente de información externa a dicha organización.

2.4.2 Fase de Diseño

En el diseño de un Almacén de Datos, al igual que en la metodología de diseño de una base de datos clásica se subdivide en el diseño conceptual, diseño lógico y diseño físico que se presentan a continuación con más detalle.

1. **Diseño conceptual:** los pasos necesarios para completar el diseño conceptual del Almacén de Datos son:
 - **Paso 1: seleccionar un *proceso* o actividad de la organización** para el que se desea diseñar un almacén de datos. Usualmente es una actividad de la organización soportada por un OLTP de la cual se puede extraer información con el propósito de construir el almacén de datos.
 - **Paso 2: establecer el gránulo de representación del proceso.** El gránulo determina el nivel de detalle de la información que se va a almacenar en el Almacén de Datos. Se trata de una decisión crítica ya que determina tanto el significado de los hechos como las dimensiones básicas del esquema (información diaria, semanal, mensual, etc.).
 - **Paso 3: diseñar las dimensiones que caracterizan el proceso.** De cada dimensión se debe decidir los atributos (propiedades) relevantes para el análisis del proceso y encontrar las jerarquías naturales que existen entre esos atributos.
 - **Paso 4: decidir cuáles van a ser los atributos o medidas de los hechos,** es decir, hay que determinar qué información se desea almacenar en cada fila de la tabla de hechos.

El resultado de esta etapa es un esquema conceptual usualmente gráfico que puede diseñarse haciendo uso de alguno de los modelos conceptuales que existen. En nuestro caso usaremos el Diagrama de Clases del UML (Rumbaugh, Jacobson y Booch, 1998).

2. **Diseño lógico:** en esta etapa, el esquema conceptual obtenido se traduce a un esquema relacional en el caso de que la tecnología que se vaya a usar sea ROLAP.
3. **Diseño físico:** esta última fase tiene como objetivo buscar la mejor optimización de los tiempos de respuestas de las futuras consultas que se realicen en las tablas de dimensiones y hechos del AD. Uno de los factores que hacen que un Almacén de Datos destaque por su calidad de diseño, es que el usuario pueda hacer consultas sobre una gran cantidad de datos con un rango de miles de millones de registros en el menor tiempo posible. Para conseguir la optimización de los tiempos de las consultas del sistema a desarrollar, se pueden utilizar los siguientes recursos para el diseño físico:
 - Definición de índices.
 - Particionamientos de las tablas.

Por otro lado, tanto el esquema conceptual como el esquema lógico obtenidos en la fase de diseño pueden verse de dos formas distintas, lo que se llama el esquema en estrella y el esquema en copo de nieve que se explicarán a continuación.

2.4.2.1 Esquema en estrella

Si usamos el diagrama de clases para representar el esquema conceptual, las componentes del AD se conectan en un esquema con forma de estrella donde:

- El hecho objeto de análisis está en el centro representado por una clase débil de todas o algunas de las clases que representan las dimensiones.

- Cada una de las dimensiones en una punta de la estrella y está representada por una clase fuerte:

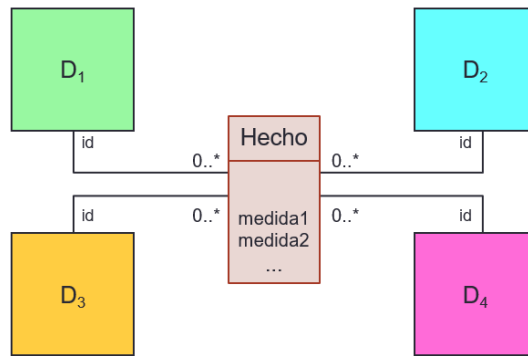


Ilustración 4 Esquema conceptual en estrella

La principal característica del esquema en estrella es que todas las propiedades de cada dimensión se incluyen en una única clase. Cuando este esquema es traducido al esquema lógico, se llega a un esquema relacional en el que hay una única tabla por cada dimensión por lo que el esquema no suele estar en 3FN. Este hecho provoca un esquema relacional con redundancia, pero esta situación es tolerable por los motivos siguientes:

- El ahorro de espacio que se generaría en caso de normalizar las tablas no es significativo.
- Las consultas (uniones y cruces entre tablas) en el esquema en estrella son simples. La normalización dificulta la complejidad de las consultas a realizar (se multiplican los JOIN).
- Las redundancias no pueden generar inconsistencias en un Almacén de Datos ya que, la información albergada es de carácter histórico. Por lo que debe cumplir la propiedad de no ser volátil.

El esquema lógico en estrella que se corresponde con el anterior esquema conceptual en estrella es el siguiente, en esta representación gráfica la clave primaria se destaca en negrita y las claves ajenas mediante flechas:

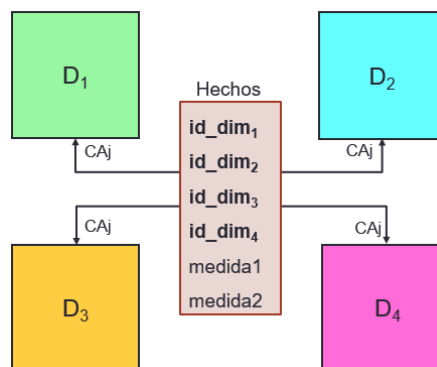


Ilustración 5 Esquema relacional en estrella

2.4.2.2 Esquema en copo de nieve

En este esquema las componentes del Almacén de Datos se conectan entre sí de forma que semejan un copo de nieve en el que:

- El hecho objeto de interés está en el centro representado por una clase débil de todas o algunas de las clases que representan las dimensiones.
- Cada dimensión está representada por un conjunto de clases conectadas entre sí.

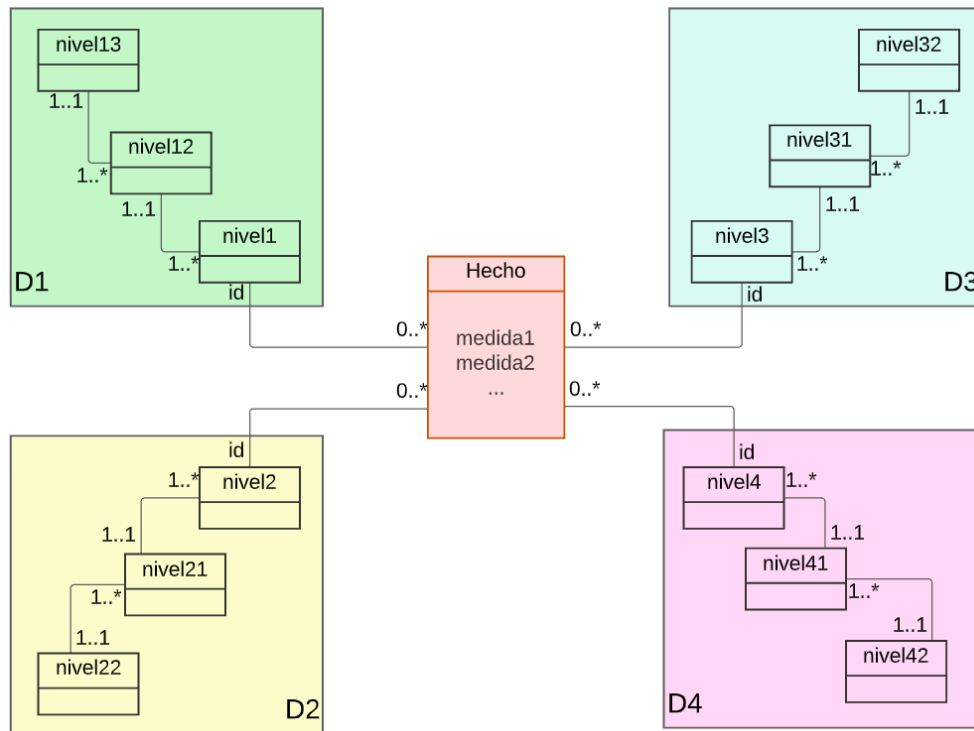


Ilustración 6 Esquema conceptual en copo de nieve

El uso de este esquema llevaría a un esquema relacional en 3FN ya que en este caso las propiedades de cada dimensión se distribuyen en distintas clases que se definen según las dependencias funcionales existentes entre los atributos que las representan, cada una de estas clases se convertirá en una tabla en el esquema relacional. Esta solución, aunque más pura y elegante desde un punto de vista relacional, ocasiona que la complejidad de las consultas aumente y que el desarrollo sea difícil tanto para el diseñador del esquema como para aquél que necesite entender la arquitectura presentada.

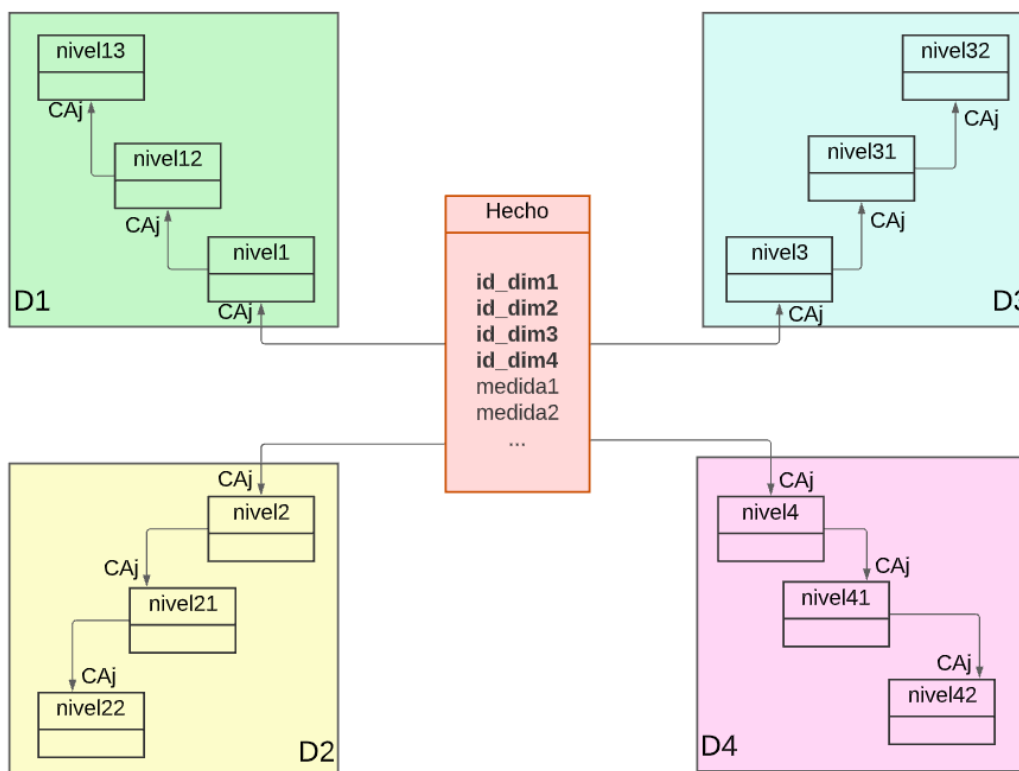


Ilustración 7 Esquema relacional en copo de nieve

2.4.3 Fase de Implementación

Una vez concluida la fase del diseño del almacén de datos, comienza la última fase que corresponde a la implementación del Almacén de Datos. En dicha fase podemos encontrar las siguientes subtarefas:

1. Cargar el banco de datos en el Almacén de Datos mediante un sistema ETL.
2. Preparar las vistas con una herramienta OLAP para que el usuario pueda interactuar con el Almacén de Datos.
3. Establecer agregación de datos.
4. Implementar informes con una herramienta OLAP.
5. El proceso ETL de un almacén de datos es uno de los más costosos por lo que en el apartado siguiente se le presta especial atención.

2.5 Herramientas ETL

En esta última sección del marco teórico se explican los conceptos fundamentales sobre este tipo de herramientas.

Las ETL tomaron gran protagonismo en la década de los setenta cuando, las grandes empresas de aquel entonces tenían la necesidad de integrar información de diferentes tipos que provenían de diferentes repositorios o bases de datos o, simplemente, garantizar que los datos,

que se cargaban en el sistema, tuvieron un grado de confianza razonable a base de someterlos a un proceso de depuración, perfilación y auditoría para una mayor calidad.

Para conseguir este propósito, las organizaciones tuvieron que depositar su confianza en el proceso ETL ya que, en aquellos momentos, se trataba de uno de los únicos métodos de integración que había. Actualmente, sigue teniendo su papel como componente central y fundamental dentro de las organizaciones ya que dichas herramientas están estrechamente implicadas en facilitar a las empresas la obtención de la ventaja competitiva frente a la toma de decisiones.

2.5.1 Fases de un proceso ETL

Este proceso consta de tres partes, como bien definen sus siglas, por su significado en inglés (*Extract, Transform y Load*) que en español se traduce a: Extracción, Transformación y Carga o Transporte. Por consiguiente, se explicarán cada una de estas fases o etapas que constituyen a las ETL.



Ilustración 8 Proceso ETL: Extracción, Transformación y Transporte de datos

1. **Extracción:** teniendo en cuenta que en esta fase se realiza la extracción de la información, el origen de datos o *source* de la arquitectura puede ser muy diverso. Sin embargo, este software no solamente es capaz de llevar a cabo este cometido, sino que también es capaz de soportar grandes volúmenes de datos heterogéneos, ya sean ficheros de tipo CSV, ficheros de tipo XSL, bases de datos tipo PostgreSQL, bases de datos Oracle, etc. Es por esta misma razón, por la que se trata de herramientas altamente calificadas para la lectura de datos de diferentes orígenes.

Hay que tener en cuenta que una extracción se puede plantear y efectuar de dos formas:

- Total: cada ejecución de extracción de datos se ejecuta en una única llamada.
 - Parcial: la extracción de datos se efectúa parcialmente a través de lotes de datos.
2. **Transformación:** una vez que los datos se han extraído de sus diferentes fuentes, será necesario aplicar un lavado de cara a los datos para garantizar su consistencia ya que, posiblemente, luego de efectuar la primera etapa de captación de datos del proceso, puede haber datos que hayan sufrido cambios o simplemente haya incompatibilidades en su estructura: caracteres inválidos, registros duplicados, valores nulos, datos errados y diversas problemáticas con las que la ingeniería de datos se enfrenta a diario.

Para poder hacer frente a las anomalías de datos antes mencionadas es necesario tener presente en cada momento, el formato de las fuentes de datos y aplicar, después de la extracción, una serie de transformaciones sobre los datos. Estas transformaciones que se plantean y se estudian en el departamento pertinente en las organizaciones, pueden consistir en:

- Filtrar filas de datos en función de una premisa.
- Eliminar valores duplicados o nulos.
- Llevar a cabo reglas de búsqueda y remplazo en función de un conjunto de caracteres.
- Unir, combinar o separar datos de diferentes fuentes.
- Cambiar el tipo de dato de las columnas de una determinada tabla.
- Calcular datos nuevos a partir de los datos de entrada mediante tecnologías de modelado aceptadas por el sistema empleado.

Todas estas transformaciones son algunas de las muchas técnicas que se pueden aplicar a los datos extraídos en bruto.

3. **Transporte:** una vez que se han depurado los datos, llega el momento de transportar esa información procesada y de carácter homogéneo. Este proceso finaliza cuando los datos llegan a su destino, es decir, cuando se lleva a cabo la carga masiva de datos. Generalmente ese destino es un Almacén de Datos o base de datos perteneciente a la organización.

Una vez concluida esta última fase del proceso ETL, los datos están listos para que, cualquier personal cualificado pueda acceder, consultar y explotar los datos a través de la elaboración de informes. Su propósito común es el de obtener conocimientos a partir de información de carácter histórico, en el caso de que el repositorio final se trate de un Almacén de Datos.

Luego de que el transporte de datos haya finalizado, es necesario aplicar una serie de procesos sobre el Almacén de Datos entre los cuales están:

- Obtención de datos agregados a partir de datos existentes en la fuente original.
- Indexación de datos mediante la creación de índices (total o parcial)

Generar estos procesos significa realizar un gran número significativo de operaciones de tal modo que se manejan grandes volúmenes de datos.

2.5.2 Funciones de un sistema ETL

Posteriormente a la creación de un Almacén de Datos, los sistemas ETL vuelven a tener protagonismo ya que las organizaciones necesitan que estos cumplan dos cometidos de gran importancia:

1. **Carga inicial:** una vez que la información de un proceso organizativo ha sido extraída correctamente de las fuentes (internas o externas de la organización) y posteriormente se ha podido llevar a cabo su limpieza mediante operaciones de transformación, se carga

dicha información histórica en el Almacén de Datos. Este último proceso puede consumir mucho tiempo.

2. **Mantenimiento periódico:** para que un Almacén de Datos pueda evolucionar con el tiempo y, sobre todo, se considere en todo momento un reflejo fiel de la organización a la que sirve, debe ser actualizado de forma periódica. Por esta misma razón se debe considerar el proceso de mantenimiento como un punto crítico que no debe de faltar en el sistema. En pocas palabras, el encargado de llevar a cabo rigurosos mantenimientos sobre un Almacén de Datos es el propio sistema ETL.

La frecuencia o con la que un sistema ETL hace mantenimiento sobre un AD se determina a partir del gránulo del Almacén de Datos y los requisitos del usuario (*load widow*²). La actualización de datos puede llegar de forma inmediata, diaria, semanal, etc. Este ciclo hay que establecerlo con cautela ya que la organización necesita que el Almacén de Datos esté disponible en todo momento para que los analistas hagan uso de él.

Centrándonos en las operaciones que ejecuta el sistema ETL para realizar el mantenimiento, se puede apreciar que dichas operaciones afectan a un volumen de datos menor en comparación a la carga inicial. Esto ocurre porque la inserción o la eliminación de datos se aplica frecuentemente sobre la tabla de hechos y pocas veces se lleva a cabo modificaciones en las dimensiones. Estos cambios, generalmente, se tratan de eliminar o archivar datos obsoletos que ya no son relevantes para el análisis. Así pues, implica la ejecución de un menor número de operaciones posteriores a la carga inicial.

Otros aspectos que hay que tener en cuenta sobre el mantenimiento de un sistema ETL son:

- El equipo encargado del desarrollo de un Almacén de Datos es responsable de la construcción del sistema ETL que se usó para la implementación de dicho AD.
- Un sistema ETL está diseñado y enfocado únicamente para cada uno de los Almacén de Datos de la organización.
- Para hacer uso de un sistema ETL se puede optar por utilizar herramientas del mercado o programas que se han implementado para un uso específico.

2.5.3 Selección de herramienta ETL

Una vez repasada la teoría de estas magnificas herramientas de integración, y luego de matizar que se tratan de un foco central e imprescindible para cualquier estrategia de análisis de datos, es necesario llevar a cabo una búsqueda de las herramientas ETL más útiles del mercado. Posteriormente, se llevará a cabo un estudio cuyo fin es el de escoger la herramienta ETL que más se ajuste a nuestras necesidades para la elaboración de este proyecto. Las posibles opciones, que se han seleccionado, son las siguientes:

- **Power BI:** dentro de todas las posibles soluciones que posee esta herramienta, tenemos un excelente motor de transformación de datos denominado Power Query.
 - Ventajas:

² Tiempo disponible para el transporte de datos y los procesos posteriores.

- Power Query está basado en el Lenguaje M, que es un lenguaje de programación que nos proporciona un gran abanico de posibilidades a la hora de llevar a cabo la limpieza y la reestructuración de datos.
- Para aquellos usuarios que hayan estado en contacto con las funciones de Excel, no tendrán dificultades al depurar los datos mediante Power Query.
- No es estrictamente necesario dominar la programación en lenguaje M para abordar todas sus funciones. La interfaz permite realizar las mismas funciones indicando en todo momento los pasos que se están realizando.
- Es accesible mediante aplicaciones móviles tanto para las plataformas de IOS como las de Android.
- Desventajas:
 - Hay dificultades en la migración de datos fuera del ecosistema de Microsoft.
- **Talend:** se trata de una solución empresarial *open source*, es decir, de código abierto, que está basada en estándares que abarcan diversas gestiones e implementaciones de integridad de datos mediante una única herramienta.
 - Ventajas:
 - Ofrece un gran volumen de opciones en la integración de datos mediante tecnologías externas, como fuente de datos, listas para usar: ADFS, Azure, Dropbox, Google Drive, etc.
 - Tiene la disponibilidad de la versión *open source* totalmente gratuita.
 - La mayoría de los elementos de la interfaz, para elaborar la transformación de datos, son fáciles de usar con funcionalidades de arrastrar y soltar.
 - La velocidad de migración de datos (transporte) es notablemente correcta para ser una herramienta gratuita.
 - Desventajas:
 - El soporte del servicio tiene dificultades para ofrecer soluciones a planteamientos complejos.
 - Pueden surgir ciertos problemas en planteamientos sobre *Big Data*.
 - Hay algunos componentes de integración que son difíciles de entender.

En cuanto a las **similitudes** de ambas herramientas seleccionadas, tenemos que:

- Ambas herramientas tienen interfaces muy intuitivas de tal forma que, cualquier persona con una noción básica de interfaces sería capaz de usarlas para implementaciones de nivel principiante.
- Ambas herramientas tienen un coste de 0€ para sus versiones básicas sin servicio en la nube.

En conclusión, ambas herramientas ETL son un claro ejemplo de soluciones propuestas para abarcar problemáticas en la toma de decisiones de una organización. Sin embargo, para la elaboración de la parte de extracción, transformación y carga de datos se seleccionará nuevamente a la herramienta de Microsoft. Los motivos son los siguientes:

- Al igual que se mencionó en la comparación anterior para la selección de una herramienta Business Intelligence, en las asignaturas relacionadas con esta materia en la UPV, se hace hincapié y se profundiza esta rama de Power BI para la presentación de los procesos ETL.
- La elaboración del Almacén de Datos propuesto resultará más sencilla ya que se utilizará una única herramienta capacitada para realizar tanto la parte de integración de datos como la de creación de informes.
- Hay mayor experiencia personal en Power Query a pesar de que, ambas herramientas han sido usadas en el ámbito laboral.

3 Plan de trabajo

Con el fin de ordenar y sintetizar las tareas que se van a aplicar en este proyecto, es necesario desglosar cada una de las fases que constituyen a este mismo. Por esta razón, en este apartado se describe cada una de estas facetas acompañadas del coste temporal necesario para su elaboración.

Es conveniente mejorar en todo lo posible la comprensión de este apartado. Así pues, se utilizará un diagrama de Gantt para ilustrar la estimación de la duración y las dependencias entre las tareas que hay que desarrollar.

3.1 Planificación de las fases

Para estimar el tiempo que se va a dedicar a cada una de las fases del proyecto, es necesario destacar que dichas fases se repartirán en jornadas de 8 horas. Una vez aclarado este matiz, podemos apreciar que el proyecto está compuesto por las siguientes fases:

1. Elección de la fuente de datos: búsqueda de información sobre los efectos de la pandemia en los portales web, que ofrezcan fuentes de datos gratuitas. Se estima que estas tareas puedan abarcar 10 jornadas.
2. Diseño Conceptual Multidimensional: obtención del diagrama de clases que represente adecuadamente el sistema de información. Se estima que estas tareas puedan abarcar 7 jornadas.
3. Diseño Lógico Relacional Multidimensional (ROLAP): traducción del esquema conceptual a un esquema relacional. Se estima que estas tareas puedan abarcar 6 jornadas.
4. Extracción, Transformación y Carga de datos (ETL): a partir de los datos obtenidos realizar las tareas necesarias para cargar las tablas definidas en la fase anterior. Se estima que estas tareas puedan abarcar 9 jornadas.
5. Explotación: diseño de informes. Se estima que estas tareas puedan abarcar 3 jornadas.

A continuación, se presenta el diagrama de Gantt en donde se muestran los días en los que se han llevado a cabo las tareas. Además de la fecha de inicio y finalización de cada una de ellas, se especifica el número de jornadas que han ocupado dichas tareas, donde el valor "1" significa una jornada completa, es decir, 8 horas y el valor "½" alude a media jornada, es decir, 4 horas.

ALMACÉN DE DATOS – IMPACTO COVID-19	ELECCIÓN DE LA FUENTE DE DATOS	DISEÑO CONCEPTUAL MULTIDIMENSIONAL	DISEÑO LÓGICO RELACIONAL MULTIDIMENSIONAL	EXTRACCIÓN, TRANSFORMACIÓN Y CARGA DE DATOS	EXPLOTACIÓN
INICIO	01/10/21	30/10/21	12/12/21	08/01/22	29/01/22
FINALIZACIÓN	21/11/21	11/12/21	18/12/21	20/02/22	06/03/22
JORNADAS	10	7	6	9	3
ESTADO	Terminado	Terminado	Terminado	Terminado	Terminado
01/10/21	1				
02/10/21	1				
16/10/21	1				
20/10/21	1				
23/10/21	1				
30/10/21	½	½			
06/11/21	½	½			
13/11/21	1				
14/11/21	1				
18/11/21	1				
20/11/21	½	½			
21/11/21	½	½			
27/11/21		1			
28/11/21		1			
04/12/21		1			
05/12/21		1			
11/12/21		1			
12/12/21			1		
13/12/21			1		
15/12/21			1		
16/12/21			1		
17/12/21			1		
18/12/21			1		
08/01/22				1	
09/01/22				1	
15/01/22				1	
16/01/22				1	
22/01/22				1	
23/01/22				1	
29/01/22				½	½
12/02/22				½	½
19/02/22				1	
20/02/22				1	
05/03/22					1
06/03/22					1

Tabla 1 Diagrama de Gantt: Planificación de las fases y estimación temporal

3.2 Presupuesto

Una vez que se han obtenido las estimaciones del coste temporal mediante las jornadas totales gastadas, es necesario calcular el coste económico que se debe invertir para la elaboración de dicho proyecto en un contexto laboral. Para obtener este coste, se toma como referencia el sueldo medio de un desarrollador de datos procedente de España.


Software	Descripción	Coste (€)
	Conjunto de herramientas empresariales <i>open source</i> que están basadas en estándares que abarcan diversas gestiones de integridad de datos.	0,0€
Personal necesario	Número de jornadas	Coste promedio (€)
1	35	1.920,35€
Coste total		1.920,35€

Tabla 2 Costes económicos del proyecto. Recuperado de <https://es.indeed.com/career/salaries/sql?from=whatwhere>

4 Análisis del problema

4.1 Temática seleccionada

La idea de realizar este proyecto surge debido a la aparición de un virus, originario en Wuhan, a finales del año 2019, llamado SARS-CoV-2 o mejor conocido como COVID-19, que ocasionó que el mundo se sumergiera en una de las peores pandemias que ha podido experimentar el ser humano a lo largo de la historia. Esta nueva gripe, que actúa de forma más agresiva de lo habitual, fue declarada por la OMS como una emergencia de salud pública a nivel internacional ya que puso en jaque a todos los sistemas públicos del mundo. Debido a la facilidad con la que el virus se propaga, ocasionó que, en tiempo récord, hubiera miles de contagios y fallecimientos, hospitales colapsados, supermercados desabastecidos y economías al borde del abismo.

Hoy en día, este brote sigue ocasionando estragos, aunque en menor medida que antes. Es por ello por lo que esta situación ha desencadenado el interés por conocer cuál ha sido la evolución del virus a nivel mundial, desde sus inicios, mediante un Almacén de Datos.

4.2 Origen del banco de datos

Con respecto a la búsqueda de un banco de datos público enfocado al tema de la pandemia mundial, se investigaron distintas fuentes de datos de diferentes portales webs:

- **World Health Organization (WHO):** organismo de la ONU para establecer políticas de prevención, promoción e intervención a nivel mundial de la salud.
- **Kaggle:** comunidad de científicos de datos más conocida en Estados Unidos, subsidiaria de Google LLC.
- **Our World in Data:** plataforma online desarrollada por la Universidad de Oxford para presentar cambios en las condiciones de vida de todo el mundo a través de datos y resultados empíricos.

Finalmente, dentro de la última opción mencionada, se localizó una fuente de datos lo suficientemente interesante, como para poder empezar a dar forma al Almacén de Datos enfocado al COVID-19.

Our World in Data

Articles by topic Search... Latest About Donate

Statistics and Research

Coronavirus Pandemic (COVID-19)

Research and data: Hannah Ritchie, Edouard Mathieu, Lucas Rod s-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie MacDonald, Diana Beltekian, Saloni Dattani and Max Roser
 Web development: Lars Yencken, Daniel Bachler, Ernst van Woerden, Daniel Gavrilo, Marcel Gerber, Matthieu Bergel, and Jason Crawford

The data on the coronavirus pandemic is updated daily. Last update: an hour ago. Reuse our work freely Cite this research

Coronavirus > By country Data explorer Deaths Cases Tests Hospitalizations Vaccinations Mortality risk Excess mortality Policy responses Exemplars

- Data Explorer**: Explore all metrics – including cases, deaths, testing, and vaccinations – in one place.
- Country Profiles**: Get an overview of the pandemic for any country on a single page.
- Download Dataset**: Download our complete dataset of COVID-19 metrics on GitHub. It's open access and free for anyone to use.
- Vaccinations**: Explore our global dataset on COVID-19 vaccinations.
- US Vaccinations**: See state-by-state data on vaccinations in the United States.
- Cases**: Explore the data on confirmed COVID-19 cases for all countries.
- Deaths**: Explore the data on confirmed COVID-19 deaths for all countries.
- Testing**: Explore our data on COVID-19 testing to see how confirmed cases compare to actual infections.
- Hospitalizations**: See data on how many people are being hospitalized for COVID-19.
- Policy Responses**: See how government policy responses – on travel, testing, vaccinations, face coverings, and more – vary across the world.
- Mortality Risk**: Learn what we know about the mortality risk of COVID-19 and explore the data used to calculate it.
- Excess Mortality**: Compare the number of deaths from all causes during COVID-19 to the years before to gauge the total impact of the pandemic on deaths.

Ilustraci3n 9 Portal web ourworldindata.org. Recuperado de <https://ourworldindata.org/coronavirus>

El fichero de datos resultante (en formato CSV) con el cual se va a trabajar, se ha recuperado a trav s de un enlace web del mismo portal de *Our World in Data*. Este enlace nos dirige a un repositorio de GitHub (plataforma de desarrollo colaborativo para alojar proyectos m s conocida de Estados Unidos) en donde, aparte de estar el *dataset*, nos encontramos su documentaci3n con su descripci3n.

Why GitHub? Team Enterprise Explore Marketplace Pricing

Search Sign in Sign up

owid / covid-19-data Public Sponsor Notifications Star 4.6k Fork 2.9k

<> Code Issues 6 Pull requests 1 Discussions Actions Security Insights

master 5 branches 0 tags Go to file Code

edomt data(vax): update 1ea6974 2 hours ago 11,635 commits

- .github chore(ci): go back to using the official Netlify GH action 2 days ago
- public data(vax): update 2 hours ago
- scripts data(vax): update 2 hours ago
- .gitignore data(tests): update 26 days ago
- README.md docs(readme): fix broken links 3 months ago

README.md

COVID-19 Dataset by Our World in Data

About: Data on COVID-19 (coronavirus) cases, deaths, hospitalizations, tests • All countries • Updated daily by Our World in Data

ourworldindata.org/coronavirus

coronavirus covid-19 covid sars-cov-2

Sponsor this project <https://ourworldindata.org/donate>

Ilustraci3n 10 Portal web github.com, origen del banco de datos COVID-19 Dataset. Recuperado de <https://github.com/owid/covid-19-data>.

Este banco de datos alberga más de ciento treinta y dos mil entradas de datos (filas) que incluyen todos los datos históricos sobre la pandemia desde la fecha del 1 de enero del año 2020, hasta la fecha de la última descarga del datase o hasta la fecha de la última vez que se actualizaron los datos desde la herramienta Power BI. Dicho banco de datos recibe nuevas actualizaciones de datos diarias ya que los casos sufridos por COVID-19 siguen continuando hasta nuestras fechas.

En este fichero, se puede encontrar información de los países más afectados por el COVID-19, el número de personas que fueron ingresados por esta enfermedad en unidades de cuidados intensivos, la fecha en la que se inició el brote en cada territorio geográfico, etc.

5 Diseño y desarrollo de la solución

En este apartado se irán analizando y documentando cada una de las acciones a realizar en cada una de las fases del diseño del Almacén de Datos, hasta alcanzar la solución planteada.

5.1 Diseño conceptual

Con el objetivo de conseguir realizar el esbozo del modelo multidimensional, en esta fase, se ha llevado a cabo un análisis del banco de datos procedente de la página web *Our World in Data*, revisando cada una de las columnas del fichero. El objetivo, desde un principio, es poder depurar dicha fuente. Para ello, no se han tenido en cuenta algunas columnas que carecían de valor o cuyo significado no contribuye a nuestro Almacén de Datos. En la fase final del proceso (etapa ETL) del diseño del Almacén de Datos, estas columnas serán eliminadas del modelo final.

Después de analizar la información obtenida, se ha detectado que el fichero tiene información almacenada a dos niveles, diario y semanal; aunque podría haberse diseñado el AD para que almacenara toda esta información en un mismo repositorio, este hecho hubiera dificultado después la manipulación de los datos ya que en todo momento habría habido que distinguir con qué información, diaria o semanal, se deseaba trabajar. Debido a esto se ha optado por diseñar dos almacenes de datos que comparten una dimensión, de esta forma es mucho más sencilla la realización de los informes.

Teniendo esto en cuenta:

- **AD diario:** se establecer el hecho principal IMPACTO_DIARIO, como objeto de interés principal. Una vez que se establecido el hecho con sus respectivas medidas o atributos de hechos, se han sacado las dimensiones ZONA_GEOGRAFICA y TIEMPO_DIARIO con sus respectivos atributos dimensionales. Cabe destacar que cada casuística perteneciente al hecho principal corresponde a datos asociados al impacto sufrido por dicha enfermedad. Estos datos fueron reportados diariamente en una ubicación geográfica determinada.
- **AD semanal:** se establecer el hecho principal IMPACTO_SEMANAL, como objeto de interés principal. Este hecho aparte de agrupar casuísticas a nivel de semanas y no de días, nos ofrece información de los ingresos que hubo en los hospitales de una zona determinada a causa de la enfermedad. La dimensión ZONA_GEOGRAFICA es similar al caso diario, por lo que será compartida. Los atributos de la dimensión temporal cambiarán ya que estarán representando casos a nivel de semana, por esta razón, dicha dimensión tendrá el nombre de TIEMPO_SEMANAL.

Una vez clasificada la información mediante dos esquemas con matices temporales diferentes es necesario integrar los datos en una única arquitectura con dos hechos. Para ello, se ha esbozado el esquema final a través de un diagrama de clases de UML que se conectan través de la dimensión ZONA_GEOGRAFICA ya que no hubo ninguna diferencia, en esta dimensión, en ninguno de los dos casos. El esquema conceptual resultante es el siguiente:

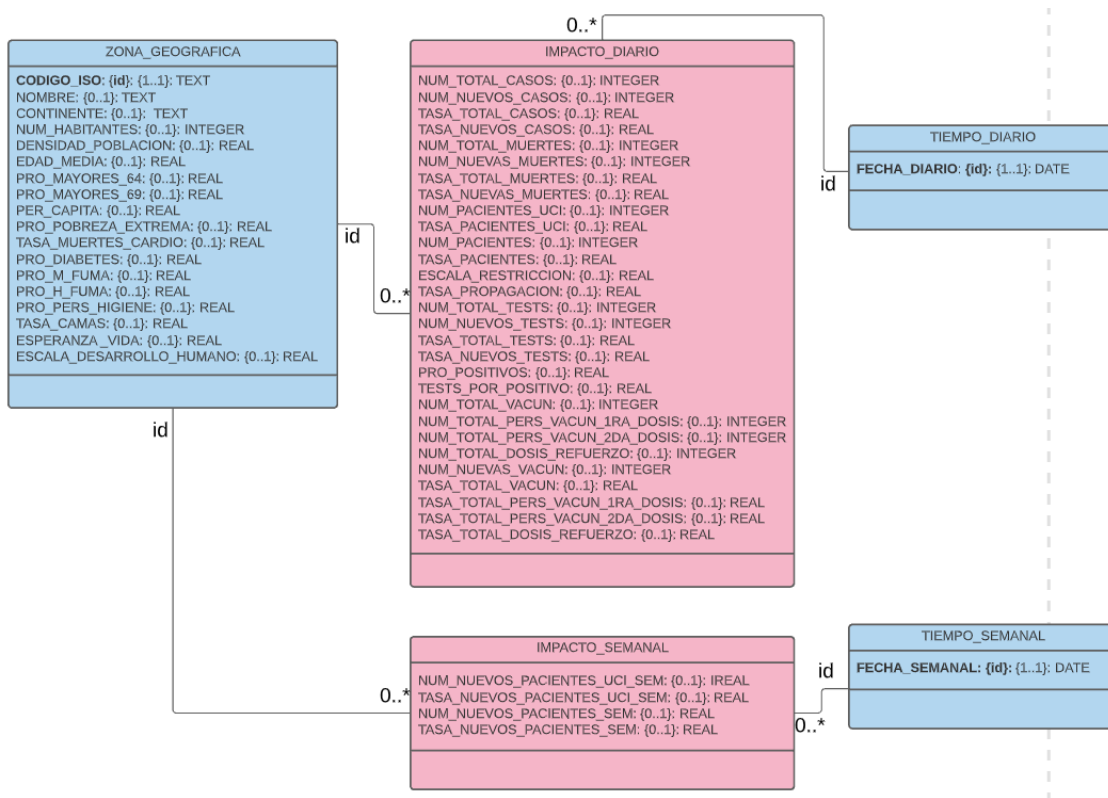


Ilustración 11 Esquema conceptual representado mediante un diagrama UML

En dicho diagrama de clases, se puede observar que, tanto el primer hecho IMPACTO_DIARIO como el segundo hecho IMPACTO_SEMANAL se representan mediante clases débiles a través de la cardinalidad “id” entre la clase del hecho y las clases de las dimensiones. La razón por la cual se establece estas cardinalidades es para que cada una de las casuísticas, pertenecientes a los hechos, puedan identificarse mediante la fecha, cuando se produjo el impacto, y la ubicación geográfica en donde ocurrió el censo causado por el COVID-19.

Otro aspecto relevante del esquema es que, no se ha establecido una dimensión HOSPITAL, para los atributos con la abreviatura “PACIENTES”. Esto es debido a que se tratan de valores numéricos, que varían en función del tiempo y de la zona geográfica en donde se sufrió el impacto. Es por esta razón, que deben estar situados como medidas en los hechos principales. Si se hubiesen tratado de atributos que no dependen ni del tiempo ni de la zona geográfica, hubiese sido de mayor interés definir dicha dimensión.

A continuación, se mostrarán y se explicarán cada una de las clases del modelo conceptual:

Clase	Descripción
IMPACTO_DIARIO	Primer hecho que relaciona la información temporal y la zona geográfica en donde ocurre el caso COVID-19 en un día determinado.
IMPACTO_SEMANAL	Segundo hecho que relaciona la información temporal y la zona geográfica en donde ocurre el caso COVID-19 cada semana.

Clase	Descripción
TIEMPO_DIARIO	Dimensión temporal con la fecha en la que se produce el caso COVID-19 con carácter diario.
TIEMPO_SEMANAL	Dimensión temporal con la fecha en la que se produce el caso COVID-19 con carácter semanal.
ZONA_GEOGRAFICA	Dimensión geográfica que contiene la ubicación territorial en donde ocurre el caso COVID-19.

Tabla 3 Descripción de cada una de las clases del modelo

Por cada clase definida, se va a explicar cada uno de sus atributos. Cabe mencionar que se va adjuntar tanto el nombre original de cada una de las variables, como el nombre que toma en el modelo resultante a implementar.

Para las dimensiones temporales:

TIEMPO_DIARIO	
Atributo	Descripción
<i>date</i> FECHA_DIARIO	Día en el que se produjo la observación.

Tabla 4 Dimensión TIEMPO_DIARIO con las descripciones de sus atributos.

TIEMPO_SEMANAL	
Atributo	Descripción
<i>date</i> FECHA_SEMANAL	Primer día de la semana en la que se produjo la observación.

Tabla 5 Dimensión TIEMPO_SEMANAL con las descripciones de sus atributos

Para la dimensión geográfica, tenemos:

ZONA GEOGRÁFICA	
Atributo	Descripción
<i>iso_code</i> CODIGO_ISO	ISO 3166-1 alpha-3 – código de tres letras identificativo de cada país.

ZONA GEOGRÁFICA	
Atributo	Descripción
<i>location</i> NOMBRE	Nombre de la ubicación geográfica.
<i>continent</i> CONTINENTE	Continente de la ubicación geográfica.
<i>population</i> NUM_HABITANTES	Número de habitantes de la ubicación geográfica (últimos valores disponibles).
<i>population_density</i> DENSIDAD_POBLACION	Número de personas dividido por área (medido en kilómetros cuadrados) de la zona geográfica.
<i>median_age</i> EDAD_MEDIA	Edad media de la población, estimación de la ONU para 2020.
<i>aged_65_older</i> PRO_MAYORES_64	Proporción de la población con 65 años o más. (%)
<i>aged_70_older</i> PRO_MAYORES_69	Proporción de la población con 70 años o más en 2015. (%)
<i>gdp_per_capita</i> PER_CAPITA	Indicador económico que mide la relación existente entre el nivel de renta de un país y su población. (\$)
<i>extreme_poverty</i> PRO_POBREZA_EXTREMA	Proporción de la población que vive en pobreza extrema. Datos del 2010. (%)
<i>cardiovasc_death_rate</i> TASA_MUERTES_CARDIO	Tasa de mortalidad por enfermedad cardiovascular en 2017 (número anual de muertes por cada 100K habitantes).
<i>diabetes_prevalence</i> PRO_DIABETES	Proporción de la población que sufre de diabetes (20 a 79 años) en 2017. (%)
<i>female_smokers</i> PRO_M_FUMA	Proporción de mujeres que fuman. (%)
<i>male_smokers</i> PRO_H_FUMA	Proporción de hombres que fuman. (%)

ZONA GEOGRÁFICA	
Atributo	Descripción
<i>handwashing_facilities</i> PRO_PERS_HIGIENE	Proporción de la población con instalaciones básicas para lavarse las manos. (%)
<i>hospital_beds_per_thousand</i> TASA_CAMAS	Número de camas de hospital por cada 1K habitantes.
<i>life_expectancy</i> ESPERANZA_VIDA	Media de años de vida de los habitantes de la zona geográfica durante el 2019.
<i>human_development_index</i> ESCALA_DESARROLLO_HUMANO	Índice compuesto por tres factores básicos que miden el desarrollo humano: una vida larga y saludable, conocimientos y un nivel de vida decente (intervalo de 0,0 hasta 1,0) ³ .

Tabla 6 Dimensión ZONA_GEOGRAFICA con las descripciones de sus atributos

Para los hechos principales, tenemos:

IMPACTO_DIARIO	
Variables	Descripciones
<i>Total_cases</i> NUM_TOTAL_CASOS	Número de casos totales confirmados de COVID-19 hasta ese día.
<i>New_cases</i> NUM_NUEVOS_CASOS	Número de casos nuevos confirmados de COVID-19 diariamente.
<i>total_cases_per_million</i> TASA_TOTAL_CASOS	Número de casos totales confirmados de COVID-19 hasta ese día por cada 1M de habitantes.
<i>New_cases_per_million</i> TASA_NUEVOS_CASOS	Número de nuevos casos confirmados de COVID-19 diariamente por cada 1M de habitantes.

³ Valores superiores estrictamente a 0,799 significan que hay un desarrollo humano sobresaliente. Valores que van desde 0,700 hasta 0,799 significan que hay un desarrollo humano alto. Valores que van desde 0,550 hasta 0,699 significan que hay un desarrollo humano medio. Valores inferiores estrictamente a 0,550 significan que el desarrollo humano es bajo.

IMPACTO_DIARIO	
Variables	Descripciones
<i>total_deaths</i> NUM_TOTAL_MUERTES	Número de muertes totales confirmadas de COVID-19 hasta ese día.
<i>New_deaths</i> NUM_NUEVAS_MUERTES	Número de muertes nuevas confirmadas de COVID-19 diariamente.
<i>total_deaths_per_million</i> TASA_TOTAL_MUERTES	Número total de muertes confirmadas de COVID-19 hasta ese día por cada 1M de habitantes.
<i>new_deaths_per_million</i> TASA_NUEVAS_MUERTES	Número de muertes nuevas confirmadas de COVID-19 diariamente por cada 1M de habitantes.
<i>icu_patients</i> NUM_PACIENTES_UCI	Número de pacientes con COVID-19 en unidades de cuidados intensivos (UCI) diariamente.
<i>icu_patients_per_million</i> TASA_PACIENTES_UCI	Número de pacientes con COVID-19 en unidades de cuidados intensivos (UCI) diariamente por cada 1M de habitantes.
<i>hosp_patients</i> NUM_PACIENTES	Número de pacientes con COVID-19 en el hospital diariamente.
<i>hosp_patients_per_million</i> TASA_PACIENTES	Número de pacientes con COVID-19 en el hospital diariamente por cada 1M de habitantes.
<i>stringency_index</i> ESCALA_RESTRICCION	Índice de rigurosidad de respuesta tomado por el gobierno diariamente. Medida basada en nueve indicadores de respuesta que incluyen cierres de escuelas, cierres de lugares de trabajo y prohibiciones de viaje (intervalo de 0 hasta 100) ⁴ .

⁴ Valores superiores estrictamente a 74,99 denotan un nivel de medidas de restricción extremo. Valores que van desde 50,0 hasta 74,99 denotan un nivel de restricción alto. Valores que van desde 25,0 hasta 49,99 denotan que hubo un nivel de restricción medio. Valores inferiores estrictamente a 24,99 denotan pocas medidas de restricción.

IMPACTO_DIARIO	
Variables	Descripciones
<i>reproduction_rate</i> TASA_PROPAGACION	Estimación en tiempo real de la tasa de reproducción efectiva (R) de COVID-19 diariamente. ⁵
<i>total_tests</i> NUM_TOTAL_TESTS	Número de tests totales para COVID-19 hasta ese día.
<i>new_tests</i> NUM_NUEVOS_TESTS	Número de nuevos tests para COVID-19 diariamente (sólo calculados para días consecutivos).
<i>total_tests_per_thousand</i> TASA_TOTAL_TESTS	Número de tests totales para COVID-19 hasta ese día por cada 1K habitantes.
<i>new_tests_per_thousand</i> TASA_NUEVOS_TESTS	Número de nuevos tests para COVID-19 diariamente por cada 1K habitantes.
<i>positive_rate</i> PRO_POSITIVOS	Proporción de tests que han dado positivo en COVID-19 diariamente. (La inversa de TESTS_POR_POSITIVO).
<i>test_per_case</i> TESTS_POR_POSITIVO	Número de tests realizados por cada nuevo caso confirmado de COVID-19 diariamente. (La inversa de PRO_POSITIVOS).
<i>total_vaccinations</i> NUM_TOTAL_VACUN	Número total de las dosis la de vacuna COVID-19 administradas hasta ese día.
<i>people_vaccinated</i> NUM_TOTAL_PERS_VACUN_1RA_DOSIS	Número total de personas que recibieron al menos una dosis de vacuna hasta ese día.
<i>people_fully_vaccinated</i> NUM_TOTAL_PERS_VACUN_2DA_DOSIS	Número total de personas que recibieron todas las dosis prescritas por el protocolo de vacunación hasta ese día.

⁵ Si R es menor estrictamente que 1, la infección no causará nuevos contagios. Por consiguiente, la enfermedad disminuirá y eventualmente, desaparecerá. Si R es igual que 1, cada infección existente provocará nuevos contagios. Por consiguiente, la enfermedad se mantendrá viva y estable pero no causará ningún brote ni ninguna epidemia. Si R es superior estrictamente que 1, cada infección existente provocará nuevos contagios. Por consiguiente, la enfermedad se transmitirá rápidamente entre los habitantes, provocando un brote o epidemia.

IMPACTO_DIARIO	
Variables	Descripciones
<i>total_boosters</i> NUM_TOTAL_DOSIS_REFUERZO	Número total de dosis de refuerzo de la vacuna COVID-19 administradas hasta ese día (dosis administradas más allá del número prescrito por el protocolo de vacunación)
<i>new_vaccinations</i> NUM_NUEVAS_VACUN	Número de nuevas dosis de vacuna COVID-19 administradas diariamente (sólo calculadas para días consecutivos).
<i>total_vaccinations_per_hundred</i> TASA_TOTAL_VACUN	Número total de dosis de vacuna COVID-19 administradas por cada 100 habitantes hasta ese día.
<i>people_vaccinated_per_hundred</i> TASA_TOTAL_PERS_VACUN_1RA_DOSIS	Número total de personas que recibieron al menos una dosis de vacuna por cada 100 habitantes hasta ese día.
<i>people_fully_vaccinated_per_hundred</i> TASA_TOTAL_PERS_VACUN_2DA_DOSIS	Número total de personas que recibieron todas las dosis prescritas por el protocolo de vacunación por cada 100 habitantes hasta ese día.
<i>total_boosters_per_hundred</i> TASA_TOTAL_DOSIS_REFUERZO	Número total de dosis de refuerzo de la vacuna COVID-19 administradas hasta ese día por cada 100 habitantes.

Tabla 7 Hecho principal IMPACTO_DIARIO con las descripciones de sus atributos

IMPACTO_SEMANAL	
Atributo	Descripción
<i>weekly_icu_admissions</i> NUM_NUEVOS_PACIENTES_UCI_SEM	Número de pacientes COVID-19 recién admitidos en unidades de cuidados intensivos (UCI) semanalmente.
<i>weekly_icu_admissions_per_million</i> TASA_NUEVOS_PACIENTES_UCI_SEM	Número de pacientes COVID-19 recién admitidos en unidades de cuidados intensivos (UCI) semanalmente por cada 1M de habitantes.
<i>weekly_hosp_admissions</i> NUM_NUEVOS_PACIENTES_SEM	Número de pacientes con COVID-19 recién admitidos en hospitales semanalmente.

IMPACTO_SEMANAL	
Atributo	Descripción
<i>weekly_hosp_admissions_per_million</i> TASA_NUEVOS_PACIENTES_SEM	Número de pacientes con COVID-19 recién admitidos en hospitales semanalmente por cada 1M de habitantes.

Tabla 8 Hecho principal IMPACTO_SEMANAL con las descripciones de sus atributos

5.2 Diseño lógico

Una vez presentado el modelo multidimensional a través de un modelo en estrella. Se ha transformado el modelo, esbozado mediante un diagrama UML, a un diagrama relacional. Dicho modelo relacional quedaría de esta forma:

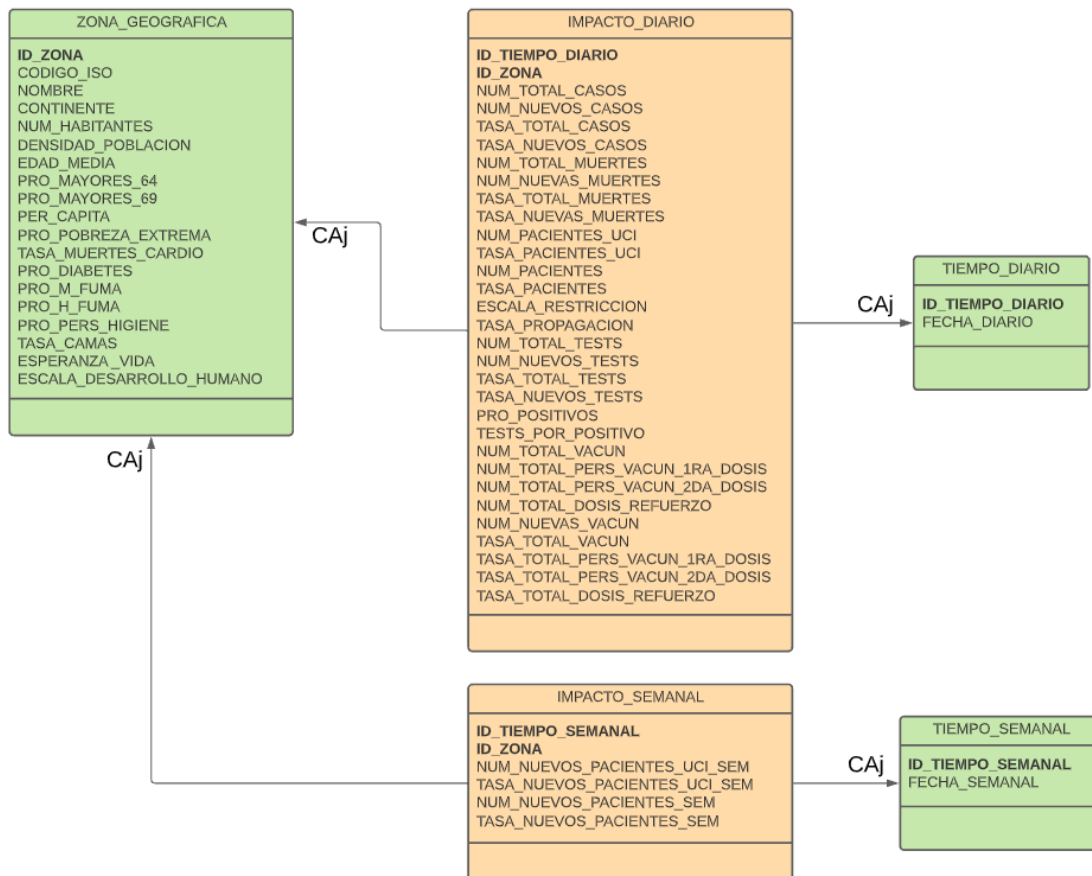


Ilustración 12 Esquema relacional

Este esquema gráfico relacional se corresponde con el siguiente esquema lógico:

IMPACTO_DIARIO (ID_TIEMPO_DIARIO: INTEGER, ID_ZONA: INTEGER, NUM_TOTAL_CASOS: INTEGER,)

```

NUM_NUEVOS_CASOS: INTEGER,
TASA_TOTAL_CASOS: REAL,
TASA_NUEVOS_CASOS: REAL,
NUM_TOTAL_MUERTES: INTEGER,
NUM_NUEVAS_MUERTES: INTEGER,
TASA_TOTAL_MUERTES: REAL,
TASA_NUEVAS_MUERTES: REAL,
NUM_PACIENTES_UCI: INTEGER,
TASA_PACIENTES_UCI: REAL,
NUM_PACIENTES: INTEGER,
TASA_PACIENTES: REAL,
ESCALA_RESTRICCION: REAL,
TASA_PROPAGACION: REAL,
NUM_TOTAL_TESTS: INTEGER,
NUM_NUEVOS_TESTS: INTEGER,
TASA_TOTAL_TESTS: REAL,
TASA_NUEVOS_TESTS: REAL,
PRO_POSITIVOS: REAL,
TESTS_POR_POSITIVO: REAL,
NUM_TOTAL_VACUN: INTEGER,
NUM_TOTAL_PERS_VACUN_1RA_DOSIS: INTEGER,
NUM_TOTAL_PERS_VACUN_2DA_DOSIS: INTEGER,
NUM_TOTAL_DOSIS_REFUERZO: INTEGER,
NUM_NUEVAS_VACUN: INTEGER,
TASA_TOTAL_VACUN: REAL,
TASA_TOTAL_PERS_VACUN_1RA_DOSIS: REAL,
TASA_TOTAL_PERS_VACUN_2DA_DOSIS: REAL,
TASA_TOTAL_DOSIS_REFUERZO: REAL
)
    CP: {ID_TIEMPO_DIARIO, ID_ZONA}
    Caj: {ID_TIEMPO_DIARIO} => TIEMPO_DIARIO
    Caj: {ID_ZONA} => ZONA_GEOGRAFICA

```

IMPACTO_SEMANAL (

```

ID_TIEMPO_SEMANAL: INTEGER,
ID_ZONA: INTEGER,
NUM_NUEVOS_PACIENTES_UCI_SEM: REAL,
TASA_NUEVOS_PACIENTES_UCI_SEM: REAL,
NUM_NUEVOS_PACIENTES_SEM: REAL,
TASA_NUEVOS_PACIENTES_SEM: REAL
)
    CP: {ID_TIEMPO_SEMANAL, ID_ZONA}
    Caj: {ID_TIEMPO_SEMANAL} => TIEMPO_SEMANAL
    Caj: {ID_ZONA} => ZONA_GEOGRAFICA

```

```

TIEMPO_DIARIO (
  ID_TIEMPO_DIARIO: INTEGER,
  FECHA_DIARIO: DATE
)
  CP: {ID_TIEMPO_DIARIO}
  ÚNICO: {FECHA_DIARIO}
  VNN: {FECHA_DIARIO}

TIEMPO_SEMANAL (
  ID_TIEMPO_SEMANAL: INTEGER,
  FECHA_SEMANAL: DATE
)
  CP: {ID_TIEMPO_SEMANAL}
  ÚNICO: {FECHA_SEMANAL}
  VNN: {FECHA_SEMANAL}

ZONA_GEOGRAFICA (
  ID_ZONA: INTEGER,
  CODIGO_ISO: TEXT,
  NOMBRE: TEXT,
  CONTINENTE: TEXT,
  NUM_HABITANTES: INTEGER,
  DENSIDAD_POBLACION: REAL,
  EDAD_MEDIA: REAL,
  PRO_MAYORES_64: REAL,
  PRO_MAYORES_69: REAL,
  PER_CAPITA: REAL,
  PRO_POBREZA_EXTREMA: REAL,
  TASA_MUERTES_CARDIO: REAL,
  PRO_DIABETES: REAL,
  PRO_M_FUMA: REAL,
  PRO_H_FUMA: REAL,
  PRO_PERS_HIGIENE: REAL,
  TASA_CAMAS: REAL,
  ESPERANZA_VIDA: REAL,
  ESCALA_DESARROLLO_HUMANO: REAL
)
  CP: {ID_ZONA}
  ÚNICO: {CODIGO_ISO}
  VNN: {CODIGO_ISO}

```

Se puede observar en el esquema lógico que, para cada dimensión se ha establecido un identificador artificial de tipo entero, tanto para la dimensión ZONA_GEOGRAFICA (ID_ZONA) como para las dimensiones temporales TIEMPO_DIARIO (ID_TIEMPO_DIARIO) y

TIEMPO_SEMANAL (ID_TIEMPO_SEMANAL). Cada uno de estos identificadores están asociados con su respectivo clave del sistema operacional. En el caso de ID_TIEMPO_DIARIO e ID_TIEMPO_SEMANAL hacen referencia a la FECHA_DIARIO y FECHA_SEMANAL respectivamente, del mismo modo ocurre con ID_PAIS que se atribuye a CODIGO_ISO.

Los motivos que justifican la sustitución de las claves del sistema operacional de las dimensiones por unas claves de tipo entero autoincremental (claves sin significado), son las siguientes:

- Las claves de tipo entero generadas artificialmente (4 bytes de tamaño) son lo suficientemente grandes como para poder almacenar los datos de las dimensiones de cualquier tipo de tamaño (2^{32} valores distintos).
- Se evita el aumento excesivo del tamaño de las tablas de hechos.
- Se evitan futuros errores de inconsistencia o esfuerzos innecesarios a la hora de modificar las claves primarias del sistema operacional.

5.3 Diseño físico

Centrándonos en esta última faceta del diseño, podemos sacar la conclusión de que no es necesario aplicar ningún tipo de estrategia para agilizar las consultas que se realicen sobre el Almacén de Datos a construir. Esto es debido a que el planteamiento de un modelo multidimensional simplifica de por sí la estructuración de los datos y al mismo tiempo reduce las tareas que hay que aplicar en el diseño físico. Esta simplificación de datos se puede apreciar en la reducción del número de tablas y relaciones necesarias para ejecutar dichas consultas.

A su vez, tanto DAX como Power Query, que son funcionalidades que se encuentran nativamente en Power BI, presentan optimizaciones para este tipo de casos. Especialmente si se trata de un modelo que utilice el esquema en forma de estrella.

6 Implementación de la solución

Una vez terminada la fase de diseño del Almacén de Datos, es necesario llevar a cabo la construcción de dicho almacén mediante el uso de la herramienta Power BI Desktop de Microsoft. Específicamente, se hace uso de su editor *Power Query*, siendo una de las muchas funcionalidades que tiene este software.

Este editor permite ejecutar *scripts* o consultas de manipulación de datos siguiendo un orden secuencial de cada una de las operaciones realizadas. Por esta misma razón, nos adentramos en la última fase del diseño, el procedimiento ETL.

En este apartado se ven y se justifican cada una de las operaciones realizadas (extracción, transformación y carga) dentro de las consultas encargadas de formar y devolver las tablas que constituyen el modelo multidimensional planteado.

Una vez que se accede al editor *Power Query* del proyecto nos encontramos diversas tareas, éstas son: un parámetro y ocho consultas organizadas en carpetas.

En dichas carpetas, se pueden apreciar consultas tanto con la carga de datos habilitada (transportan las tablas dimensionales al modelo final multidimensional) como las deshabilitadas, que simplemente son subconsultas que están dentro de la ecuación para poder construir las tablas dimensionales a partir de las consultas con la carga habilitada.

Primeramente, tenemos el parámetro RUTA dentro de la carpeta PARAMETROS. En este, como bien dice su nombre, se especifica la ruta en donde se hospedarán dos ficheros de tipo CSV, ambos en el mismo directorio:

- Por un lado, el banco de datos denominado “owid-covid-data.csv”, en el caso de que la carga de datos se haga de forma estática, es decir, descargando el último fichero de la plataforma GitHub.
- Por otro lado, está el fichero de datos denominado “TRADUCCION.csv”, que contiene todos los nombres de las zonas geográficas y de los continentes, tanto en el idioma original del banco de datos, es decir, en inglés como el de su respectiva traducción al español.

Seguidamente, tenemos las consultas de las dimensiones (con la carga de datos habilitada) Su objetivo como *scripts* es el de formar las tablas dimensionales correspondientes al tiempo y al lugar del modelo final. Dichas consultas son: TIEMPO_DIARIO, TIEMPO_SEMANAL y ZONA_GEOGRAFICA. Todas ellas se encuentran dentro de la carpeta DIMENSIONES.

Por otro lado, tenemos las consultas de los hechos (con la carga de datos habilitada) Su objetivo como *scripts* es el de formar las tablas constituyentes a los objetos de estudio del modelo final. Dichas consultas son: IMPACTO_DIARIO e IMPACTO_SEMANAL. Todas ellas se encuentran dentro de la carpeta HECHOS.

Finalmente, tenemos las subconsultas (con la carga deshabilitada): TRADUCCION, BANCO_ESTATICO y BANCO_DINAMICO. Dichas subconsultas se ubican dentro de la carpeta denominada FUENTES.

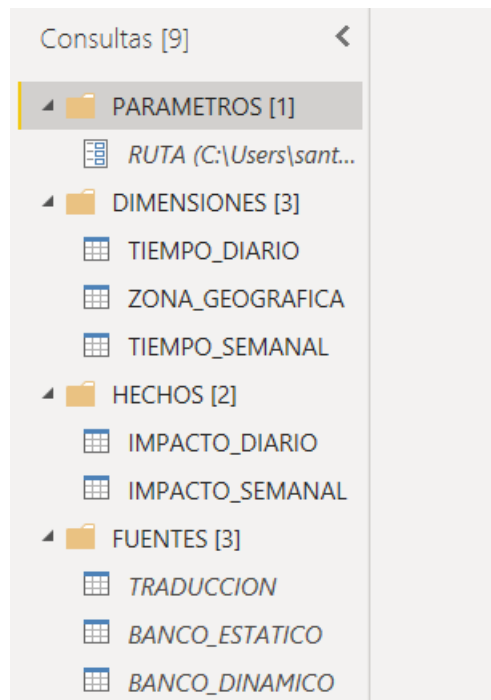


Ilustración 13 Consultas Power Query del modelo multidimensional

Luego de haber presentado cómo se han organizado los ficheros del proyecto, se va a analizar con más detenimiento las ocho consultas que constituyen al modelo final mediante los hechos y las dimensiones que los caracterizan.

6.1 Fuentes

Como se ha comentado anteriormente, la extracción de datos del modelo multidimensional, que corresponde a la primera etapa del proceso ETL, se realiza a través de estas tres consultas situadas en la carpeta FUENTES.

Por un lado, tenemos la consulta TRADUCCION que simplemente extrae los datos de tipo texto del fichero “TRADUCCION.csv”. Hay que destacar que, para poder sacar partido de este fichero, es necesario que dicho fichero contenga la columna “CODIGO_ISO”. Dicha columna nos permitirá, posteriormente, referenciar cada una de las traducciones con su respectiva zona geográfica.

Seguidamente, tenemos las consultas BANCO_ESTATICO y BANCO_DINAMICO. Sus cometidos son los mismos, dotar al Almacén de Datos de toda la información procedente del portal *Our World in Data*. La única diferencia es que, tanto la consulta estática como la consulta dinámica, lo hacen de diferente forma.

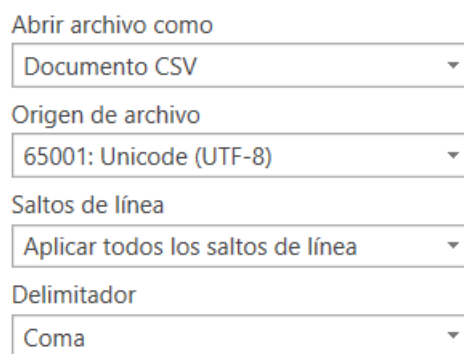
Como bien lo indica su nombre, la consulta BANCO_ESTATICO extrae los datos a nivel local a partir del fichero de datos “*owid-covid-data.csv*” que, previamente se ha descargado de *GitHub* y se ha ubicado en la ruta especificada en el parámetro RUTA.

Mientras que la consulta BANCO_DINAMICO extrae los datos de forma remota, es decir, directamente del servidor *GitHub* sin llevar a cabo ningún tipo de descarga manual por parte del

usuario. La URL, que se obtiene a través de la plataforma *GitHub*, nos permite acceder a los datos en limpio, es decir, sin procesamientos ni *taglists* de HTML del repositorio de *Our World in Data*. Paralelamente, mediante una llamada API de Power BI, se consigue extraer en tiempo real todos los datos procedentes de dicha URL, cuyo dominio es “raw.githubusercontent.com/”.

Posteriormente, en ambas consultas, es necesario reajustar la configuración de entrada para evitar errores de extracción de datos a la hora procesar y elaborar las consultas en *Power Query*:

- Para separar las columnas de datos de los diferentes valores, que constituyen a la fuente de datos, se utiliza la coma como delimitador de datos.
- Se tiene que establecer el formato de codificación UTF-8 ya que es retro compatible con ASCII y es la codificación con la que viene por defecto los datos de *Our World in Data*.



The image shows a configuration window for opening a file. It contains four dropdown menus:

- Abrir archivo como:** Documento CSV
- Origen de archivo:** 65001: Unicode (UTF-8)
- Saltos de línea:** Aplicar todos los saltos de línea
- Delimitador:** Coma

Ilustración 14 Configuración de la extracción de datos del origen

Se ha optado por tener dos consultas para la extracción de los datos del origen. En el caso de que la consulta BANCO_DINAMICO deje de recibir retroalimentación de datos por parte del servidor *GitHub*, el usuario tiene la alternativa de descargar dichos datos de forma local y extraerlos con la consulta BANCO_ESTATICO. De esta forma nos aseguramos de que, el Almacén de Datos reciba en todo momento los datos necesarios para desempeñar su función.

6.2 Dimensiones

Centrándose en la elaboración de las tablas dimensionales del Almacén de Datos, se parte de los datos de las consultas BANCO_DINAMICO y BANCO_ESTATICO. Por defecto se utiliza la primera opción.

Un aspecto relevante que hay que tener en cuenta, es que las consultas TIEMPO_DIARIO y TIEMPO_SEMANAL están constituidas, en su gran mayoría, por las mismas operaciones de transformación de datos. Posteriormente, en la fase de ampliación de la solución tendrán algunas diferencias más añadidas mediante la creación de columnas calculadas.

Tanto la consulta TIEMPO_DIARIO como TIEMPO_SEMANAL están constituidas inicialmente de la columna correspondiente a las fechas de cada uno de los impactos del banco de datos. El resto de las columnas se obviarán ya que no hay otras columnas con carácter temporal y solamente se necesita de esta columna para identificar cada uno de los impactos por COVID-19.

Inicialmente, dichas columnas, en las que se especifica las fechas de los sucesos, vienen por defecto con el tipo de datos: Texto. Por ende, se tiene que aplicar una instrucción de cambio de tipo a Date. Más adelante, estas columnas se renombran como FECHA_DIARIO, para el caso de la consulta TIEMPO_DIARIO y FECHA_SEMANAL, para la consulta TIEMPO_SEMANAL.

En el caso de la consulta TIEMPO_SEMANAL, antes de aplicar dicho renombramiento en la columna que denota la fecha. Es necesario aplicar una columna personalizada para obtener la fecha cuyo día corresponda al primer día de la semana. Para conseguir dicho resultado se utiliza la función M Date.StartOfWeek. Una vez que ha llevado a cabo esta tarea, esta columna resultante se renombra como FECHA_SEMANAL.

Columna personalizada

Agregue una columna que se calcula a partir de otras columnas.

Nuevo nombre de columna

Fórmula de columna personalizada ⓘ

Columnas disponibles

date

<< Insertar

[Información sobre fórmulas de Power Query](#)

Ilustración 15 Función del Lenguaje M Date.StartOfWeek para obtener el primer día de la semana

Enfocándonos de nuevo en ambas consultas temporales, podemos apreciar que las columnas, que se han generado, se tratan de las claves primarias del sistema operacional. Por esta misma razón, es necesario establecer mediante operaciones de transformación de datos, sus respectivas restricciones de integridad para que se cumplan los conceptos relacionales en el modelo resultante:

- Las claves primarias no pueden tener valores nulos. Por esta misma razón, se tiene que aplicar un filtro para eliminar en dichas columnas aquellos valores distintos a “null” y, por consiguiente, los valores vacíos.
- Las claves primarias identifican de forma única registros de su respectiva tabla. Es por esto por lo que no puede haber valores duplicados. Por ello, se introduce otro filtrado para eliminar valores repetidos en dichas columnas.

```
#"Filas filtradas nulas" =
Table.SelectRows("#Columnas con nombre cambiado", each [FECHA_DIARIO] <> null and [FECHA_DIARIO] <> ""),
#"Duplicados quitados" =
Table.Distinct("#Filas filtradas nulas"),
```

Ilustración 16 Operaciones para cumplimentar las restricciones de clave primaria de la consulta TIEMPO_DIARIO

Finalmente, se agrega en ambas consultas su respectiva columna de índices denominada ID_TIEMPO_DIARIO e ID_TIEMPO_SEMANAL para referenciar cada una de las claves primarias

del sistema operacional. Estas columnas, toman el relevo para convertirse finalmente en las claves primarias de tipo entero de las tablas temporales.

Por otro lado, dentro de este conjunto dimensional se encuentra la consulta ZONA_GEOGRAFICA. Esta consulta, inicialmente, toma del banco de datos todas aquellas columnas que correspondan a atributos relacionados con la ubicación en donde ocurren los impactos de COVID-19. Seguidamente, todas estas columnas se renombran con los siguientes nombres:

- CODIGO_ISO
- NOMBRE
- CONTINENTE
- NUM_HABITANTES
- DENSIDAD_POBLACION
- EDAD_MEDIA
- PRO_MAYORES_64
- PRO_MAYORES_69
- PER_CAPITA
- PRO_POBREZA_EXTREMA
- TASA_MUERTE_CARDIO
- TASA_DIABETES
- PRO_M_FUMA
- PRO_H_FUMA
- PRO_PERS_HIGIENE
- TASA_CAMAS
- ESPERANZA_VIDA
- ESCALA_DESARROLLO_HUMANO.

De igual manera a como ocurre con las consultas temporales, los datos vienen inicialmente con el tipo Texto. Así pues, es vital cambiar el tipo de datos de cada uno de los atributos a su tipo correspondiente. Es en este punto, donde surge una peculiaridad con respecto a las columnas con valores decimales, como es el caso de las tasas o las proporciones. Debido a que el origen del banco de datos es anglosajón (*Our World in Data* es un portal desarrollado por la Universidad de Oxford) se utiliza como delimitador o separador decimal el punto para separar la parte entera de la parte decimal. Por este motivo hay que establecer una operación de cambio de tipo con configuración regional por cada columna que denote una tasa o proporción. Con este paso conseguimos cambiar el punto decimal por la coma decimal, que es la que usualmente se usa en muchos países hispanoamericanos.

Paralelamente, teniendo en cuenta que el idioma del origen de datos está en inglés y el resto del modelo se ha planteado en castellano, se ha optado por traducir los nombres de los continentes y las zonas geográficas al último idioma comentado. Para ello, se ha aplicado una operación de combinación de consultas. Mediante el CODIGO_ISO de las zonas geográficas se logra adjuntar las columnas en español NOMBRE y CONTINENTE de la consulta TRADUCCION a la consulta ZONA_GEOGRAFICA.

Combinar

Seleccione una tabla y las columnas coincidentes para crear una tabla combinada.

ZONA_GEOGRAFICA



CODIGO_ISO	NUM_HABITANTES	DENSIDAD_POBLACION	EDAD_MEDIA	PRO_MAYORES_64	PRO_MAYOR
AFG	39835428	54,422	18,6	2,581	
OWID_AFR	1373486472	null	null	null	
ALB	2872934	104,871	38	13,188	
DZA	44616626	17,348	29,1	6,211	
...

TRADUCCION

CODIGO_ISO	CONTINENTE ING	NOMBRE ING	CONTINENTE	NOMBRE
AND	Europe	Andorra	Europa	Andorra
AIA	North America	Anguilla	Norteamérica	Anguila
ATG	North America	Antigua and Barbuda	Norteamérica	Antigua y Barbuda
ABW	North America	Aruba	Norteamérica	Aruba
BHS	North America	Bahamas	Norteamérica	Bahamas

Tipo de combinación

Externa izquierda (todas de la primera, coincidencias...)

Use las coincidencias aproximadas para comparar la combinación.

Ilustración 17 Operación de combinación de consultas para traducir los valores de ZONA_GEOGRAFICA

Al igual que ocurre con las claves primarias del sistema operacional de las consultas temporales. Es necesario cumplimentar las restricciones de la clave CODIGO_ISO mediante la ejecución de operaciones de transformación de datos. Gracias a este cometido, evitamos cualquier tipo de inconsistencia a la hora de identificar cada una de las zonas geográficas en donde ocurre el impacto sufrido por COVID-19. De esta forma, se obtienen la columna de índices ID_ZONA que corresponde a la nueva clave primaria de la tabla resultante.

Posteriormente, es necesario depurar los datos de la consulta ZONA_GEOGRAFICA mediante una operación de filtrado ya que existen filas adicionales junto al resto de estas que llevan a cabo el recuento de todos los datos a nivel continental y mundial. Esto puede generar fallos de cálculo en los resultados de los informes generados por el sistema debido a redundancias. El filtro trata de quedarse con aquellos casos en los que la columna CONTINENTE no tenga una cadena vacía ni tampoco *null*. Toda ubicación geográfica es aceptada en el modelo siempre y cuando pertenezca a un continente establecido.

AB _C CODIGO_ISO	AB _C NOMBRE	AB _C CONTINENTE
OWID_AFR	Africa	<i>null</i>
OWID_ASI	Asia	<i>null</i>
OWID_EUR	Europe	<i>null</i>
OWID_EUN	European Union	<i>null</i>
OWID_HIC	High income	<i>null</i>
OWID_INT	International	<i>null</i>
OWID_LIC	Low income	<i>null</i>
OWID_LMC	Lower middle income	<i>null</i>
OWID_NAM	North America	<i>null</i>
OWID_OCE	Oceania	<i>null</i>
OWID_SAM	South America	<i>null</i>
OWID_UMC	Upper middle income	<i>null</i>
OWID_WRL	World	<i>null</i>

Ilustración 18 Filas de datos de carácter sumatorio que se han suprimido mediante un filtro en la consulta ZONA_GEOGRAFICA

Esta depuración de datos se podría haber llevado a cabo sin aplicar la operación de filtrado que se ha explicado anteriormente. Únicamente, cambiando el tipo de combinación (de externa izquierda a interna) en la operación de combinación de consultas para las traducciones de las ubicaciones geográficas, ya que el fichero “TRADUCCIONES.csv” no contempla la traducción de estas filas de carácter sumatoria al no tener un CODIGO_ISO preasignado. Pero se ha optado por hacerlo con dicha operación de filtrado para tener presente en todo momento que esas filas se han suprimido para evitar sumatorios de datos innecesarios.

Estas filas posteriormente se pueden volver a calcular haciendo sumatorios en la implementación del Almacén de Datos sin que los resultados de los datos base se vean alterados.

6.3 Hechos

Finalmente, encontramos las consultas IMPACTO_DIARIO e IMPACTO_SEMANAL. El conjunto de operaciones en ambos casos es similar, pero presentan algunas diferencias.

La primera diferencia es que la consulta IMPACTO_DIARIO toma las columnas cuyos valores almacenan el impacto que hubo de COVID-19 en un día determinado, mientras que IMPACTO_SEMANAL lo hace con aquellas columnas que almacenan valores por cada semana transcurrida. Posteriormente, estas columnas se renombradas siguiendo el patrón de nombres establecido en las anteriores tablas del modelo.

En el caso de la consulta IMPACTO_DIARIO tenemos:

- NUM_TOTAL_CASOS
- NUM_NUEVOS_CASOS
- NUM_TOTAL_MUERTES
- NUM_NUEVAS_MUERTES
- TASA_TOTAL_CASOS

- TASA_NUEVOS_CASOS
- TASA_TOTAL_MUERTES
- TASA_NUEVAS_MUERTES
- TASA_PROPAGACION
- NUM_PACIENTES_UCI
- TASA_PACIENTES_UCI
- NUM_PACIENTES
- TASA_PACIENTES
- NUM_NUEVOS_TESTS
- NUM_TOTAL_TESTS
- TASA_TOTAL_TESTS
- TASA_NUEVOS_TESTS
- PRO_POSITIVOS
- TESTS_POR_POSITIVO
- NUM_TOTAL_VACUN
- NUM_TOTAL_PERS_VACUN_1RA_DOSIS
- NUM_TOTAL_PERS_VACUN_2DA_DOSIS
- NUM_TOTAL_DOSIS_REFUERZO
- NUM_NUEVAS_VACUN
- TASA_TOTAL_PERS_VACUN_1RA_DOSIS
- TASA_TOTAL_VACUN
- TASA_TOTAL_PERS_VACUN_2DA_DOSIS
- TASA_TOTAL_DOSIS_REFUERZO y ESCALA_RESTRICCION

Por otro lado, en la consulta IMPACTO_SEMANAL tenemos:

- NUM_NUEVOS_PACIENTES_UCI_SEM
- TASA_NUEVOS_PACIENTES_UCI_SEM
- NUM_NUEVOS_PACIENTES_SEM
- TASA_NUEVOS_PACIENTES_SEM.

En este punto de la implementación se encuentra la otra diferencia de ambas tablas. Del mismo modo que se ha llevado a cabo una función del lenguaje M para poder sacar el primer día de la semana en la consulta TIEMPO_SEMANAL. Se tiene que aplicar la misma fórmula para obtener el mismo resultado en la consulta IMPACTO_SEMANAL. Esta acción, posteriormente, nos facilitará el trabajo para poder establecer la relación entre la tabla IMPACTO_SEMANAL y TIEMPO_SEMANAL.

De igual manera como ocurre con las consultas temporales y las consultas relacionadas con la ubicación, es necesario cambiar el tipo de datos de cada uno de los atributos a su tipo correspondiente. Volvemos a encontrarnos la misma peculiaridad con respecto a las columnas con valores decimales en ambas consultas. Por ello, se vuelve a establecer una operación de cambio de tipo con configuración regional por cada tasa o proporción.

En este punto, es de suma importancia relacionar todas las tablas dimensionales con las tablas de los hechos tanto de carácter semanal como diario. Aplicando de nuevo una combinación de consultas en IMPACTO_DIARIO e IMPACTO_SEMANAL, se puede lograr dicha acción. De este modo, también se logra reagrupar los identificadores de las dimensiones que a

su vez actúan como claves ajenas y claves primarias en las consultas enfocadas a los hechos. En la configuración de la combinación se puede apreciar que se ha seleccionado el tipo de combinación interna ya que no nos interesa tener filas con campos nulos.

6.4 Ampliación mediante DAX

Para enriquecer las dimensiones del esquema, se han creado nuevas columnas calculadas a partir de las columnas nativas de la fuente original. Para ello se ha hecho uso de DAX, que se trata de una colección de operaciones y funciones que incorpora nativamente Power BI.

Con respecto a la tabla TIEMPO_DIARIO, se han aplicado fórmulas de DAX para sacar las siguientes columnas calculadas:

- NOMBRE_DIA
- NOMBRE_MES
- NUM_AÑOS
- NUM_DIA
- NUM_DIA_SEM
- NUM_MES

Por otro lado, en la tabla TIEMPO_SEMANAL, se han aplicado las mismas formulas para adaptarlas a su contexto temporal y así conseguir las siguientes columnas:

- NOMBRE_MES
- NUM_AÑO
- NUM_MES
- NUM_SEM_AÑO

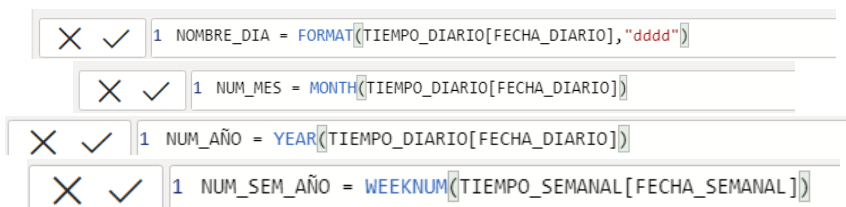


Ilustración 19 Funciones DAX para sacar el nombre, el mes, el año y el número de la semana del año de un campo DATE

Mediante la aplicación de dichas columnas derivadas podemos incorporar jerarquías a las tablas dimensionales.

En el caso de las tablas temporales, encontramos “Jerarquía diaria” para el caso de TIEMPO_DIARIO y “Jerarquía semanal” para la otra tabla temporal TIEMPO_SEMANAL.

En la primera jerarquía, que corresponde a la jerarquía diaria, se puede apreciar que la columna con una granularidad mayor es NUM_AÑO seguidamente de esta tenemos otras columnas con granularidad mucho más fina, como es el caso de NOMBRE_MES y finalmente tenemos FECHA_DIARIO.

Por otro lado, en la segunda jerarquía se puede apreciar que la columna con un mayor nivel de granularidad es NUM_AÑO y, a medida que se va afinando la granularidad, tenemos la columna NUM_SEM_AÑO y FECHA_SEMANAL.

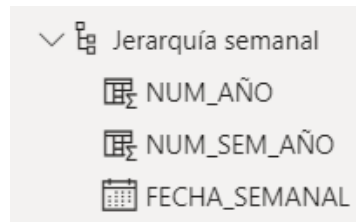


Ilustración 20 Jerarquía de la tabla TIEMPO_SEMANAL

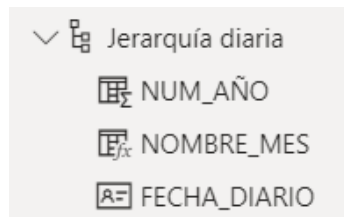


Ilustración 21 Jerarquía de la tabla TIEMPO_DIARIO

En el caso de la tabla ZONA_GEOGRAFICA, hay que destacar que también se ha aplicado otra jerarquía sobre las columnas CONTINENTE y NOMBRE. En esta jerarquía se puede distinguir que CONTINENTE ocupa la primera posición y seguidamente en un nivel inferior se encuentra NOMBRE.

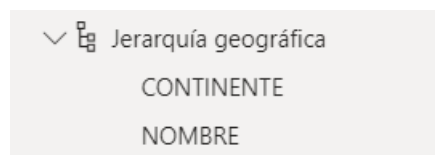


Ilustración 22 Jerarquía de la tabla ZONA_GEOGRAFICA

El cometido de dichas jerarquías es el de establecer una estructura jerárquica con diferentes niveles entre diversos atributos. Con ello se consigue que el usuario pueda navegar entre los diferentes niveles existentes en la jerarquía aplicando dos tipos de operaciones. Dichas operaciones repercuten en el nivel de agregación en el que se presentan los datos. Ambos casos son los siguientes:

- **Agregación (ROLL):** permite sustituir el criterio de agrupación utilizado en el análisis por uno de mayor granularidad. Se agregan los grupos de la consulta actual.
- **Disgregación (DRILL):** permite reemplazar el criterio de agrupación utilizado en el análisis por uno de menor granularidad. Se disgregan los grupos de la consulta actual.

Una vez que se han aplicado cada una de estas operaciones que forman parte de la faceta final de la ampliación de la solución, se concluye el desarrollo del Almacén de Datos. Con ello, se obtiene finalmente el respectivo modelo con el que se llevarán a cabo las explotaciones de datos.

Finalmente, se puede apreciar que la representación definitiva del Almacén de Datos es la siguiente:

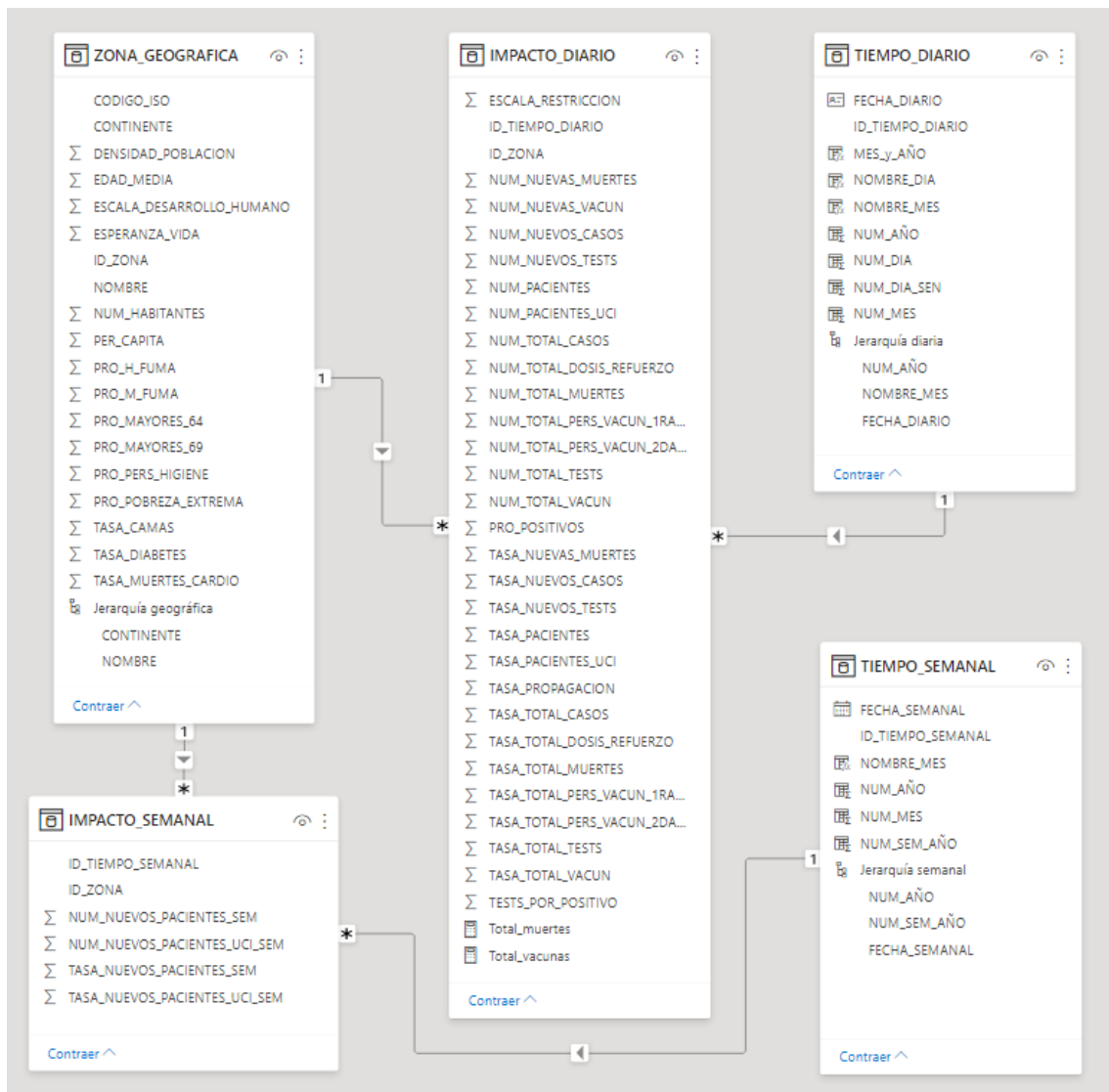


Ilustración 23 Representación gráfica del modelo final del Almacén de Datos en Power BI

7 Explotación del Almacén de Datos

Aunque el análisis de los datos cae fuera del propósito del trabajo realizado, en este apartado se van a generar algunos informes de prueba a partir de los datos almacenado, para ello se va a hacer uso de la interfaz gráfica de Power BI.

Gracias al generador de informes, que posee la herramienta de Microsoft, se van a poder recuperar y combinar los datos de distintas formas posibles con el objetivo de obtener una vista clara de los resultados extraídos de los informes.

En el primer ejemplo, se ha utilizado un gráfico de columnas agrupadas para mostrar, como valores de estudio, el número total de casos que hay de COVID-19 en cada uno de los países almacenados en el modelo. Para ello, se ha aplicado un sumatorio a la columna NUM_NUEVOS_CASOS.

Por otro lado, en la leyenda del gráfico se ha seleccionado el número del año en el que ocurren dichos sucesos. A su vez, también se ha usado dos paneles de segmentación de datos para poder acotar los datos en función de los meses y del continente. Como no se ha seleccionado ningún mes ni continente, el valor que se muestra es el total de cada uno de los años hasta ahora a nivel internacional.

Como se puede observar en dicho gráfico, el año 2021 corresponde con el año en el que más casos por COVID-19 se produjo en gran parte de los países. Sin embargo, cabe resaltar que este valor se debe a que, en el año mencionado anteriormente, el COVID-19 tuvo más margen de tiempo para hacer estragos.

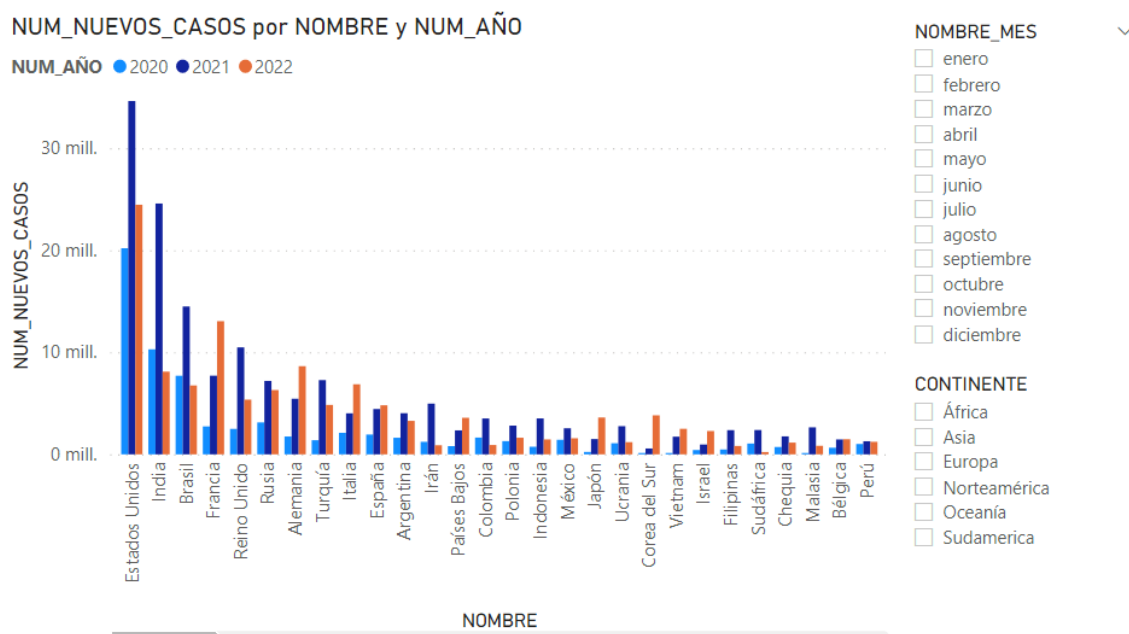


Ilustración 24 Gráfico de columnas agrupadas del número total de casos COVID-19 por cada zona geográfica junto a dos paneles de segmentación

Si acotamos los valores seleccionando un mes y un continente en concreto mediante los paneles de segmentación de datos, éstos varían. Por ejemplo, si seleccionamos enero y el continente europeo, se puede observar que se produjo un mayor número de casos en 2022 que en el año 2021 en los países pertenecientes a dicho continente.

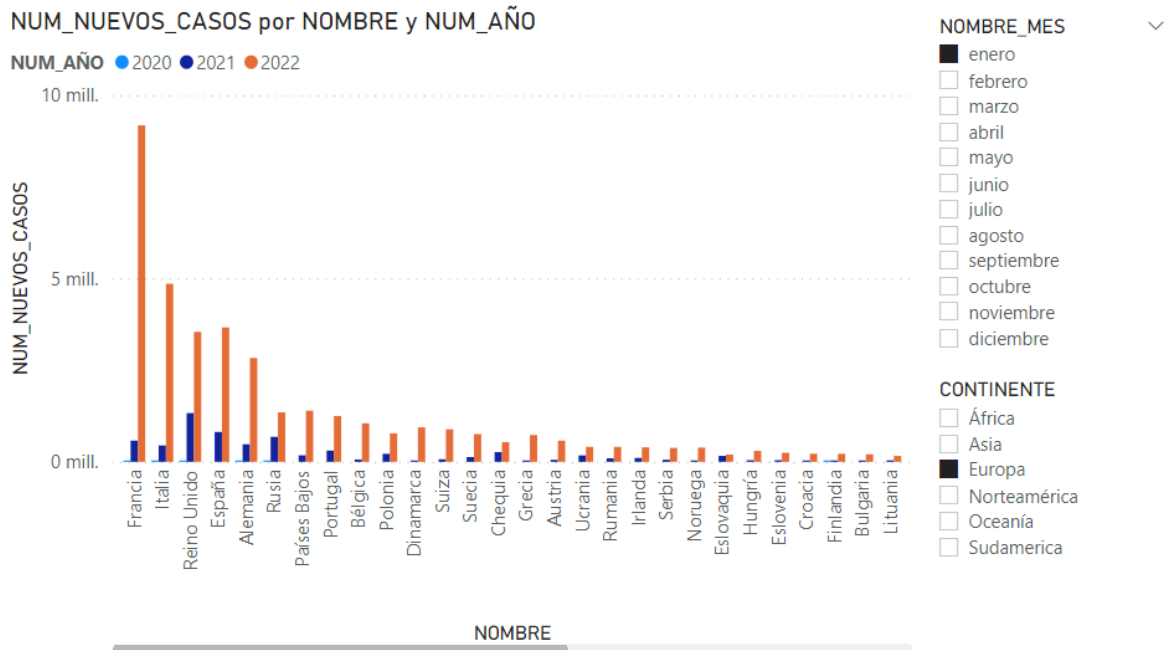


Ilustración 25 Gráfico de columnas agrupadas del número total de casos COVID-19 acotado por el mes de enero y el continente de Europa

En el segundo ejemplo, se ha utilizado un mapa coroplético⁶ para mostrar el número total de muertes en función del país. Se ha incluido en la leyenda de este tipo de gráfica interactiva el nombre del país para que todos los territorios se diferencien en función del color. A su vez también se ha incorporado de nuevo un panel de segmentación de los continentes para facilitar la navegación por el mapa dinámico. Por otro lado, también se ha incorporado las variables correspondientes a la esperanza de vida de cada país y el porcentaje de población mayores de 69 años estrictamente.

En este sentido, de forma general, aquellos países con una alta esperanza de vida y con una población más envejecida, como los países europeos, tienen un mayor número de defunciones. Esto es debido a que se trata del grupo poblacional que más se ha visto afecto por el virus.

⁶ Tipo de mapa temático en el que las áreas se sombreen de diferentes colores en función de distintos valores de una variable estadística.

NUM_TOTAL_MUERTES, ESPERANZA_VIDA y PRO_MAYORES_69 por NOMBRE y NOMBRE

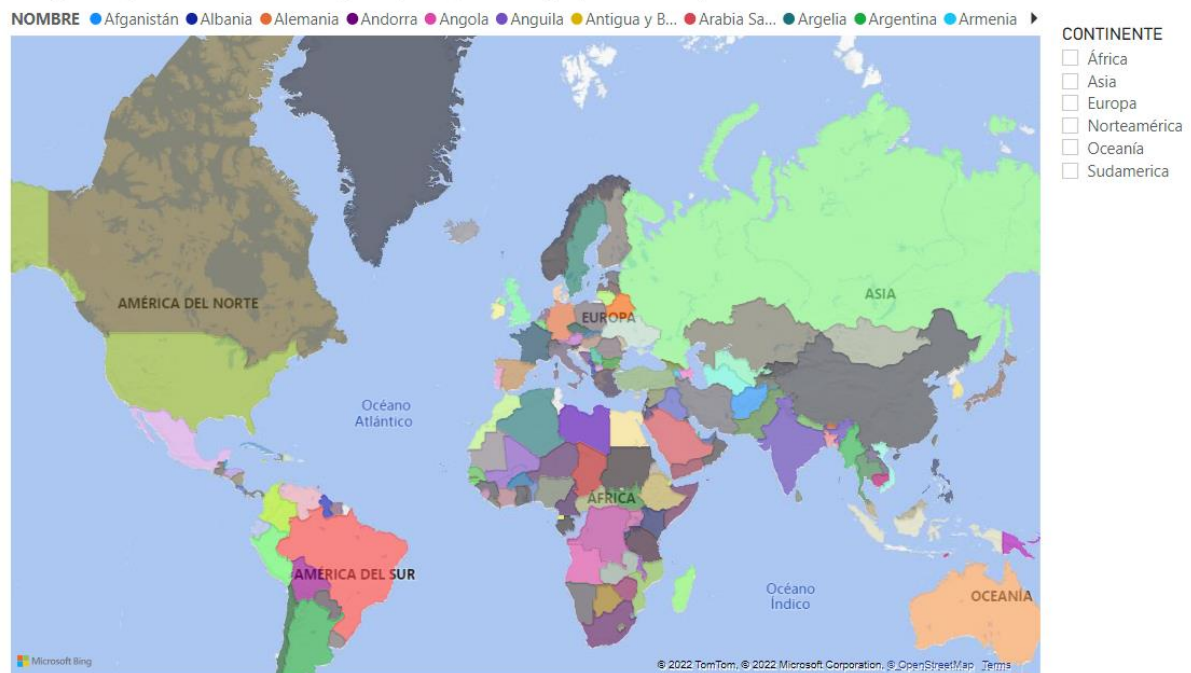


Ilustración 26 Mapa coroplético del número total de muertes que hay en cada uno de los países almacenados en el modelo junto a un panel de segmentación por continentes.

En el tercer y último ejemplo de informe, se puede apreciar que se ha optado por usar un gráfico circular. En dicho gráfico, se representa el total de camas hospitalarias que dispone cada continente por cada mil habitantes. Al mismo tiempo, en cada continente, se ha incluido la media de las personas que se encuentran en una situación de pobreza extrema. Ante estos datos, se puede observar que Europa es el continente con más camas en relación con el número de personas que habitan en él. Concretamente, hay 222 camas por cada mil habitantes, es decir, un 42,72% del total de camas. En este sentido, Europa también es el país con un menor porcentaje de pobreza extrema (0,90%).

Lo curioso es África, siendo el país en el que se sufre un mayor nivel de pobreza extrema (34,10%), es el tercer país con un mayor número de camas hospitalarias por cada mil habitantes, 53 camas por cada mil habitantes. Incluso por delante de Norteamérica, la cual dispone de 53 camas por cada mil habitantes, siendo el segundo continente con un menor porcentaje de personas en situación de pobreza extrema (5,72%).

TASA_CAMAS y Promedio de PRO_POBREZA_EXTREMA por CONTINENTE

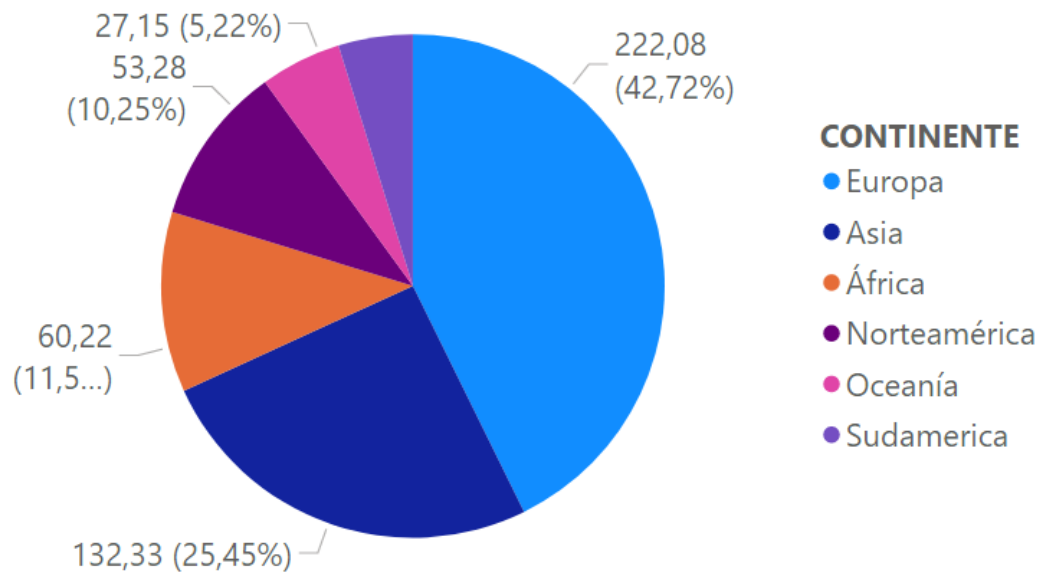


Ilustración 27 Gráfico circular del total de la tasa de camas que hay en cada continente y el respectivo porcentaje de la población que vive en extrema pobreza

8 Conclusiones

Tras todo el análisis expuesto anteriormente, en este apartado se va a llevar a cabo una valoración en la que se medirá si los objetivos que se exponen en el inicio de la memoria han sido alcanzados.

Por un lado, se ha conseguido diseñar y desarrollar un Almacén de Datos cuyo banco de datos se centra en el impacto que está teniendo el COVID-19 internacionalmente. Para ello, entre las diversas herramientas software presentadas se ha seleccionado aquella que se adapte mejor en nuestro caso. En cuyas características destaca que es gratuita y libre de costes.

La gran mayoría de las variables del banco de datos, procedente de la plataforma *Our World in Data*, se han adaptado correctamente desde el inicio del diseño para la elaboración de la estructura del modelo. Sin embargo, algunas de estas variables tuvieron que ser descartadas debido a que representaban valores fuera del contexto del proyecto.

La implementación del modelo se ha podido desarrollar exitosamente en lo que respecta al proceso ETL (mediante Power Query) y la ampliación del modelo (mediante la aplicación de las fórmulas de DAX). Ambos procedimientos se han podido completar gracias a la herramienta Power BI. Esta alternativa software permite que los accesos a los registros sean prácticamente instantáneos. De ahí que, la generación de informes no requiera de tiempos de espera.

En cuanto a la extracción de datos del proceso ETL, podemos afirmar que se ha podido implementar la retroalimentación de datos del Almacén de Datos. Con ello, el sistema puede ampliar los datos históricos hasta el día en el que el usuario haya ejecutado dicha función desde el panel principal de Power BI.

En virtud de lo estudiado, podemos afirmar que la elaboración de este proyecto nos ha permitido conocer y estudiar con detalle cada una de las pautas que hay que seguir para poder diseñar e implementar un Almacén de Datos desde cero. Así pues, nos ha enseñado que es de suma importancia entender con detalle los datos con los que se va a trabajar y el hecho de construir un modelo multidimensional correctamente.

De acuerdo con los objetivos cumplidos y los informes que hemos podido generar con Power BI a partir de nuestros datos, podemos sostener que esta herramienta es una de las más potentes y fáciles de usar para la explotación de datos de un modelo multidimensional.

A nivel personal, considero que, a partir del desarrollo de este proyecto, he mejorado mis conocimientos y aptitudes relacionados con la construcción de un Almacén de Datos. Por otro lado, gustaría destacar que he podido disfrutar de la implementación del proyecto ya que Power BI ha sido una herramienta que he usado tanto en la universidad como en el trabajo. No obstante, ha resultado ser más complejo de lo que esperaba, especialmente por la falta de tiempo que he tenido al compaginarlo con el trabajo y por la falta de experiencia para la elaboración de la memoria.

9 Bibliografía

- Casamayor, J. C. (2021). Tema 4: Diseño de Almacenes de Datos. En *Sistemas de Información Estratégicos Parte I: Almacenes de Datos* (4 ed.).
- Casamayor, J. C. (2021). Tema 5: Mantenimiento de Almacenes de Datos. Herramientas ETL. En *Sistemas de Información Estratégicos Parte I: Almacenes de Datos* (5 ed.).
- Laudon, K., & Laudon, J. (1996). *Administración de los sistemas de información*. México: Prentice-Hall Hispanoamericana.
- Mota, L., & Casamayor, J. C. (2021). Tema 1: Los Sistemas de Información Estratégicos. En *Sistemas de Información Estratégicos. Parte I: Almacenes de Datos* (1 ed.).
- Mota, L., & Casamayor, J. C. (2021). Tema 2: Introducción a los Almacenes de Datos. En *Sistemas de Información Estratégicos Parte I: Almacenes de Datos* (2 ed.).
- Mota, L., & Casamayor, J. C. (2021). Tema 3: Explotación de Almacenes de Datos: herramientas OLAP. En *Sistemas de Información Estratégicos Parte I: Almacenes de Datos* (3 ed.).
- Mota, L., & Vicent, M. (2019). Tema 2: Procesamiento de transacciones y mantenimiento de la integridad. En *Diseño y Gestión de bases de datos* (2 ed.).
- Mota, L., & Vicent, M. (2019). Tema 7: Normalización. En *Diseño y gestión de bases de datos* (7 ed.).
- Naeem, T. (2020). *Astera*. Obtenido de <https://www.astera.com/es/type/blog/data-warehouse-concepts>
- Naeem, T. (3 de febrero de 2020). *Conceptos de Data Warehouse: enfoque de Kimball vs Imon*. Obtenido de Astera: <https://www.astera.com/es/type/blog/data-warehouse-concepts/>
- Rumbaugh, J., Jacobson, I., & Booch, G. (1998). *The Unified Modeling Language Reference Manual*. Massachusetts: Addison Wesley Logman.
- Talend. (2022). *What is Database Integration?* Obtenido de <https://www.talend.com/resources/what-is-database-integration/#:~:text=La%20integraci%C3%B3n%20de%20bases%20de%20datos%20es%20el%20proceso%20empleado,actualizada%20en%20toda%20una%20organizaci%C3%B3n>



10 Anexo

10.1 Objetivos de desarrollo sostenible

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. Fin de la pobreza.				X
ODS 2. Hambre cero.				X
ODS 3. Salud y bienestar.	X			
ODS 4. Educación de calidad.				X
ODS 5. Igualdad de género.				X
ODS 6. Agua limpia y saneamiento.		X		
ODS 7. Energía asequible y no contaminante.				X
ODS 8. Trabajo decente y crecimiento económico.				X
ODS 9. Industria, innovación e infraestructuras.				X
ODS 10. Reducción de las desigualdades.	X			
ODS 11. Ciudades y comunidades sostenibles.				X
ODS 12. Producción y consumo responsables.				X
ODS 13. Acción por el clima.				X
ODS 14. Vida submarina.				X
ODS 15. Vida de ecosistemas terrestres.				X
ODS 16. Paz, justicia e instituciones sólidas.		X		
ODS 17. Alianzas para lograr objetivos.				X



Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

Los Objetivos de Desarrollo Sostenible que se pueden relacionar con este proyecto final de carrera son especialmente el de salud y bienestar y el de reducción de las desigualdades en la medida en que permite identificar aquellas fortalezas y escaseces que, en estos ámbitos, presenta cada país.

Al mostrar los valores relacionados con la mortalidad, la esperanza de vida, los recursos sanitarios, desarrollo humano, etc., se puede determinar cuál ha sido el impacto del COVID-19 en cada país, en función de su situación social, política y económica. A modo de ejemplo, la pandemia, independientemente de que haya afectado a todos los continentes, ha incidido de forma negativa a unas zonas geográficas más que a otras en función de los recursos sanitarios, tanto personal como material, el nivel de organización social y política de la zona, etc., no sólo a nivel de salud, sino a nivel económico y social.

En este sentido, aunque a partir de la elaboración y resolución de estos datos no se establecen medidas para mejorar esos objetivos, sí permite identificar cuáles son las limitaciones o debilidades que presentan cada una de las zonas desarrolladas con el objetivo de incidir en la mejora de éstas, al mismo tiempo que refleja las destrezas, fortalezas o puntos fuertes que han tenido cada uno de ellos para abordar esta problemática para saber qué aspectos en la organización deberían cambiar y cuáles mantener para conseguir estos dos Objetivos de Desarrollo Sostenible.

Del mismo modo, este mismo impacto percibido en las estadísticas que se muestran en el proyecto en cuestión se relaciona con el grado o nivel de logro de otros Objetivos de Desarrollo Sostenible como son el del agua limpia y saneamiento, así como, el de paz, justicia e instituciones sólidas. En el primer caso, porque, en regla general, hay una relación entre el número de muertes y la consecución o no de dicho objetivo, es decir, al igual que en los anteriores objetivos, aquellos países donde el agua limpia es escasa se ven representados por una mayor tasa de mortalidad durante la pandemia, aunque con algunas excepciones.

En relación con el objetivo de la paz, justicia e instituciones sólidas también tiene una relación con las estadísticas mostradas en el trabajo. En una situación tan crítica como la acaecida durante estos últimos años, las decisiones políticas que se lleven a cabo son determinantes no sólo para reducir o evitar la propagación del COVID-19 de forma insostenible, sino también para aminorar el miedo y la inseguridad ciudadana, controlar los medios de comunicación y en definitiva para asegurar la seguridad del país.

En una situación como la vivida actualmente la confianza que la sociedad tenga hacia el Estado es sustancial, de ello dependerá la legitimidad que se les otorgue para cumplir o no con las directrices pactadas. Aquellos países con menor nivel de confianza hacia la autoridad también suelen coincidir con aquellos países con más desorden social, más conflictos internos y una mayor percepción de la inseguridad, todo ello, agravado durante la pandemia dada la desigual distribución de los recursos económicos y sanitarios necesarios para su supervivencia.