UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica

Universitat Politècnica de València

Análisis e integración de la herramienta de gestión de metadatos Enterprise Data Catalog en el proyecto Alumbra

Trabajo Fin de Grado

Grado en Ingeniería Informática

Autor: David Marín Gutiérrez

Tutor: Laura Mota Herranz

Tutor Externo: Gabriel Enríquez Molina

2017/2018

Resumen

Analizar e integrar la herramienta de gestión de metadatos *Enterprise Data Catalog (EDC)* en el proyecto Alumbra de la *Consellería de Sanitat Universal i Salut Pública*.

Se dará una vista general del proyecto Alumbra y de su estructura para entender el contexto de trabajo y la importancia de la integración de una herramienta de gestión de metadatos.

Por un lado, se criticará la interfaz, y por otro, las funcionalidades de la herramienta EDC que tengan un impacto importante en el proyecto.

Por último, se concluirá si la adquisición de la herramienta aporta un valor positivo para el proyecto Alumbra

Palabras clave: Metadatos, documentación automática, bases de datos, linaje.

Resum

Analitzar e integrar la ferramenta de gestió de metadades *Enterprise Data Catalog* en el projecte Alumbra de la Consellería de Sanitat Universal i Salut Pública.

Es donarà una visió general del projecte i de la seua estrucutura per a entendre el context de treball i la importància de la integració d'una ferramenta de gestió de metadades.

Per un costat, es criticara la seua interficie, i per altre, les funcionalitats de la ferramenta que tinguen un impacte important en el projecte.

Per últim, es conclourà si l'adquisició de la ferramenta és rendible i aporta un valor positu per a Alumbra.

Paraules clau: Meta-dades, documentació automàtica, bases de dades, llinatge.

Abstract

Analyze and integrate the metadata manager tool *Enterprise Data Catalog* in the Alumbra project of the Consellería de Sanitat Universal i Salut Pública.

An overview of the project and his structure will be given to understand the work context and the importance of the metadata management tools.

On one hand, its interface will be criticized, and on the other hand, the functionalities of the *Enterprise Data Catalog* that have an important impact to the project will be analyzed.

Finally, it will be concluded if the acquisition of the tool is profitable and contributes positively to the Alumbra project.

Keywords: Metadata, automatic documentation, databases, lineage.

Tabla de contenidos

Índice de ilustraciones

Índice de tablas

1 Introducción

En este apartado se exponen las motivaciones que han llevado a la realización de este proyecto, así como los objetivos y su estructura.

1.1 Motivación

El proceso de digitalización es a día de hoy, inevitable, por ello la mayoría de empresas e instituciones guardan sus datos en formato digital. La globalización ha aumentado en gran medida la cantidad de información que se necesita almacenar para ofrecer un servicio. Por otro lado, la tendencia del mercado en cuanto al tamaño de las empresas se refiere, es la extinción de las pymes y el auge de multinacionales o instituciones, que engloban a su vez instituciones más pequeñas.

Se pueden encontrar ejemplos en todos los sectores, Amazon es uno de los primeros que viene a la mente. La cantidad de información que necesita guardar en sus bases de datos es ingente. Algo similar, aunque en menor medida, ocurre en el proyecto Alumbra, donde se pretende unificar la información de todos los hospitales y centros de salud públicos de la *Comunitat Valenciana*.

Cuando se trabaja con volúmenes de datos tan grandes, se genera la necesidad de documentar todo aquello que se modifique. Las tareas de documentación exigen de un coste temporal muy alto lo que conlleva que una parte importante del tiempo invertido por los trabajadores consista en explicar por qué y cómo han modificado un elemento. Es por todo esto que las herramientas de gestión de los datos son cada vez más importantes en las empresas e instituciones.

Las funcionalidades que ofrecen las herramientas de gestión de datos como son el linaje, el análisis de la calidad y la búsqueda del dato ayudan en gran medida a los desarrolladores, sin embargo, el impacto no es tan importante como el que puede tener la nueva funcionalidad que aporta la herramienta de estudio.

Enterprise Data Catalog (EDC), herramienta de gestión de metadatos[1], ofrece la funcionalidad de documentación automática que, una vez estudiada y verificada, puede aportar un gran valor al proyecto.

[1] Metadatos: Son aquellos datos que acompañan al dato y lo describen (ejemplo: el tipo de dato y su formato).

1.2 Motivación personal

La empresa *Capgemini*, donde estoy realizando las prácticas, me ofreció participar en un proyecto de bases de datos y *Big Data*, temas que siempre me han llamado la atención dentro del mundo de la informática y que estaba interesado en aprender.

La idea de poder ayudar en un proyecto tan importante como es Alumbra y de que el resultado de mi trabajo pueda tener un impacto real y notable en el rendimiento de mis compañeros, lo que se traduce, en una mejora visible para los usuarios finales (cualquier persona vinculada al sistema de sanidad pública de la Generalitat Valenciana), sin duda me hizo decantarme por la elección de este trabajo.

Tener la oportunidad de aprender a usar la nueva herramienta del mercado de gestión de metadatos EDC, marcó mi elección en gran medida. Sin olvidar las demás herramientas que se utilizan en el proyecto, como *PowerCenter* (PWC) y *Oracle Business Inteligencie Enterprise Edition* (OBIEE) con las que también me he familiarizado.

1.3 Objetivo

El objetivo principal del trabajo es concluir si la adquisición de la herramienta sería rentable para Alumbra, es decir, si las funcionalidades que tiene EDC, son de utilidad para la gestión del proyecto, haciendo una mención especial a la funcionalidad de documentación automática, clave en la elección de la herramienta.

Por otro lado, el trabajo busca verificar si se puede integrar la herramienta de estudio con todas las herramientas que ya se utilizan en el proyecto Alumbra (PWC, OBIEE y la base de datos de Oracle) para conseguir el mayor impacto posible.

De los dos objetivos anteriores se derivan dos objetivos necesarios:

- análisis del estado del arte, donde se descartarán otras herramientas y se justificará la elección de EDC y,
- el análisis de la herramienta y sus funcionalidades más destacables para Alumbra.

1.4 Estructura del trabajo fin de grado

El trabajo está dividido en seis apartados, El primer apartado, es la introducción, donde se describe el contexto histórico y social y se da una primera idea sobre el trabajo que se va a desarrollar. También se ven los objetivos a cumplir, la estructura de la memoria y las razones por las que se ha elegido este trabajo.

En el segundo apartado, se presenta el proyecto en el que se desarrolla todo el trabajo, se da una vista general de Alumbra y otra más específica del entorno de trabajo en el que se realizan las pruebas y el análisis.

En el tercer apartado se analiza el problema que conllevan proyectos en los que se trabaja con volúmenes de datos muy grandes. También se habla, por otro lado, de los inconvenientes legales y de seguridad que afectan al desarrollo del estudio.

En el cuarto apartado, destinado a la herramienta de estudio EDC, se presentan varios puntos muy importantes para la comprensión de la importancia que puede llegar a tener su integración en el proyecto. Por una parte, se presenta la herramienta de estudio detallando sus funcionalidades más importantes. Las herramientas que se utilizan actualmente en el proyecto también aparecen descritas en este apartado, ya que es de ellas de donde se alimenta EDC (otro subapartado, la integración). Para ayudar a comprender cada definición de cada herramienta y sus funciones, en ocasiones los subapartados se acompañan de ilustraciones que facilitan esta tarea.

El quinto apartado está centrado en el análisis de la herramienta, ya no con una visión teórica de sus funcionalidades como en el apartado anterior, sino un análisis práctico, sobre un campo de pruebas real del proyecto Alumbra de dónde se extraen los resultados en los distintos subapartados.

En el sexto apartado, se resume el resultado del trabajo, se habla de qué objetivos se han cumplido, tanto personales en el ámbito de las prácticas, como los descritos en el apartado de la introducción.

Para concluir con la memoria, se detalla la bibliografía con todos los libros, artículos y documentos, que se han utilizado a lo largo del trabajo.

2 El proyecto Alumbra

<<La misión del sistema de información ALUMBRA es dar soporte de forma unificada a todas las iniciativas de análisis de información de la Conselleria de Sanitat, constituyendo un sistema único, accesible y fiable donde proveer a toda la organización, de la información necesaria para servir de ayuda en la toma de decisiones.>> [1]

Con la puesta en marcha del proyecto se han conseguido distintas mejoras en la sociedad valenciana.

- Se reduce el impacto medioambiental negativo, fomentando el envío de información entre dispositivos electrónicos y evitando el uso del papel.
- Se garantiza un acceso territorial igualitario al servicio en la Comunitat Valenciana.
- Agiliza y facilita la toma de decisiones en cualquier nivel (sobre un paciente, sobre un centro de salud, hospital, etc.).
- Cubre el proceso completo de un paciente (atención primaria, consultas externas, urgencias, hospitalización y listas de espera).
- Se consigue <<*integrar información heterogénea que reside dispersa*>> [1] (en distintos centros de salud, por ejemplo).
- Reporta información estadística valiosa para la Conselleria de Sanitat.

En Alumbra se delega el análisis de la información de Abucasis (sistema de gestión del paciente de ambulatorios), Orion-Clinic (sistema de gestión del paciente hospitalario) y Orion-Logis (sistema de gestión de la logística, almacenes y central de compra).

El proyecto lleva en desarrollo desde 2007, actualmente trabajan 33 personas, contando a los responsables de la *Conselleria*, y el resto de personal de *Capgmenini*, el equipo de soporte, desarrolladores, analistas y jefe de proyecto. Su desarrollo se lleva a cabo en las instalaciones del antiguo hospital La Fe de Valencia.

Alumbra se caracteriza por sus elementos innovadores de los que destacamos los siguientes:

- Uso de herramientas punteras en análisis de información.
- Uso de herramientas punteras en integración de información.
- Simplicidad en el Mapa de Herramientas.
- Optimizados de procesos (permite escalar el sistema).
- Puesta en marcha de un equipo que dé soporte a usuarios basado en la metodología ITIL (Biblioteca de Infraestructura de Tecnologías de la Información).

2.1 Entorno de trabajo

Este subapartado se divide en dos bloques, el humano que hace referencia al equipo con el que se trabaja en Alumbra y el técnico donde se describe un esquema de la arquitectura de Alumbra.

2.1.1 Entorno humano

En el proyecto Alumbra trabajan 33 personas, incluidas las tres trabajadoras de la Generalitat Valenciana responsables del proyecto. Las 30 personas restantes son empleadas de *Capgemini.*

Hay ocho equipos de trabajo, cinco de ellos dirigidos por un analista y con uno o varios desarrolladores a su cargo, un equipo de gestión de dos personas (en las que se encuentra el jefe de proyecto de *Capgemini) y dos* equipos de soporte, uno de PWC y otro que da soporte a las peticiones de los usuarios de Alumbra.

Para el desarrollo del estudio es necesario trabajar en conjunto con el equipo de soporte de PWC, el analista de OBIEE y el jefe de proyecto, ya que preparar el entorno de la herramienta requería conocimientos en todas las áreas. En mi caso he trabajado con el equipo de soporte de PWC y con el de OBIEE, moviéndome del primero al segundo cuando había finalizado la preparación de la parte de PWC.

2.1.2 Entorno técnico

La base de datos con la que se trabaja en el proyecto es demasiado compleja para realizar las pruebas con la herramienta EDC, por lo tanto, es necesario simplificarla y trabajar con una partición de la misma. En este caso de estudio se realizan las pruebas con el esquema US_SIDO22_PAR, donde “PAR” es una abreviatura de la palabra paritorios, uno de los esquemas más intuitivos de los que se dispone.

En la Ilustración 1 se representa de forma gráfica el proceso por el que pasan los datos desde que son introducidos en un hospital, hasta que se solicitan para obtener un informe.

El proceso de tratamiento de datos en el esquema de PAR (teniendo en cuenta que se trata del área de paritorios), empieza en un hospital, donde se registran datos como, la fecha de entrada y salida, el centro, el nombre de la paciente, si solicita o no la epidural, etc. Toda esta información se carga en la primera base de datos, llamada *Stages* de tipo temporal, es decir, que los datos se borran cada cierto tiempo.

Es en los tres pasos siguientes donde la herramienta PWC tiene el protagonismo, transformando, filtrando, extrayendo y cargando de nuevo los datos que posteriormente serán necesarios para la creación de informes.

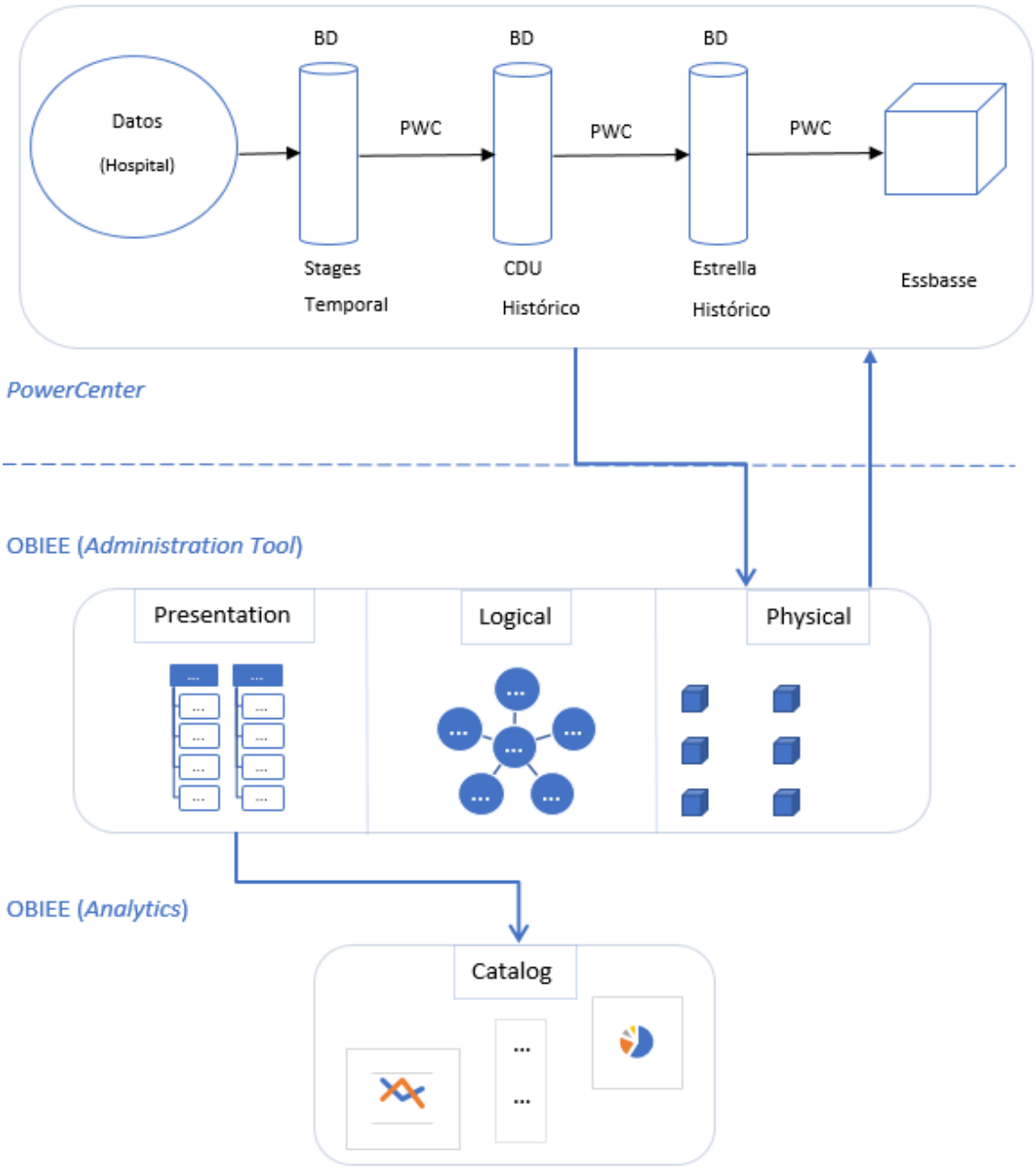


Ilustración 1: Esquema del entorno de trabajo

En el último paso de PWC, se cargan los datos en cubos de *Essbase*[2], estructuras que en teoría agilizan el proceso de recuperación de los datos cuando se hacen consultas sobre ellos.

Por último, OBIEE importa los esquemas de la base de datos, llamada CDU y de los cubos de *Essbase*, para finalmente crear un informe en su plataforma de *Analytics.*

[2] <<*Essbase* es un servidor de base de datos multidimensional, considerado como "servidor analítico">> [12], complementa a OBIEE para mejorar su rendimiento.

El objetivo de este proceso es conseguir un almacen de datos donde se pueda consultar de forma eficiente cualquier información relacionada con las actividades de los usuarios y centros de salud, para así ayudar en el análisis y la toma de decisiones a cualquier nivel de la *Consellería de Sanitat Universal i Salut Pública*.

2.2 Caso de uso

La persona responsable de recoger toda la información de una paciente, rellena el formulario con un nombre completo, edad, grupo sanguíneo, sexo etc. Una vez lo tiene completado lo carga en la primera base de datos (*Stages*).

A partir de aquí *PowerCenter* va a filtrar esa información, por ejemplo, quitando el nombre completo porque está sujeto a restricciones de privacidad, o agrupando en nuevas tablas las personas que tienen un campo en común (por rango de edad, por ejemplo). Este proceso se repite en tres ocasiones hasta que los datos quedan bien organizados para su presentación.

En este paso entra en juego OBIEE, que importa las tablas que PWC ha transformado y de nuevo se vuelven a filtrar, por ejemplo, agrupando por el tiempo medio de parto (valor que se ha extraído con las transformaciones de PWC, hora de salida – hora de entrada). Una vez se tiene la información organizada como se quiere, se procede a eliminar todo aquello que no se desea que el usuario final vea, las claves primarias[3], los códigos de los centros etc.

Por último, un usuario (médico) pide una serie de datos sobre un grupo de pacientes y OBIEE le proporciona esos datos en un informe.

2.3 Esquema PAR (Paritorios)

La razón principal por la que se ha elegido el esquema de paritorios es por su simplicidad, comparado con otros esquemas, además de ser uno de los más intuitivos. Se relaciona con un número razonable de tablas, como pueden ser, TA_D_CENTRO, TA_D_FECHA, TA_D_SEXO, TA_D_TURNO…

Dado que este esquema (ver

Ilustración *2*) es de los más simples que se disponen en el proyecto, se entiende la magnitud con la que se trabaja en Alumbra en cuanto a relaciones, tablas, registros, dependencias, etc. Es por esto que surgen ciertos problemas a la hora de gestionar el tratamiento de datos y metadatos del proyecto. Se hace casi indispensable el uso de alguna herramienta que ayude en esta tarea.

[3]Clave primaria: Conjunto de valores alfanuméricos que forman un valor único para cada registro de una tabla.

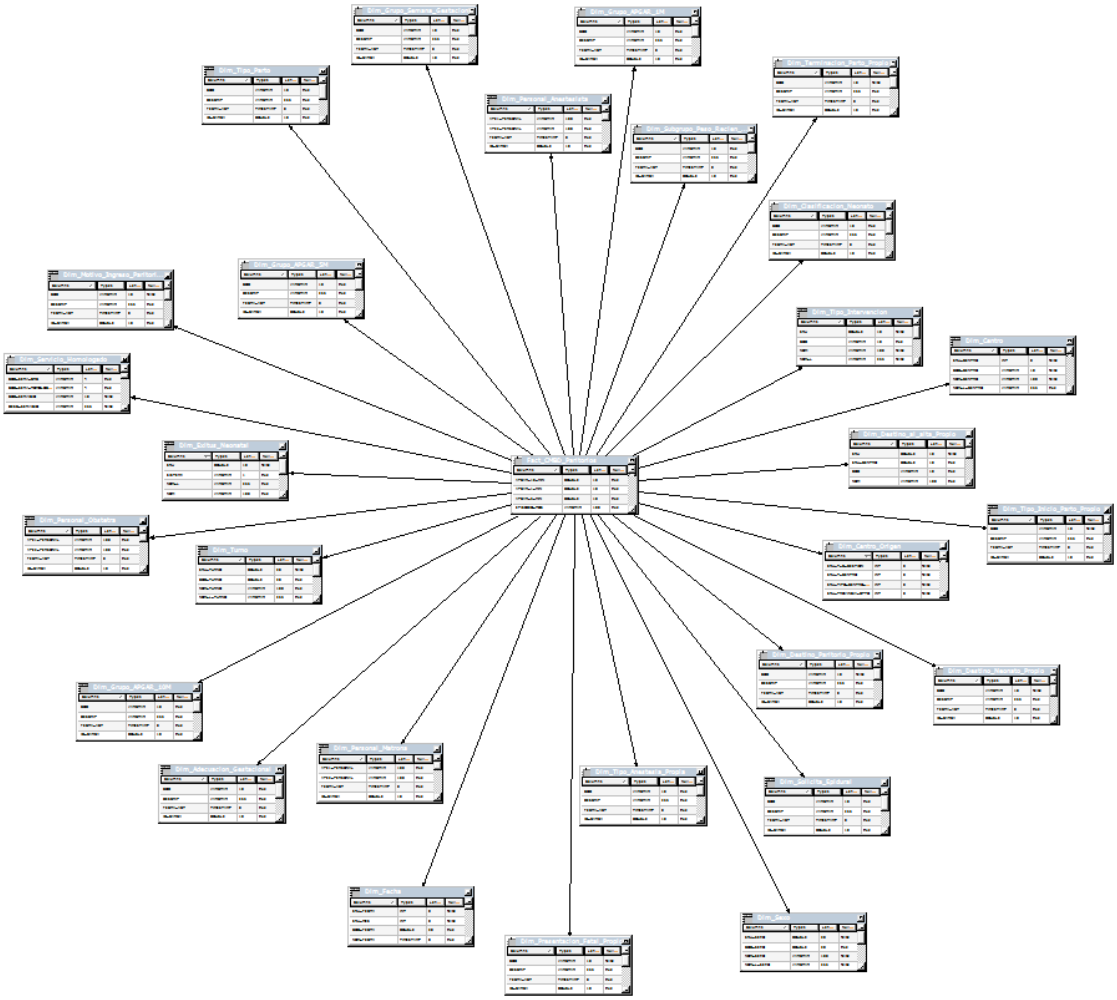


Ilustración 2: Tablas del esquema PAR

A continuación, se presenta el esquema PAR de forma más detallada, dividiendo en tres partes donde se pueden observar las tablas y sus atributos.

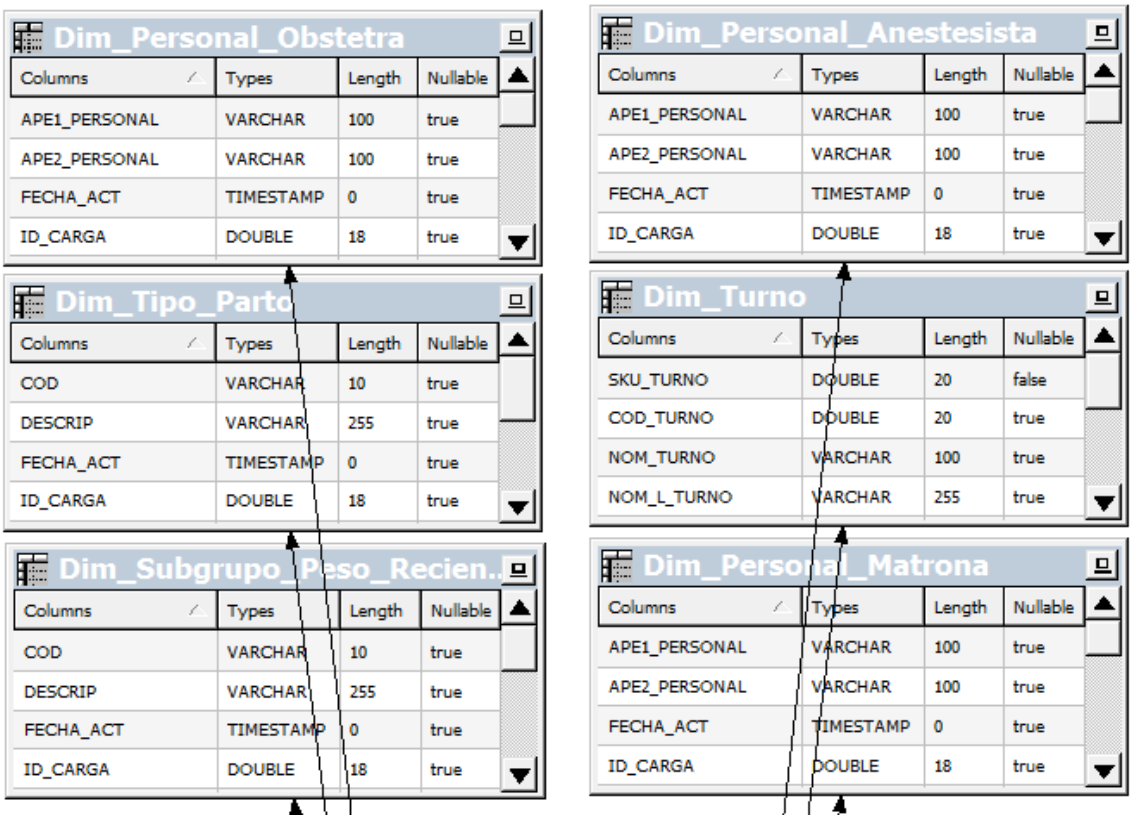


Ilustración 3: Esquema PAR 1

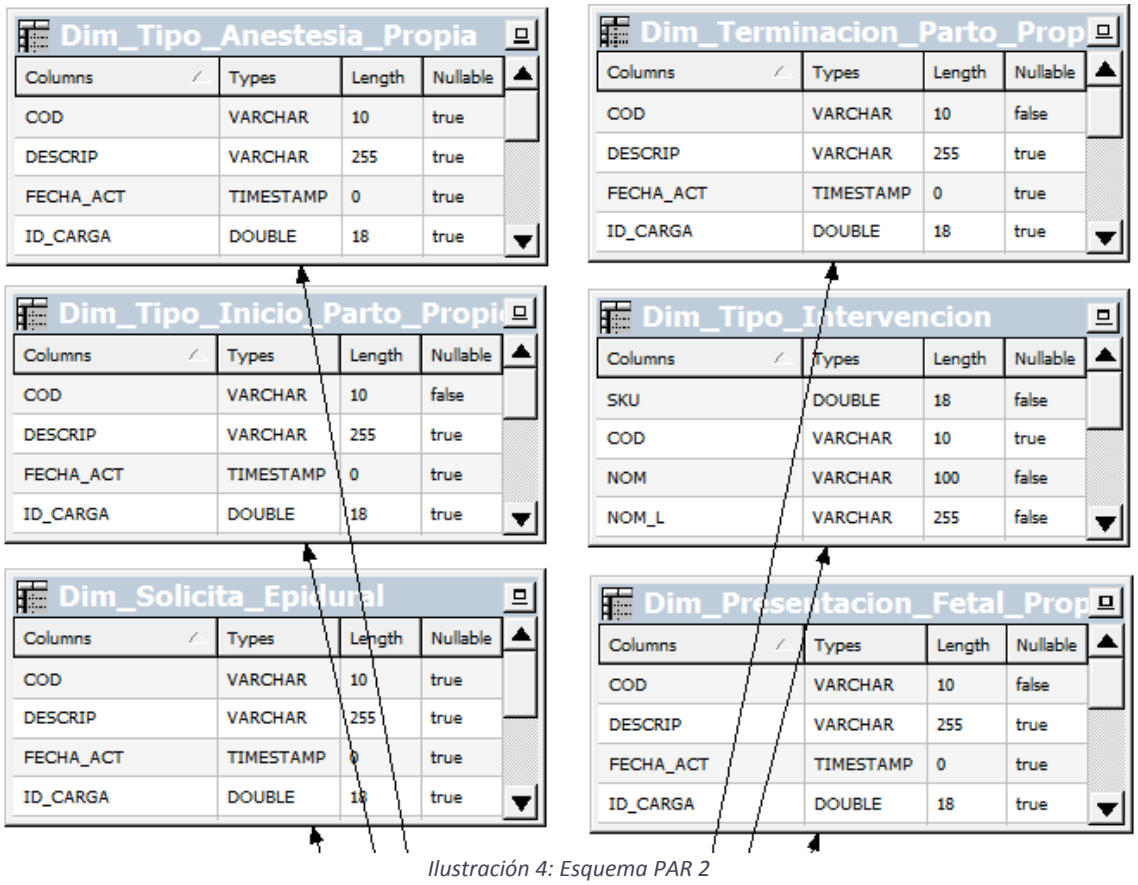


Ilustración 4: Esquema PAR 2

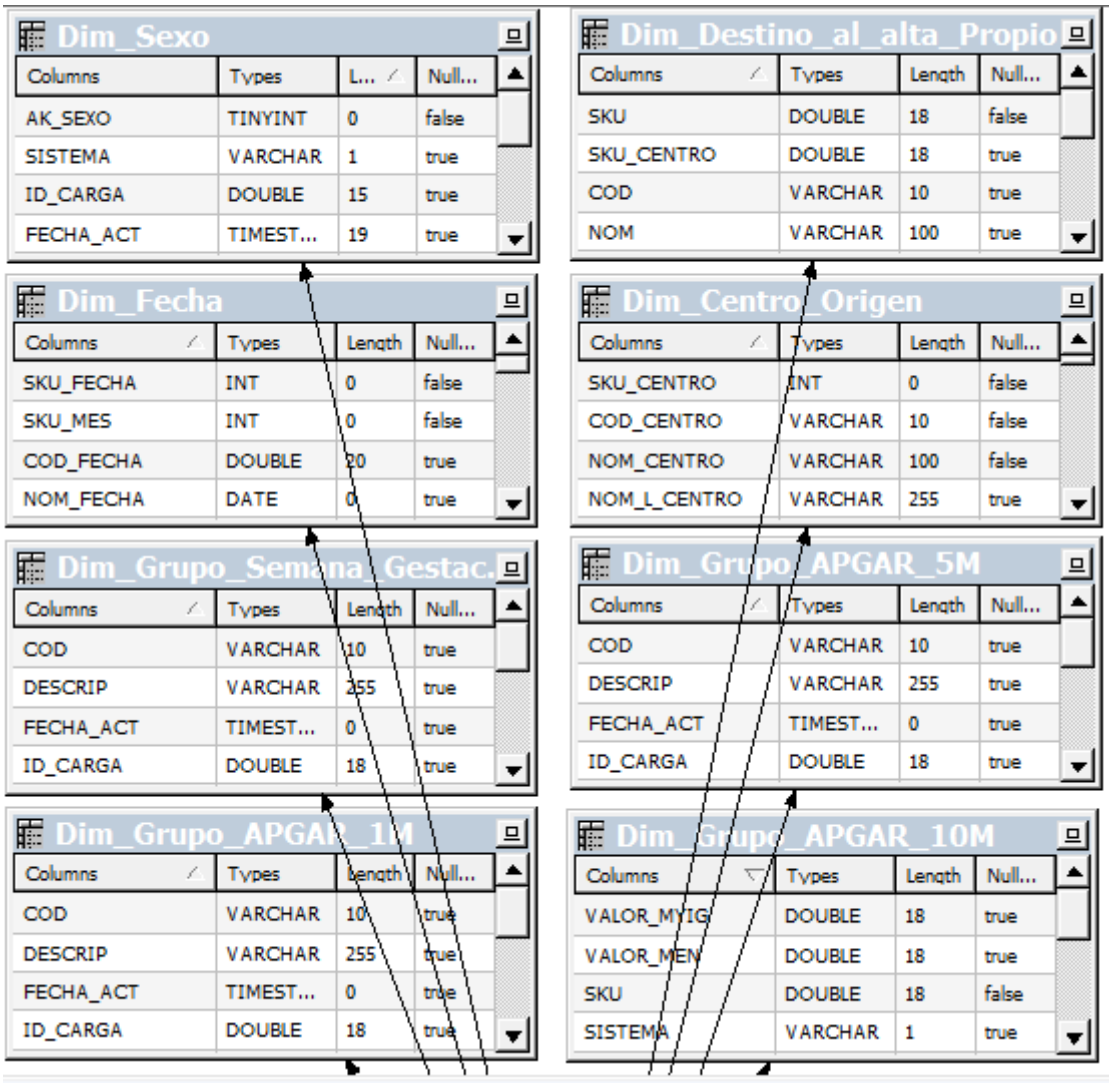


Ilustración 5: Esquema PAR 3

2.4 Estado del arte

Para poder manejar de forma eficaz un proyecto como es Alumbra, es muy recomendable usar herramientas de gestión de datos y metadatos. Unos de los proveedores del proyecto con el que se tiene una licencia para la herramienta *PowerCenter*, ofrece tres opciones que podrían resultar de utilidad para mejorar el desarrollo del proyecto:

- *PowerCenter Advanced Edition*: versión superior a la de PWC que se dispone en el proyecto,
- *Axon,* y
- *Entreprise Data Catalog*.

Antes de decidir qué herramienta podría ser la más útil para el proyecto y por lo tanto la que se va a analizar e integrar, se hizo un estudio del estado del arte, que da una visión general de qué funcionalidades comparten y en cuáles se diferencian las tres herramientas.

La versión *Advanced* de la herramienta PWC ofrece tres nuevas funciones sobre la versión estándar que se dispone en Alumbra.

- ***Metadata Manager:*** Esta función tiene como objetivo extraer los metadatos de los flujos de trabajo y cargarlos en un almacén, para así poder hacer búsquedas y realizar análisis de trazabilidad y linaje.
- ***Team-Based Development:*** Su función es la de gestionar los privilegios ligados a los datos, pudiendo restringir el acceso a ciertos usuarios.
- ***Data Analyzer:*** Permite configurar cuadros de mando, alertas, e informes, extrayendo y filtrando información de la base de datos o de almacenes de metadatos.

En la Tabla 1 se resumen las funcionalidades de las tres herramientas que se han considerado:

	PowerCenter Advanced Edition	*Axon*	*Enterprise Data Catalog*
Diccionario de datos	Sí	Sí	Sí
Trazabilidad y linaje	Sí	Sí	Sí
Seguridad y auditoria	Sí	Sí	Sí
Análisis de calidad del dato	Sí	Sí	Sí
Percepción de fallos en entidades y tablas	No	Sí	No
Alerta de errores	No	Sí	No
Documentación automática	No	No	Sí
Integración con OBIEE	-	-	Sí

Tabla 1: Comparación de las tres herramientas

La versión avanzada de PWC no ofrece ninguna funcionalidad que destaque entre las demás herramientas, por lo tanto, fue descartada casi de inmediato.

Respecto a *Axon*, como puede observarse, supera a las demás herramientas en cuanto a número de funcionalidades destacables se refiere, ofreciendo un amplio abanico (en la tabla se muestran las más relevantes). Además, esta herramienta se caracteriza por la gran amabilidad de su interfaz, pudiendo personalizarla, añadiendo o quitando funciones de la barra de herramientas y por el amplio abanico de posibilidades que ofrece en cada ventana. Ésta es la herramienta de gestión de metadatos más completa de las tres que se estudian.

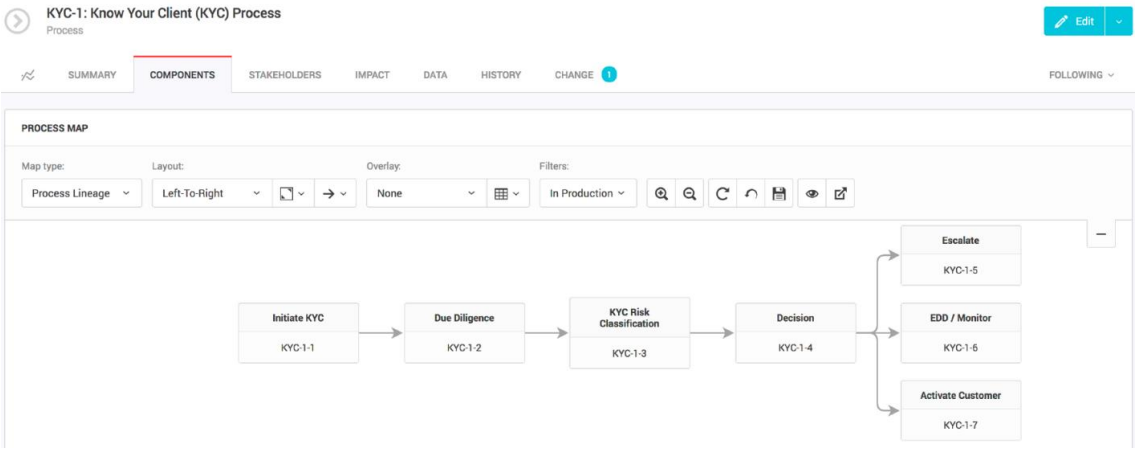


Ilustración 6: Ejemplo de trazabilidad de la herramienta Axon

Por último, la herramienta EDC no tiene un catálogo tan amplio como *Axon*, pero ofrece una funcionalidad clave, la documentación automática

Así que, pese a todas las ventajas que aporta la herramienta *Axon* frente al resto, se decidió optar por EDC. El motivo que justifica esta elección es la funcionalidad de la documentación automática. Si realmente esta función es capaz de documentar el trabajo que van realizando los desarrolladores y analistas, el tiempo que el proyecto ahorraría en estas tareas sería muy alto.

- **Crítica del estado del arte:** Dentro de los trabajos académicos presentados en la escuela de informática no se encuentra ninguno que se desarrolle centrándose en herramientas de gestión de metadatos y su integración en un proyecto de *big data* como es Alumbra.
- **Propuesta:** El estudio por tanto se centra en el análisis de una herramienta de gestión de metadatos, su funcionamiento e integración en un proyecto de *big data*.

3 Análisis del problema

Las tareas de gestión y desarrollo en proyectos con volúmenes de datos desproporcionados se vuelven complicadas debido al gran número de relaciones, dependencias y riesgos que se asumen cuando se debe modificar algún elemento del proyecto. Es por ello que las herramientas dedicadas exclusivamente a la gestión se hacen casi indispensables para el buen desarrollo de las tareas.

En el caso de Alumbra se dispone de la herramienta OBIEE, centrada en la gestión de los datos, con la que se consigue obtener un almacén de datos que agrupe toda la información del proyecto, para así poder ofrecer un servicio eficaz a los usuarios. Sin embargo, OBIEE no ayuda en las tareas cotidianas de los desarrolladores y analistas.

Un problema recurrente para los trabajadores de Alumbra es poder identificar el contexto de trabajo, las repercusiones de éste y sus riesgos. Normalmente las tareas requieren de un estudio de su entorno, previo a la modificación de los elementos, perdiendo mucho tiempo en comprender el alcance de las tareas. Aquí es donde entran en juego las herramientas de gestión de metadatos, facilitando todo este proceso y reduciendo en gran medida el tiempo invertido en la primera fase del desarrollo de las tareas.

3.1 Análisis del marco legal y ético

Los datos con los que se trabajan en Alumbra son datos personales, relativos a la salud y al historial clínico, por tanto, están considerados como datos especialmente protegidos. Es por esto que se obliga a los responsables que trabajen con este tipo de información, a adoptar las medidas necesarias para garantizar la confidencialidad y el procedimiento legal de acceso.

Según la Ley Orgánica de Protección de Datos de 1999, en su artículo 7, <<*Los datos de carácter personal que hagan referencia al origen racial, a la salud y a la vida sexual sólo podrán ser recabados, tratados y cedidos cuando, por razones de interés general, así lo disponga una ley o el afectado consienta expresamente*>> [2]

Teniendo presente lo comentado en los párrafos anteriores, la información utilizada durante la prueba no contiene ningún tipo de indicador de carácter personal (nombre, apellidos, SIP, DNI…) que pudiese implicar a una persona en concreto.

No se ha tenido en cuenta el nuevo reglamento europeo ya que la información se eliminará tras la finalización de las pruebas, que se han realizado antes del 25 de mayo de 2018, fecha en la que entra en vigor la nueva normativa RGPD (Reglamento General de Protección de Datos).

3.2 Análisis de la seguridad

Otro inconveniente a la hora de poder conectar con el responsable dela empresa *Informatica* fue el acceso al ordenador donde se han preparado todos los datos a cargar. La red de sanidad está restringida y no se puede acceder desde el exterior si no se da de alta la MAC del ordenador, y como es obvio, no se puede autorizar a un trabajador de una empresa externa para entrar en la red. Estas medidas de seguridad adoptadas por lo comentado en el subapartado anterior complican la interacción con la herramienta.

Por todo esto, para poder hacer la conexión se ha utilizado una red diferente a la de sanidad. Un USB 3G que permite tener una dirección IP pública desde la que el responsable podrá acceder de forma remota al ordenador y cargar los datos de las distintas herramientas.

3.3 Solución propuesta

La herramienta de estudio, que está centrada en la gestión de datos y metadatos, utiliza un motor de descubrimiento que automáticamente escanea, cataloga e indexa los datos de un proyecto o empresa.

Enterprise Data Catalog necesita una máquina donde hospedarse con unas características bastante exigentes (32 núcleos, 64 Gigabytes de memoria RAM, etc.). En el proyecto Alumbra, no se dispone de un servidor para realizar las pruebas que cumpla con estos requisitos, por lo tanto, la solución elegida ha sido interactuar con la herramienta de forma remota y a través de un intermediario, la empresa *Informatica*.

Se asignó un responsable por parte de la empresa *Informatica*, para cargar toda la información necesaria para realizar las pruebas, y una vez hecho se trasladaría hasta las oficinas del proyecto para probar todas las funcionalidades de la herramienta.

Para poder cargar los datos que se requieren para realizar las pruebas, es necesario que el responsable de *Informatica* pueda acceder a un ordenador, donde se trabaje en local y se disponga de los metadatos ya preparados para su carga.

4 Diseño de la solución

Durante el proceso de análisis se evaluarán las funcionalidades de la herramienta EDC, detalladas en el siguiente subapartado. Por otro lado, este punto también incluye un análisis detallado de las herramientas con las que interacciona EDC.

4.1 Enterprise Data Catalog

La herramienta de estudio ofrece un amplio abanico de posibilidades en cuanto a funciones de gestión de datos y metadatos se refiere. En este apartado se comentan las funciones más relevantes para el proyecto, haciendo mención especial a la función de documentación automática.

4.1.1 Diccionario de datos

Esta función de EDC permite recopilar todos los datos que ha recogido de las demás herramientas con las que está integrada. Con toda esa información crea un almacén de datos y metadatos, donde el usuario puede realizar búsquedas (funciona de forma similar a la búsqueda de un navegador) para encontrar cualquier elemento que se utilice en el proyecto, y a partir de ese punto, poder ver su descripción, dependencias, relaciones, privilegios, etc.

Con una barra de navegación (en la parte superior derecha), EDC permite buscar cualquier dato de forma intuitiva, listando los datos por orden de mayor coincidencia con los parámetros introducidos y pudiendo filtrar la respuesta por: nombre de la fuente, tipo de objeto, tipo de fuente, último actualizado, tamaño y propietario técnico (todos ellos en la columna de filtro a la izquierda).

4.1.2 Control de roles y privilegios

Una función básica para llevar un control sobre los datos del proyecto, es poder restringir el acceso y la modificación a los mismos. Esto se puede hacer mediante el uso de privilegios, con los cuales se pueden definir roles con más o menos derechos sobre los datos. De esta forma se consigue gestionar de forma eficiente el acceso y la modificación de los datos por parte de los desarrolladores, analistas, jefes de proyecto etc.

4.1.3 Esquema de relaciones y trazabilidad

Además de proporcionar una visión muy intuitiva de la estructura que hay alrededor de un dato, entidad, o cualquier objeto cargado en la herramienta, se puede acceder a cualquier tabla, dominio, usuario o informe, que esté relacionado con ese objeto.

A diferencia del linaje, la trazabilidad navega a través de más objetos, pero con menos profundidad. El linaje está más centrado en ver la procedencia y el destino del dato desde su origen hasta su último descendiente, sin embargo, la trazabilidad se centra en los objetos más cercanos al seleccionado que tienen algún elemento en común o que se relacionan por algún motivo (desde los usuarios con acceso, hasta objetos con dependencias directas).

4.1.4 Linaje e impacto

Se ha visto en la función de diccionario de datos, que es posible buscar un objeto (por su nombre, por ejemplo) de entre todos los que se tienen en el proyecto. Sin embargo, la función de diccionario se quedaría limitada si no se pudiese ver con qué otros objetos o datos se complementa el dato que se ha buscado. Es por esto que la función de linaje y trazabilidad perfecciona la función anterior, permitiendo ver si tiene dependencias con otros datos, de donde procede (padres, linaje), o si tiene descendencia (hijos, impacto).

EDC ofrece la posibilidad de ver el linaje de forma muy intuitiva, con una primera vista muy simple, que conforme se aumenta va detallándose más, hasta ver todos los objetos con los que tiene relación el dato buscado.

4.1.5 Calidad del dato

Gracias a esta funcionalidad, se puede ver cómo está formado un objeto, qué atributos tiene, el tamaño, el tipo de dato, su fuente, etc.

Si el objeto es una tabla, EDC proporciona información sobre el tipo de dato, el esquema al que pertenece, su nombre, el tamaño y la fecha de última modificación. Al tratarse de una tabla se puede ver el número de columnas que tiene, sus nombres, el tipo de datos de las mismas, y el dominio al que pertenecen. También se tiene la opción de profundizar más y conseguir información sobre una de las columnas, proporcionando datos estadísticos cómo el número de filas con valores distintos, nulos, u otros, o el tipo de datos (enteros, decimales, …) que se utilizan.

4.1.6 Documentación automática

Esta función que ofrece EDC es, la gran protagonista. Sin esta función, el desarrollador, analista o jefe de proyecto, debe crear un documento cada vez que completa con éxito una tarea. En

este documento se describe qué es lo que se ha modificado (hacer una consulta complicada para crear un informe, programar una nueva tabla con datos modificados, etc.). Con la función de documentación automática se consigue eliminar ese proceso, ya que el programa anota todos los cambios y los documenta.

Cada vez que se cargan los metadatos en el almacén de EDC, éste identificará los cambios y los anotará en su descripción. Además, documentará los cambios que va haciendo *PowerCenter* en una breve descripción del dato, así los desarrolladores y analistas podrán ver el proceso casi completo por el que pasa un dato desde su origen hasta su última modificación.

4.2 Diseño Detallado

A continuación, se analizan las herramientas con las que EDC se conecta para recoger sus metadatos y cargarlos en su repositorio.

4.2.1 PowerCenter

PowerCenter es una herramienta ETL (*extract, transform and load*) que se utiliza en el proyecto Alumbra. Una herramienta ETL es aquella encargada de realizar las funciones de la extracción de los datos de una base de datos, de su transformación (quitando campos de carácter sensible, por ejemplo) y de volver a cargarlos en una base de datos. En el contexto que se está considerando, *PowerCenter* extrae los datos de la base de datos *stage*, los transforma y lo carga en la base de datos CDU.

PowerCenter está estructurada en cinco (sub)herramientas. La consola de administración y el *Repository Manager* son herramientas de gestión mientras que las otras tres son destinadas al desarrollo.

Para entender la función de cada herramienta del programa *PowerCenter* es necesario tener claros los siguientes conceptos:

- <<*Un mapping se define como la representación visual de un proceso de carga. Se compone de una serie de fuentes (Sources), destinos (Targets) y transformaciones que indican cómo realizar la lectura, transformación y carga de los datos*.>> [3]
- <<*Una mapplet se define como un conjunto de instrucciones que permiten ser reutilizadas en múltiples mappings*.>> [3]
- <<*Un workflow se define como una serie de instrucciones que indican al servidor* [*de Informatica*] *cómo debe ejecutarse un proceso de carga. Un workflow se compone de un punto de inicio (Start) y una o más tareas*.>> [3]

Una vez vistos los conceptos con los que trabaja la herramienta PWC se pueden analizar las cinco herramientas que la forman.

- **Repository Server Administration Console**: Desde la consola de administración se configuran las conexiones del repositorio con la base de datos, los nodos, las copias de seguridad, los roles y privilegios y licencia de la herramienta.

  Con una vista del archivo de *logs*, la propia aplicación permite ver todas las operaciones que realiza la herramienta y los mensajes de éxito o de error, estos últimos bien documentados, ayudan a comprender qué proceso ha fallado y el motivo.

- **PowerCenter Repository Manger**: Es la primera herramienta de transformación que se usa en el proyecto Alumbra. Está centrada en la navegación entre los objetos importados o creados, no permite modificar ningún dato, simplemente se puede realizar acciones de gestión como buscar objetos, crear nuevas carpetas de trabajo, mover objetos entre carpetas y cambiarles el control de acceso.

- **PowerCenter Designer**: Es la herramienta con mayor peso dentro de PWC, ya que es ña que transforma los datos y crea los flujos de trabajo.

  - **Source Analyzer**: Define una fuente, una tabla con la que se trabaja. En esta tabla se importan los datos con unos sencillos pasos donde el usuario puede filtrar los campos de la misma y decidir qué datos pasarán a la herramienta y cuáles no.

  - **Mapplet Designer**: Una vez se tiene la fuente definida, se procede a crear un *mapping.* Esto se hace para poder hacer las operaciones, en el siguiente paso, que sean necesarias sin tener que tocar la tabla fuente.

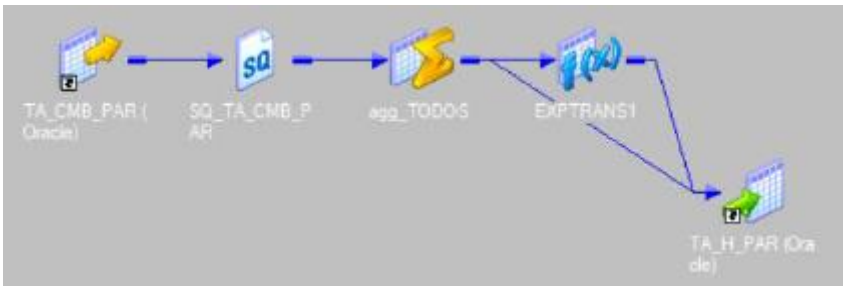


Ilustración 7: Ejemplo de mapping con Mapping Designer

  - **Transform Developer:** Aquí es donde se escogen los registros que hacen falta para una consulta, o para realizar alguna operación sobre ellos y obtener valores que no están en las tablas. Permite añadir nuevos campos a las tablas.

  - **Target Designer:** Todas las modificaciones que se han realizado en el paso anterior se deben cargar en una tabla destino. Para ello se define una tabla como tal, donde se cargan los datos elegidos de las tablas que se han modificado en el proceso.

- **PowerCenter Workflow Manager:** Con esta herramienta se puede crear un flujo de trabajo con un *mappping* asociado. La función de esta parte es la ejecución del flujo que ya se ha creado. Además, también se puede unir varios flujos de trabajo si están relacionados para que se ejecuten en una secuencia concreta (puede haber dependencias entre los flujos).

- **PowerCenter Workflow Monitor:** Gracias a esta herramienta de monitorización, el control de los flujos de trabajo es mucho más sencillo, pudiendo iniciar, parar y reiniciar

cualquier flujo que se haya creado. También proporciona información sobre cómo se ha ejecutado el proceso, si ha habido errores o si por el contrario se ha completado correctamente.

4.2.2 Oracle Business Intelligence Enterprise Edition

La herramienta de Oracle OBIEE se alimenta tanto de las bases de datos como de los cubos de *Essbase*. Trabaja en su totalidad con metadatos, razón por la que su integración con EDC es un objetivo clave para Alumbra.

Para el correcto funcionamiento de la aplicación en las pruebas, se deben seguir unos pasos para configurar correctamente los servicios de OBIEE. En primer lugar, se crea un archivo *.rpd* [4] con la herramienta de administración de OBIEE. Después se carga el rpd y se levantan los servicios de OBIEE con el *middleware Enterprise Manager*. Por último, los usuarios trabajan con la herramienta *Analytics*.

4.2.2.1 Herramienta de administración de OBIEE

Esta herramienta utiliza un archivo *.rpd*, donde se puede configurar qué tablas se quieren importar y las conexiones a la base de datos. Desde esta herramienta se tiene una visión de las tres capas que utiliza OBIEE:

- La capa física donde se importan las tablas desde la base de datos CDU y los cubos de *Essba*se.
- La capa lógica donde se filtra, transforma, agrupa y organiza la información en función de los informes que se puedan requerir.
- La capa de presentación donde se eliminan los campos que no son de interés para los informes.

En la Ilustración 8 se muestra una captura de la herramienta de administración de OBIEE con sus tres capas.

[4] Un archivo rpd contiene los esquemas y sus respectivas tablas que se han importado desde la base de datos.

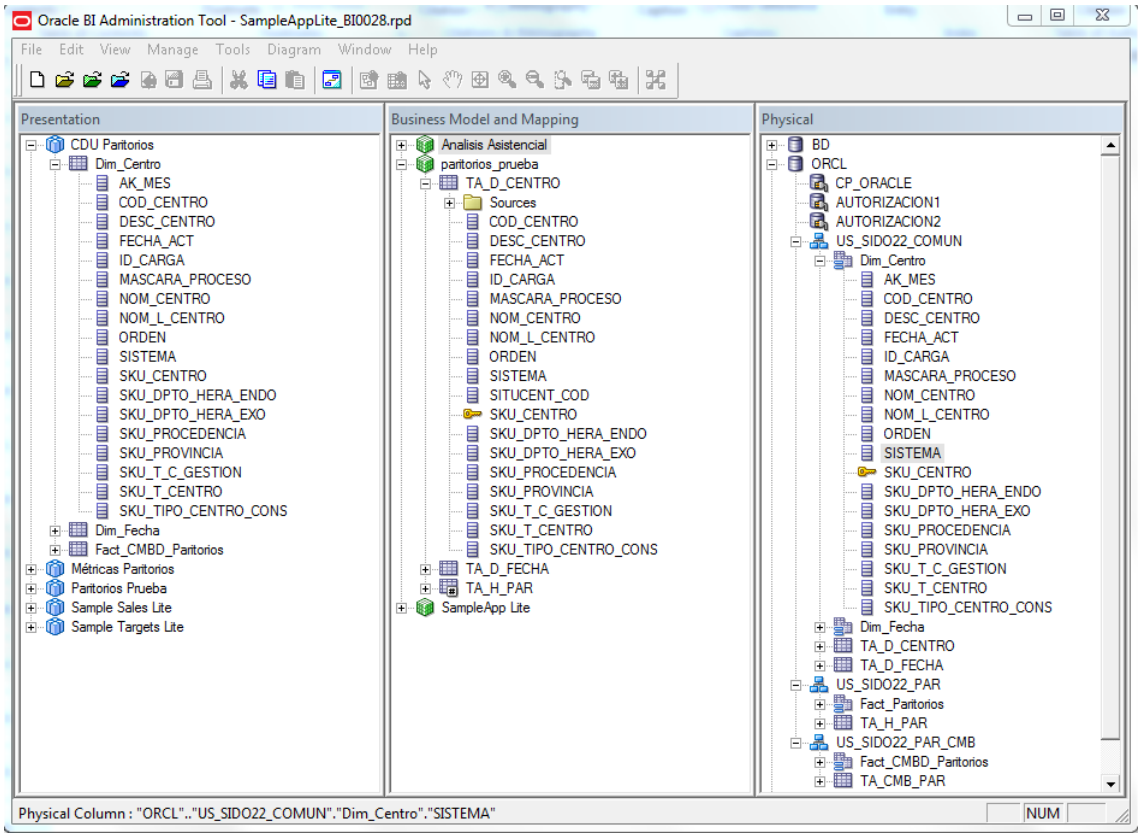


Ilustración 8: Oracle BI Administration Tool

En la Ilustración 8 se ve cómo en la capa física (ventana de la derecha) están los esquemas importados desde ORCL (nombre que se le da a la base de datos). En el caso del esquema US_SIDO22_COMUN se importan las tablas de TA_D_CENTRO y TA_D_FECHA, en este esquema se encuentran todas aquellas tablas que tienen todos los esquemas en común. De los esquemas de US_SIDO22_PAR y US_SIDO22_PAR_CMB se importan sus respectivas tablas, estos esquemas se diferencian al ser el CMB_PAR un paso de transformación y filtro anterior al H_PAR, que es el definitivo. El orden de transformación de la tabla es: TA_STG_PAR (de la base de datos Stages) pasa a TA_CMB_PAR y por último TA_H_PAR.

En la capa lógica (ventana central) es donde se trabaja con las tablas que se han importado en la capa anterior. En esta ilustración se pueden ver los campos que tiene la tabla TA_D_CENTRO del esquema US_SIDO22_COMUN. Aquí se hacen consultas, agrupaciones, se ponen restricciones para visualizar datos con derechos de privacidad, etc.

No es necesario que el usuario final visualice todos los campos de la tabla. En la capa de presentación (ventana de la izquierda) se filtra la información ya agrupada y organizada en la capa anterior, para seleccionar qué es lo que los usuarios podrán extraer en sus informes. En la imagen se puede ver que el código SITUCENT_COD ha sido eliminado ya que no aporta ningún dato valioso para los usuarios.

4.2.2.2 Oracle Enterprise Manager Fusion Middleware Control

En el proyecto Alumbra OBIEE se gobierna con un conjunto de herramientas de *Oracle* que forman parte de *Oracle Enterprise Manager Fusion Middleware Control.* Estas herramientas son las siguientes:

- ***OBIEE y* Enterprise Manager:** este *middleware* se encarga de administrar los servicios de OBIEE, pudiendo levantarlos, pararlos y reiniciarlos. También proporciona retroalimentación del estado de los servicios de la herramienta.
- **OBIEE y Analytics:** es con este *middleware* con el que los usuarios interaccionan con el resultado del proyecto. Con los permisos necesarios un usuario puede crear un informe sobre los esquemas que se le han habilitado. Cuando se requiere un informe más completo, esta herramienta también ofrece una opción avanzada de creación de informes (sólo accesible para personal autorizado). Requiere de acciones más técnicas pero que consiguen un resultado más preciso (Ilustración 9).

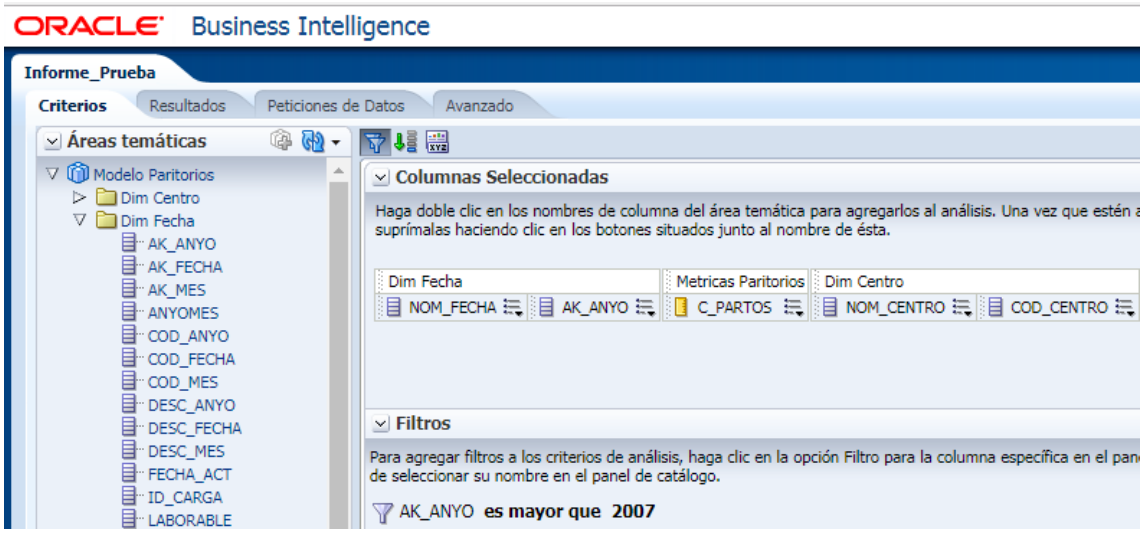
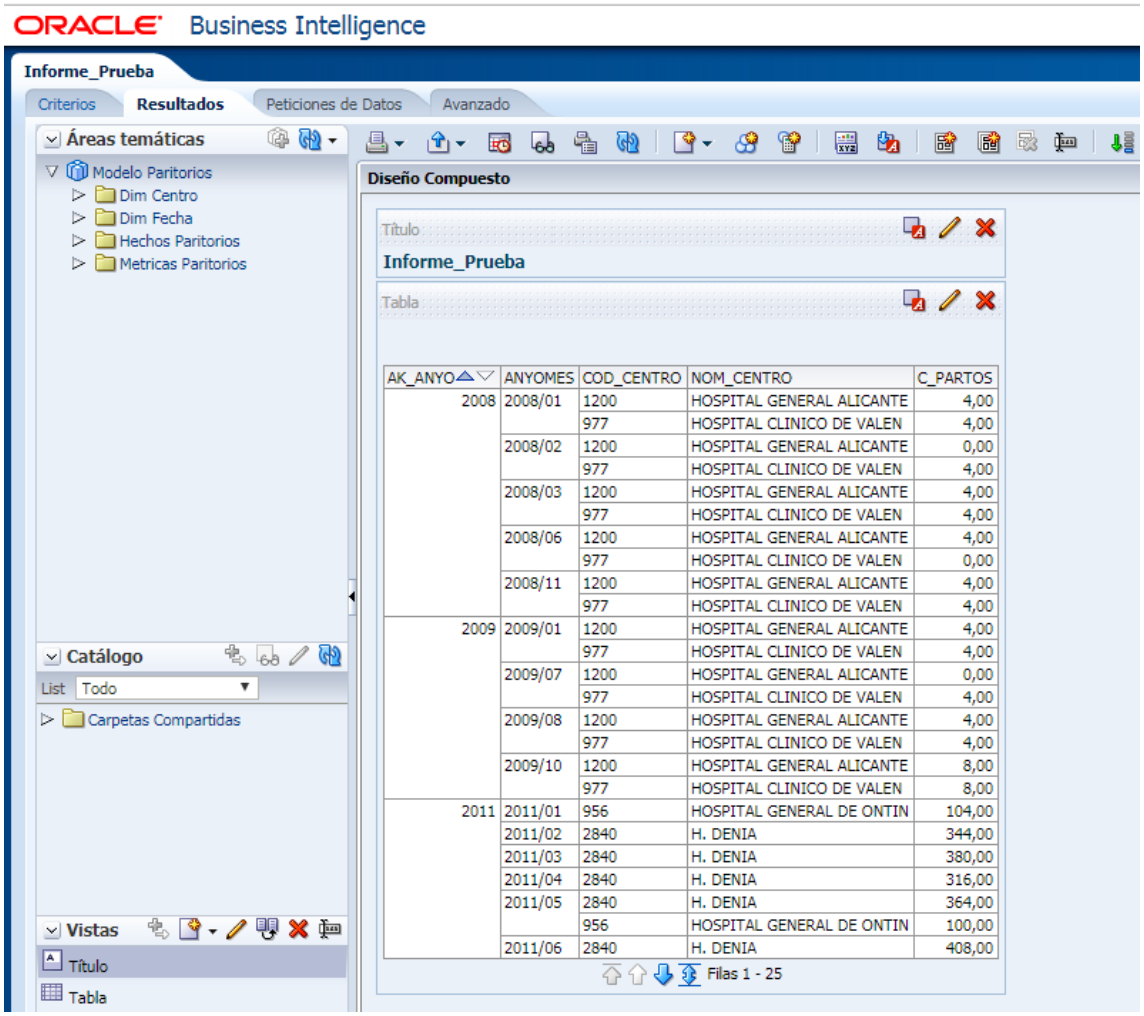


Ilustración 9: Creación de un informe en Analytics

La Ilustración 9 hace referencia a la creación de un informe, se añaden los campos (columnas) que se necesiten y se filtra la información. También es posible realizar consultas introduciendo lenguaje SQL, el ejemplo anterior sería equivalente a lanzar una consulta como la siguiente.

```
SELECT F.NOM_FEHCA, F.AK_ANYO, P.C_PARTOS, C.NOM_CENTRO, C.COD_CENTRO
FROM US_SIDO22_COMUN.TA_D_FECHA F,
     US_SIDO22_COMUN.TA_D_CENTRO C,
     US_SIDO22_PAR_CMB P
WHERE F.AK_ANYO > 2017;
```

La Ilustración 10 representa el resultado de la consulta anterior de forma gráfica.



AK_ANYO	ANYOMES	COD_CENTRO	NOM_CENTRO	C_PARTOS
2008	2008/01	1200	HOSPITAL GENERAL ALICANTE	4,00
		977	HOSPITAL CLINICO DE VALEN	4,00
	2008/02	1200	HOSPITAL GENERAL ALICANTE	0,00
		977	HOSPITAL CLINICO DE VALEN	4,00
	2008/03	1200	HOSPITAL GENERAL ALICANTE	4,00
		977	HOSPITAL CLINICO DE VALEN	4,00
	2008/06	1200	HOSPITAL GENERAL ALICANTE	4,00
		977	HOSPITAL CLINICO DE VALEN	0,00
	2008/11	1200	HOSPITAL GENERAL ALICANTE	4,00
		977	HOSPITAL CLINICO DE VALEN	4,00
2009	2009/01	1200	HOSPITAL GENERAL ALICANTE	4,00
		977	HOSPITAL CLINICO DE VALEN	4,00
	2009/07	1200	HOSPITAL GENERAL ALICANTE	0,00
		977	HOSPITAL CLINICO DE VALEN	4,00
	2009/08	1200	HOSPITAL GENERAL ALICANTE	4,00
		977	HOSPITAL CLINICO DE VALEN	4,00
	2009/10	1200	HOSPITAL GENERAL ALICANTE	8,00
		977	HOSPITAL CLINICO DE VALEN	8,00
2011	2011/01	956	HOSPITAL GENERAL DE ONTIN	104,00
	2011/02	2840	H. DENIA	344,00
	2011/03	2840	H. DENIA	380,00
	2011/04	2840	H. DENIA	316,00
	2011/05	2840	H. DENIA	364,00
		956	HOSPITAL GENERAL DE ONTIN	100,00
	2011/06	2840	H. DENIA	408,00

Ilustración 10: Resultado del informe técnico

Si todos los usuarios tuvieran que utilizar esta parte de la herramienta, la creación de informes sería una tarea demasiado enrevesada para los usuarios menos avanzados (ver Ilustración 10). Es por esto que con el middleware *Analytics* el usuario puede realizar informes de una forma mucho más intuitiva, como se puede ver en la Ilustración 11.

Ilustración 11: Filtros de informe simple en Analytics

Cómo se puede observar en la Ilustración 11, la interfaz proporciona al usuario una serie de filtros como son el año (2018), el mes (de enero a mayo), el departamento de atención (Vinaròs) y la zona de atención (Benicarló). A priori pueden parecer suficientes, pero como se ha comentado anteriormente puede quedar algo limitado para búsquedas complejas. Con los filtros anteriores se obtiene el siguiente resultado.

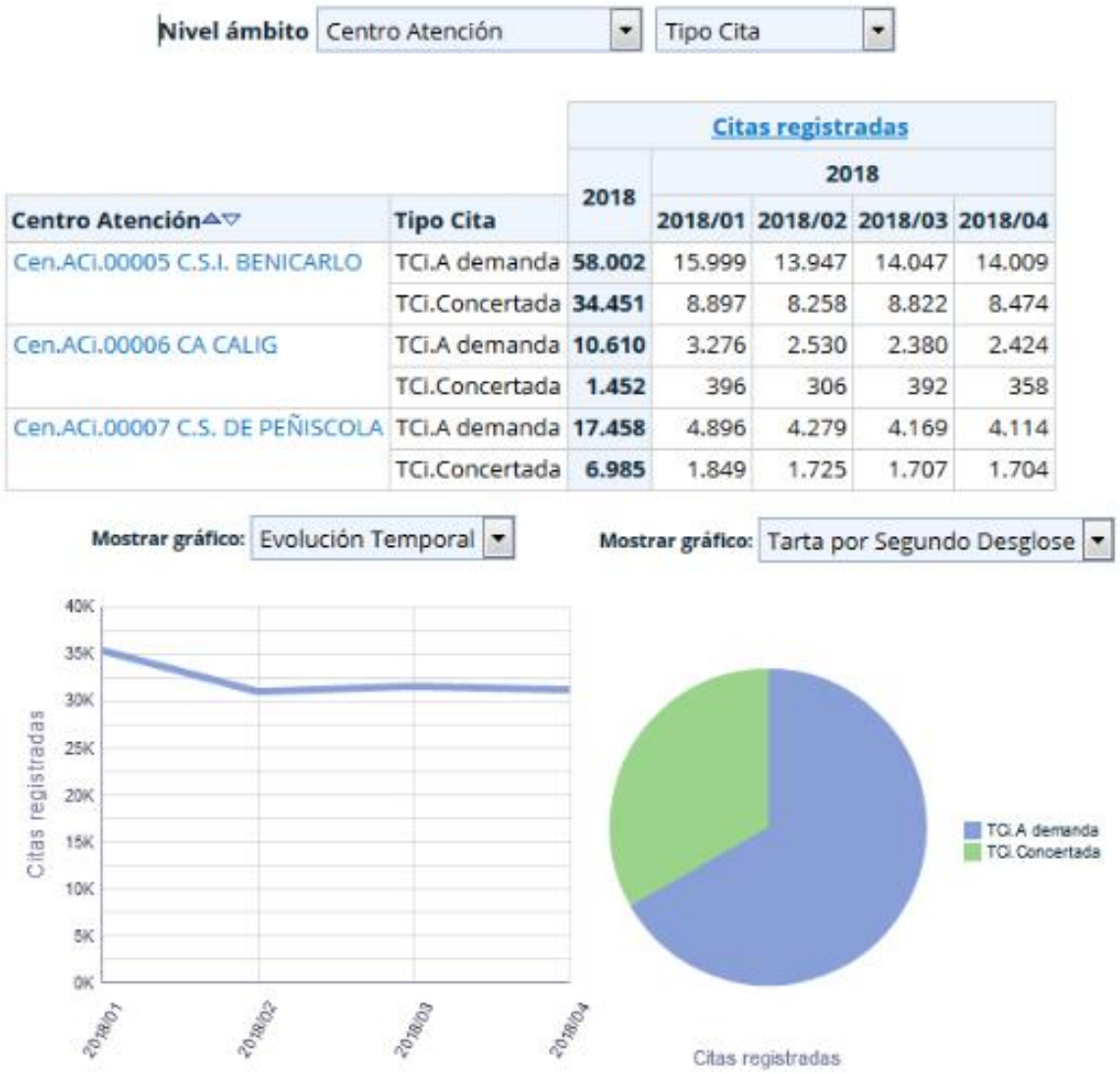


Ilustración 12: Resultado informe simple con Analytics

4.3 Desarrollo de la solución

El primer problema que se encuentra al pensar en la prueba, es el de la conectividad, como se ha explicado en el apartado de análisis de la seguridad, la red de sanidad está restringida, por ello todo el proceso de creación del entorno donde EDC recogería los metadatos se ha desarrollado en una máquina local, ubicada fuera de esta red.

Se han resuelto numerosos problemas de compatibilidad de puertos entre las herramientas de PWC y OBIEE, así como de la base de datos en cuanto a la importación del esquema PAR se refiere.

La integración de EDC con la herramienta que se utiliza en Alumbra OBIEE, ha sido el mayor problema planteado, siendo necesario la ayuda del soporte técnico de ambas herramientas para que la conexión fructificará.

5 Análisis de EDC

En este apartado se detallan los resultados de las pruebas que se han realizado con la herramienta, empezando con su integración en el proyecto y viendo en cada apartado los diferentes criterios de evaluación que se han seguido para determinar si la adquisición de la licencia aporta un valor positivo para el proyecto Alumbra.

5.1 Implantación

Uno de los primeros problemas que surgen a la hora de utilizar la herramienta en el proyecto es su integración. A priori, EDC puede recoger metadatos de cualquier fichero en cualquier formato, pero si el proceso no está predefinido, es el usuario el que crea el proceso para hacerlo.

La integración con PWC es sencilla y eficaz, algo coherente teniendo en cuenta que ambas herramientas pertenecen del mismo proveedor. Con respecto a la base de datos simplemente basta con introducir parámetros como el nombre del servicio (XE) de base de datos, el puerto que utiliza (por defecto, 1521), el nombre de la máquina que la contiene (HOST-13077) y el usuario y contraseña para acceder a ella.

Con OBIEE, sin embargo, la tarea no es trivial. En un principio no se consiguió que la herramienta cargase los metadatos, se intentaron solucionar los problemas que iban surgiendo (conectividad desde fuera de la red de sanidad, permisos y privilegios, …) sin éxito, siendo necesario pedir ayuda al equipo de soporte, perdiendo así toda referencia de cómo se ha llevado a cabo la tarea y de su complejidad.

Si esto se extrapola al resto de proyecto Alumbra, lo más probable es que para la integración con OBIEE se necesite la presencia de arquitectos tanto de EDC y Oracle como del propio proyecto. Esto retrasaría mucho la puesta en marcha y obligaría a dejar tareas del proyecto de lado durante ese periodo de tiempo.

En el esquema de la Ilustración 13 se ve que EDC recoge información de las tres herramientas del proyecto Alumbra, tanto de PWC y OBIEE, como de las bases de datos. En la prueba, esta base de datos está compuesta únicamente por el esquema PAR y CMB_PAR y algunas tablas del esquema COMUN.

EDC en ningun momento carga datos en las otras herramientas, al igual que OBIEE, sólo trabaja con metadatos.

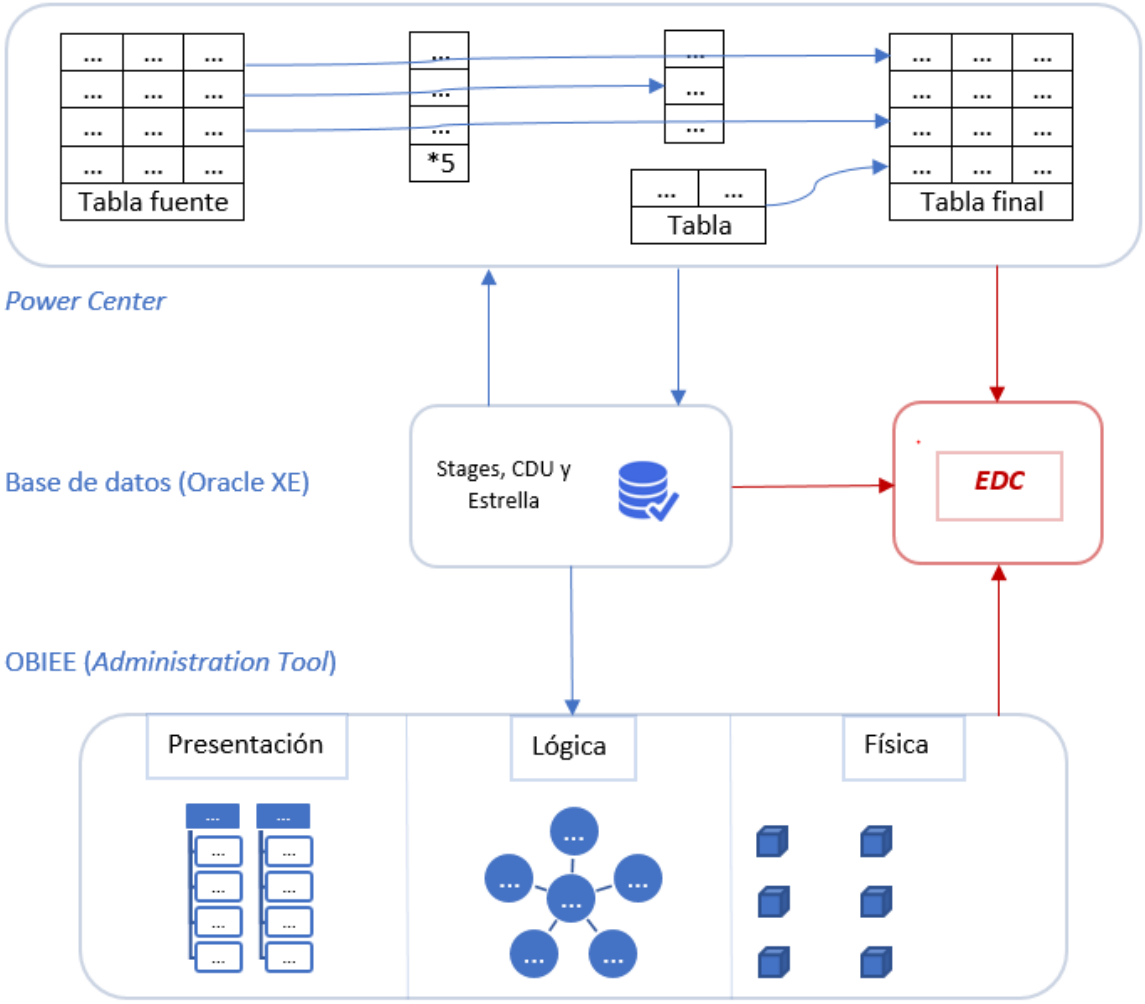


Ilustración 13: Esquema integración de EDC en Alumbra

5.2 Pruebas

A continuación, se detalla el resultado del análisis de la herramienta y de las repercusiones de sus funcionalidades en el proyecto Alumbra.

5.2.1 Interfaz

Una de las primeras críticas que se puede hacer sobre la herramienta EDC es la de su interfaz. Para analizar la usabilidad de la misma el estudio se ha realizado en función de los diez principios de Nielsen [4]. Estos principios sirven para medir qué grado de adaptación tiene en los usuarios a la herramienta, cuantos más principios cumpla, mejor será su interfaz.

- **Visibilidad del estado del sistema:** Ofrece una solución simple, aunque eficaz para abordar este problema. La herramienta usa una rueda de carga dinámica que indica al usuario si el programa está cargando alguna funcionalidad (si la rueda gira) o si, por el contrario, el programa se ha quedado congelado (la rueda no gira).
- **Correspondencia entre el sistema y el mundo real:** Ya que los usuarios que utilicen esta herramienta estarán familiarizados con los conceptos de trazabilidad, diccionario de datos, control de roles, tablas, esquemas, metadatos, etc. se puede concluir que EDC sí que mantiene una relación coherente entre el sistema y los usuarios que utilizan la herramienta.
- **Libertad y control del usuario:** EDC permite hacer y deshacer cualquier cambio que se desee, en casi cualquier momento. Teniendo en cuenta que es una herramienta destinada al análisis y la gobernación de los metadatos no hay casi ninguna acción en la que no se pueda volver a un paso anterior.

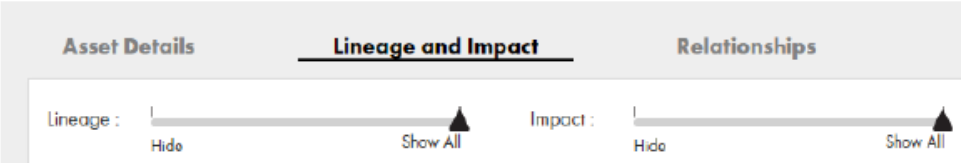


Ilustración 14: : Ejemplo del control de linaje e impacto

En la ventana de linaje tenemos la opción de ver el esquema en más profundidad y volver a una visión simple de forma muy sencilla con un elemento de tipo *slide*.

- **Prevención de errores:** A primera vista no ofrece ningún tipo de mensaje ni símbolo que ayude a los usuarios a no cometer errores durante su uso.
- **Coherencia y estándares:** La herramienta mantiene un mismo diseño en todas sus ventanas, proporcionando un sistema uniforme y homogéneo. EDC es un sistema coherente y respeta los estándares, como utilizar el color azul y no rojo para resaltar información, o como que en ningún momento se utilicen distintos diseños para un mismo objeto o acción.
- **Reconocer en vez de recordar:** El uso de iconos y figuras está muy integrado en la herramienta. En la Ilustración 15 se pueden ver cuatro iconos distintos uno representando la búsqueda (mediante una lupa) y los otros tres representando un tipo de objeto cada uno.

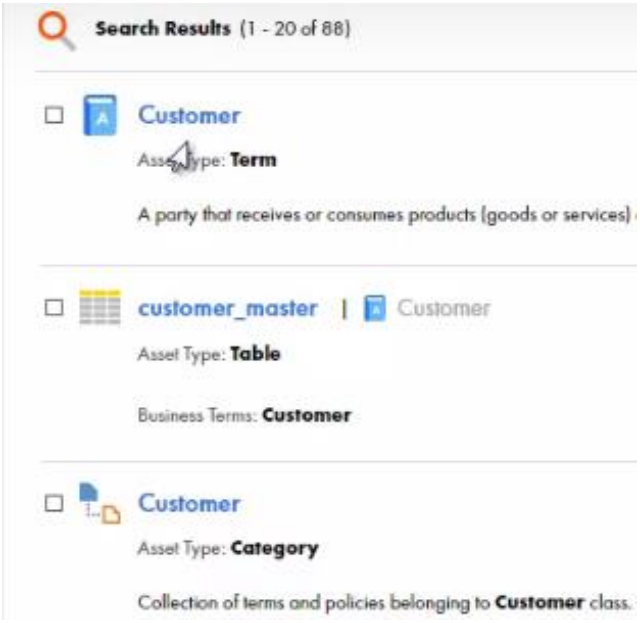


Ilustración 15: Ejemplo de diccionario de datos

- **Flexibilidad y eficiencia de uso:** La herramienta dispone de una barra superior donde el usuario puede navegar con rapidez entre distintas funcionalidades. La barra sólo está disponible una vez se tiene un objeto seleccionado, en la ventana de búsqueda no se dispone de tal menú, aunque esto no influye en la fluidez de su uso.

Ilustración 16: Ejemplo de detalles de objeto

Como se ve en la Ilustración 16, podemos navegar entre tres ventanas con un solo paso, sin necesidad de volver atrás ni acceder a un menú.

- **Diseño estético y minimalista:** EDC exprime casi al máximo este principio. Si no se solicita algún tipo de información, la herramienta estará prácticamente vacía. Sólo se permite al usuario buscar algún objeto mediante una barra de navegación.

  En cada ventana, la primera vista de las funciones es la más simple posible, como se ha visto en el ejemplo de trazabilidad y linaje.

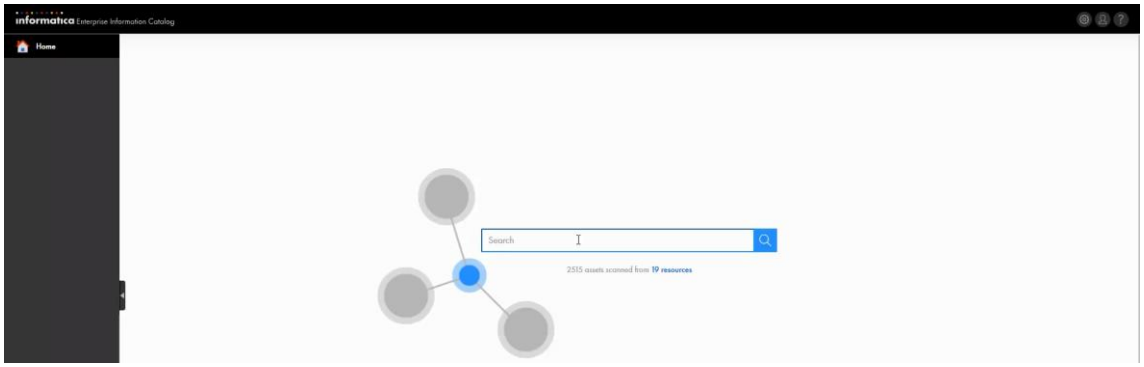


Ilustración 17: Ejemplo página de inicio

- **Ayuda y documentación:** La propia herramienta no ofrece ningún tipo de documentación. Es en internet donde los usuarios pueden encontrar tutoriales y guías sobre EDC, aunque a día de hoy son un poco escasos debido a que la herramienta no lleva mucho tiempo en el mercado.
- **Ayuda a los usuarios a reconocer, diagnosticar y recuperarse de errores:** EDC almacena en su archivo de <<*logs*>> todo lo que la herramienta hace en cada paso. Cuando hay un error aparece bien detallado en este archivo. Sin embargo, es poco práctico e intuitivo buscar entre tantas líneas de información, aunque los errores se resaltan con distinto color al resto de operaciones.

5.2.1.1 Conclusiones de la interfaz

En el subapartado anterior se puede ver que EDC cumple siete de los diez principios de Nielsen, por tanto, se puede concluir que el diseño de la interfaz de la herramienta es bueno, lo que quiere decir que los trabajadores, en teoría, se adaptarán a la herramienta en un corto periodo de tiempo.

Por otro lado, se echa en falta algún tipo de ayuda para el usuario a la hora de evitar posibles errores, como que la herramienta avise si se crean dos usuarios con el mismo nombre, y algún sistema de reconocimiento de errores más intuitivo e integrado en la propia herramienta.

La interfaz de EDC tiene cierta relevancia para el proyecto Alumbra. Cada vez que entra un nuevo desarrollador, es necesario impartir cursos de formación de las herramientas que se utilizan. Cuanto más fácil e intuitiva sea la herramienta menos costes para el proyecto y más tiempo para el desarrollo de sus tareas.

5.2.2 Crítica de las funcionalidades

En este punto del trabajo se comentarán las funciones explicadas en el apartado cuatro. Con ayuda de ilustraciones realizadas durante las pruebas, se critica el valor que aportan las funcionalidades de EDC al proyecto Alumbra.

5.2.2.1 Control de roles y privilegios

La herramienta ofrece la posibilidad de asignar privilegios a un dato o conjunto de datos de los que tiene almacenados. Cuando se crea un usuario se le asigna el rol que desempeña dentro de la empresa, desarrollador, analista, arquitecto o *manager*, estos roles tienen asociados más o menos privilegios que le permiten acceder a más o menos datos.

Esta función no requiere de tiempo ni de conocimientos específicos, consiste simplemente en rellenar un formulario con la información de los usuarios, sus privilegios y agrupar los datos según estos. Es muy importante que EDC no dé más carga de trabajo al proyecto, en Alumbra se considera la incorporación de la herramienta para facilitar el desarrollo del trabajo y reducir su carga, no para aumentarla.

Los usuarios creados se relacionarán automáticamente con los objetos a los que tengan acceso, esto quiere decir que en el momento de buscar un objeto y ver su esquema de relaciones, en este aparecerán los usuarios que pueden acceder a él.

5.2.2.2 Diccionario de datos

Esta funcionalidad ayuda tanto a desarrolladores como analistas a buscar e identificar cualquier objeto cargado en la herramienta. Facilita tareas cómo buscar elementos duplicados, identificar si un objeto ya ha sido creado o buscar un objeto sobre el que trabajar con otras funcionalidades.

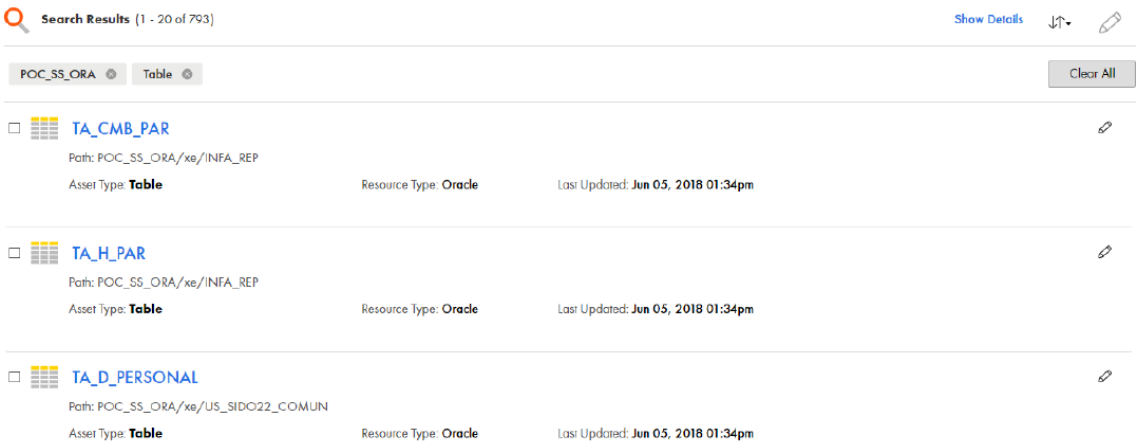


Ilustración 18: Función de diccionario de datos con la búsqueda de "tabla"

Cómo se puede ver en la Ilustración 18 se han buscado las tablas del esquema INFA_REP (nombre que recibe un esquema simplificado del esquema original US_SIDO22_COMUN junto con el esquema de paritorios US_SIDO22_PAR). EDC muestra como resultado de la búsqueda todas las tablas de que contienen ambos esquemas.

En la parte derecha de la ilustración aparece un símbolo por cada elemento resultante de la búsqueda que permite editarlo. No se trata de una modificación del dato, sino de poder cambiar atributos cómo el nombre del objeto o su tipo (en contandas ocasiones EDC identifica de forma errónea los tipos de datos que se han cargado).

EDC ofrece la posibilidad de crear tipos de datos compuestos, muy útiles para identificar conjuntos de datos con algún elemento en común. Un tipo de dato compuesto cómo puede ser <<dirección>> está compuesto por tipos cómo <<calle, número, código postal, ciudad…>>. Sin embargo, este proceso necesita mucho tiempo para llegar a tener un conjunto valioso para el proyecto Alumbra de tipos de datos compuestos, por ello, en este momento se descarta su realización.

5.2.2.3 Detalles del objeto

Esta funcionalidad tendrá poco impacto en el proyecto, aunque ayuda al usuario a tener una visión global del dato con algunos detalles e información de su contenido.

Como se puede ver en la Ilustración 19, EDC proporciona información como el nombre del objeto, en este caso una tabla llamada TA_CMB_PAR, el esquema de donde viene, INFA_REP, las columnas por las que está formada y el tipo de las mismas.

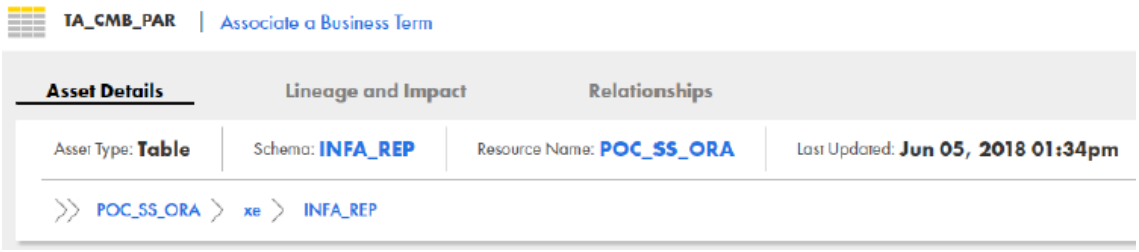


Ilustración 19: Detalles de la tabla TA_CMB_PAR

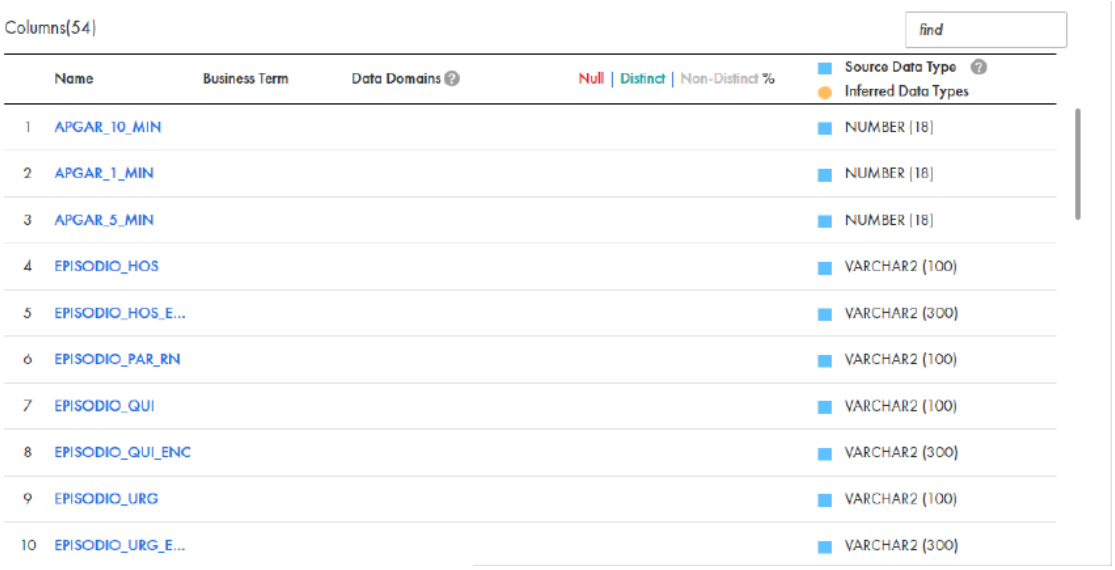


Ilustración 20: Detalles de la tabla TA_CMB_PAR

Si lo desea, el usuario puede profundizar más en su búsqueda y acceder a la información de una columna, donde EDC le proporcionará datos estadísticos, más valiosos para el proyecto que los mostrados en la ventana anterior, cómo el número de filas con un tipo concreto de valor (valor distinto o nulo, por ejemplo), o el porcentaje de datos que coinciden con el tipo de la columna.

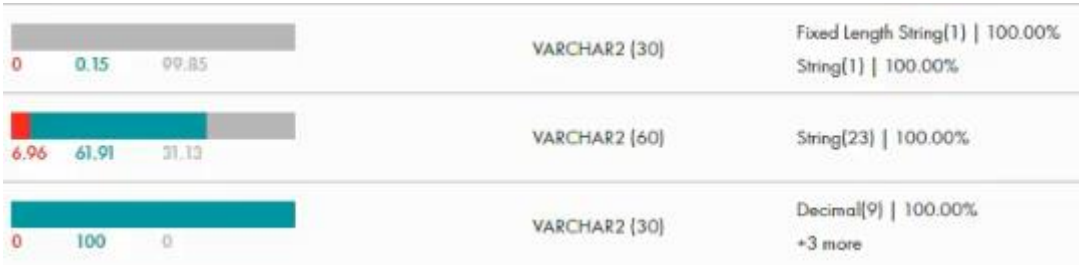


Ilustración 21: Detalle de la columna cinco

En la Ilustración 21 se puede ver que en la segunda columna el 6.96% de los datos son nulos, el 61.91% son distintos y el 31.13% restante son no-distintos, pero el 100% de ellos cumple con el tipo fuente del dato, es decir, los valores que contiene son del tipo VARCHAR2 y por tanto son correctos.

5.2.2.4 Linaje e impacto

Esta función ha cobrado un protagonismo muy importante gracias a su complementación con la documentación automática. Gracias al linaje los desarrolladores pueden ver, con el nivel de detalle que consideren oportuno, todo el proceso por el que pasa un objeto. Esto les ayuda a saber de dónde proviene el objeto, cómo ha sido modificado, gracias a la descripción de la documentación automática, y qué dependencias pueden darse si se modifica.

Sin EDC los desarrolladores recurren a los analistas para obtener una visión global del proceso y entenderlo. Con esta función esto ya no es necesario, ahorrando tiempo tanto a unos como a otros. Evidentemente también es de gran ayuda para los analistas, aunque tengan un conocimiento muy amplio del proyecto.

El método anterior para entender un proceso complejo era buscar uno a uno en los *workflows* y *mappings* (en PWC) que se utilizan en la tarea asignada y ver qué modificaban, sus dependencias y relaciones hasta llegar al dato deseado. Un proceso complejo, que requiere de un conocimiento profundo del esquema donde se lleva a cabo el desarrollo de la tarea.

Asset Lineage Summary Asset Impact Summary

	Asset Name	Asset Type	Resource Name	Resource Type
1	TA_DP_DESTINO_NEONATO	Table	POC_SS_ORA	Oracle
2	TA_DP_DESTINO_PARITORIO	Table	POC_SS_ORA	Oracle
3	TA_DP_MOTIVO_ING_PAR	Table	POC_SS_ORA	Oracle
4	TA_DP_PRESENTACION_FETAL	Table	POC_SS_ORA	Oracle
5	TA_DP_SERVICIO	Table	POC_SS_ORA	Oracle
6	TA_DP_TERMINACION_PARTO	Table	POC_SS_ORA	Oracle
7	TA_DP_TIPO_ANESTESIA	Table	POC_SS_ORA	Oracle
8	TA_DP_TIPO_INICIO_PARTO	Table	POC_SS_ORA	Oracle
9	TA_D_CENTRO	Table	POC_SS_ORA	Oracle
10	TA_D_CLASIFICACION_NEONATO	Table	POC_SS_ORA	Oracle
11	TA_D_EXITUS_NEONATAL	Table	POC_SS_ORA	Oracle
12	TA_D_FECHA	Table	POC_SS_ORA	Oracle

Ilustración 22: Linaje de la tabla TA_CMB_PAR

Asset Lineage Summary | **Asset Impact Summary**

	Asset Name	Asset Type	Resource Name	Resource Type
1	TA_H_PAR	Table	POC_SS_ORA	Oracle
2	m_CALCULA_SKU_APGAR_1	Mapping	POC_SS_PWC	PowerCenter
3	m_LOAD_DWH_PARTOS_1	Mapping	POC_SS_PWC	PowerCenter
4	m_LOAD_DWH_PARTOS_2	Mapping	POC_SS_PWC	PowerCenter

Ilustración 23: Impacto de la tabla TA_CMB_PAR

Estas dos ilustraciones (Ilustración 22 e Ilustración 23) representan el linaje e impacto, en este orden, de la forma más detallada; mediante una tabla se enumeran los demás objetos con los que se relaciona TA_CMB_PAR. Por otro lado, EDC proporciona una vista mucho más gráfica e intuitiva que puede simplificarse o desarrollarse a voluntad del usuario y que ayuda en gran medida a desarrolladores y analista.

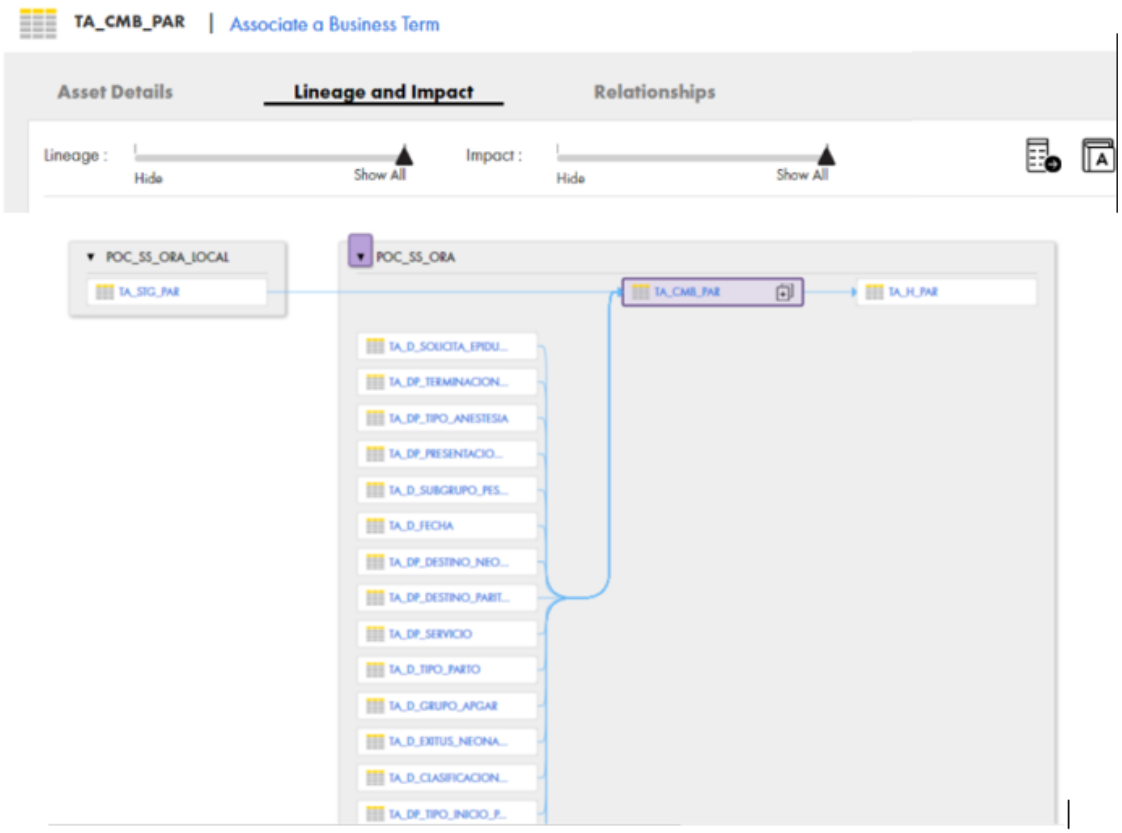


Ilustración 24: Representación gráfica de linaje e impacto de TA_CMB_PAR

Como si de un esquema de grafos se tratara, EDC une mediante flechas, los elementos relacionados entre sí. En este caso la tabla TA_CMB_PAR (resaltada en la imagen), tiene como origen la tabla TA_STG_PAR, dependencias con una serie de tablas del esquema COMUN y al final desemboca en la tabla TA_H_PAR.

5.2.2.5 Documentación automática

Lamentablemente no se pueden poner ilustraciones de los procesos que realizan los *mappings* en su interior, es trabajo de desarrollo del equipo de Alumbra y no se ha conseguido el permiso para mostrarlo.

En la Ilustración 25 se muestra el *mapping* LOAD_DWH_PARTOS_1. Se puede ver la tabla origen TA_CMB_PAR y todo el proceso hasta que los datos llegan a la tabla destino TA_H_PAR, con cambios introducidos mediante código SQL (<<SQ_>>) como se ven en los dos primeros pasos y operaciones de cálculos agregados (<<agg_>>) y de unión (<<EX>>) de *PowerCenter.*

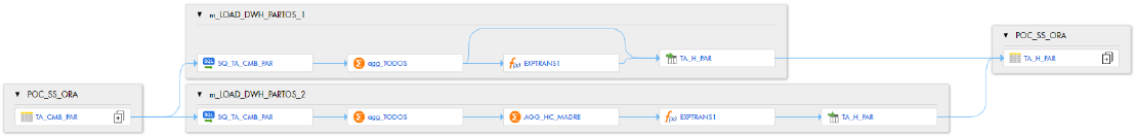


Ilustración 25: Mapping documentación automática

A continuación, se puede ver con más detalle lo descrito en el párrafo anterior. Ampliando la Ilustración 25, dividiéndola en cuatro partes.

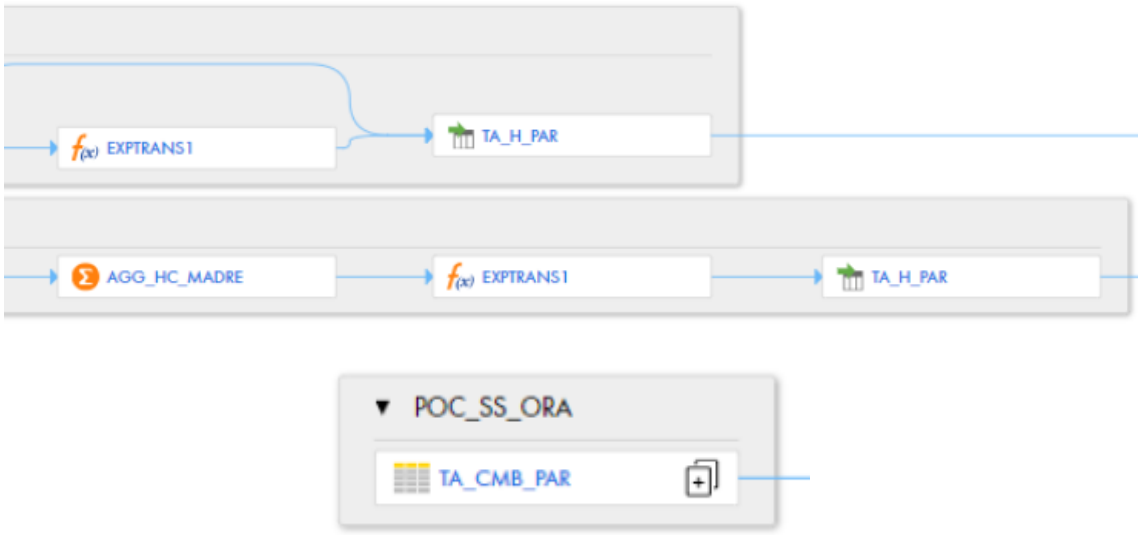


Ilustración 26: Representación gráfica de documentación automática del mapping m_LOAD_DWH_PARTOS_1

Los desarrolladores, analistas y miembros del equipo Alumbra pueden acceder a cada objeto mostrado en las dos ilustraciones anteriores y ver cómo EDC ha añadido información del proceso que hace cada tarea (SQL y operaciones de *PWC*). Aunque el proceso de ejecución de cada tarea no está muy detallado es de gran valor ya que da una primera idea del objetivo que se pretende conseguir con la ejecución del *mapping*.

El nivel de detalle de la documentación es suficiente para que aporte un valor positivo para el proyecto, aun siendo necesario documentar las tareas más complejas, ya que la documentación no aporta comentarios del motivo por el que se ha hecho, o el objetivo que se pretende conseguir.

EDC añade un campo de texto a un objeto con información sobre qué operaciones y con qué objetos las hace, la operación de agregación Tiempo_Transcurrido resta a Fecha_Fin el campo Fecha_Inicio, por ejemplo.

Un aspecto importante a tener en cuenta es que EDC no lee SQL. Como se ha visto en la ilustración anterior, el *mapping* tiene dos elementos tipo SQL, esto quiere decir que se realizan acciones mediante el uso de este lenguaje. Aunque los responsables aseguran que la herramienta en un futuro sí que podrá documentar estos cambios introducidos por código, a día de hoy, no es capaz de documentar este paso.

Afortunadamente el uso de *scripts* en lenguaje SQL no entra en el desarrollo de buenas prácticas del proyecto Alumbra y siempre se intenta evitar, aunque en demasiadas ocasiones su uso se hace necesario. Por ello, que la herramienta no sea capaz de leer estos *scripts* no supone un gran problema.

5.2.3 Rendimiento

Un punto a tener en cuenta a la hora de evaluar la herramienta es su rendimiento. Si para lograr representar cualquier función comentada en el subapartado anterior, la herramienta necesita dedicar una cantidad de tiempo muy elevada, su valor para el proyecto disminuye.

	Superficial	*Medio*	*Profundo/a*
Diccionario de datos Primera ejecución	2seg	2seg	4seg
Diccionario de datos más ejecuciones	1seg	1seg	1seg
Linaje e impacto Primera ejecución	2seg	6seg	24seg
Linaje e impacto más ejecuciones	1seg	1seg	6seg

Tabla 2: Coste temporal de dos funcionalidades

En la Tabla 2 no aparece la funcionalidad de esquema de relaciones y trazabilidad ya que se cargan a la vez que el linaje e impacto, el coste temporal de cambiar de una ventana a otra es despreciable.

Teniendo en cuenta que se trata de uno de los esquemas más fáciles del proyecto, el tiempo en las funciones de linaje e impacto y el esquema de relaciones y trazabilidad puede aumentar significativamente. Por otro lado, la herramienta requiere de prestaciones muy altas en cuanto a memoria RAM (64 Gb como mínimo) para poder ofrecer una mejora notable de rendimiento una vez ha cargado las funciones por primera vez, como se puede observar en las segundas pruebas de la tabla.

Si se extrapola a un esquema complejo, la herramienta tardaría alrededor de cinco minutos en ofrecer el linaje e impacto de una búsqueda profunda por primera vez y un minuto con quince segundos aproximadamente, en las siguientes ocasiones. A priori puede parecer demasiado tiempo, pero en comparación con OBIEE, no lo es.

Afortunadamente la herramienta no requiere de una configuración inicial compleja gracias al motor de escaneo que dispone. Un valor añadido ya que no se busca que la herramienta aporte más trabajo, si no que ayude a disminuirlo.

5.2.4 Benchmarking

Una vez se han recopilado suficientes datos se puede hacer un *benchmarking,* que consiste en comparar dos sistemas y evaluar las mejoras obtenidas del sistema modificado respecto al

original. En este caso se considera como sistema al conjunto de herramientas del proyecto Alumbra (con la integración de EDC y sin ella).

- **Aceleración:** En este caso la aceleración mide la ganancia de velocidad del sistema modificado respecto al original.
  - El tiempo empleado en realizar tareas de documentación varía mucho dependiendo de la complejidad de ésta. Tomando varios valores reales de muestra, se puede obtener el tiempo medio para realizar una tarea de documentación.
  - En Alumbra los trabajadores imputan tiempo a las tareas por horas enteras, aunque una tarea tarde en documentarse 2 horas y 48 minutos, el trabajador imputará 3 horas y si son menos de 2 horas y 30 minutos imputará 2 horas.

	Simples	*Medias*	*Complejas*
Documentar la solución de un error de informe de OBIEE	1h	3h	5h
Documentar la modificación de un *workflow*	2h	3h	6h
Documentar el desarrollo de una consulta	1h	2h	2h

Tabla 3: Coste temporal de distintas tareas

  - Los expertos del proyecto aseguran que el coste temporal de las tareas de documentación se reduciría en torno al 75% en las tareas más simples y un 20% en las más complejas. Los cambios en los costes se pueden apreciar a continuación (ver Tabla 4)

	Simples	*Medias*	*Complejas*
Documentar la solución de un error de informe de OBIEE	15 min	2h y 24 min	4h
Documentar la modificación de un *workflow*	30 minutos	45 min	4h y 48 min
Documentar el desarrollo de una consulta	1h	2h	2h

Tabla 4: Coste temporal de distintas tareas con EDC

  - Como se puede observar, las tareas de documentar una consulta no varían, esto es porque se trata de lenguaje SQL y la herramienta, a día de hoy, no está capacitada para interpretar código.

- **Mejora del rendimiento del proyecto:** Hay tareas donde son necesarias hasta 60 horas de trabajo y otras con las que cuatro horas son suficientes, pero en ambos casos las funciones de EDC reducirían notablemente el tiempo empleado. No lo haría en cuanto a desarrollo se refiere, sino que reduciría en gran medida el tiempo que se necesita para entender el contexto de la tarea, las repercusiones de la misma y los riesgos que conlleva.
- **Mejora de la rentabilidad del proyecto:** Se ha de tener en cuenta que los desarrolladores se apoyan en los analistas en la gran mayoría de tareas que realizan. En la gran parte de los casos es para tener una visión global de la tarea, sus repercusiones y relaciones de la misma, por tanto, EDC ayuda a liberar a los analistas y así poder dedicar ese tiempo en las tareas más complejas, aumentando el rendimiento sin aumentar el coste de personal.

En definitiva, EDC consigue que el rendimiento del proyecto Alumbra mejore, pudiendo realizar muchas más tareas en un mismo periodo de tiempo. Los responsables técnicos de Alumbra han valorado que la adquisición de la licencia es rentable para el proyecto, ya que no necesitarían aumentar el coste humano para mejorar el rendimiento del proyecto.

6 Conclusiones

A continuación, se exponen las conclusiones a las que se ha llegado con el trabajo realizado, tanto profesionales como personales.

6.1 Conclusiones del estudio

Gracias a las herramientas de gestión de metadatos los proyectos con un volumen de datos muy grande han mejorado el rendimiento de su gestión. En contra de los que se pensó en el inicio, la función con más repercusión para el proyecto Alumbra es la de linaje e impacto. Si se separa esta funcionalidad de la documentación automática, la primera proporciona un valor mucho mayor para el proyecto teniendo en cuenta que el desarrollo de las tareas consume mucho más tiempo que las tareas de documentación.

En cuanto al estudio de la herramienta se puede concluir que sería una buena aportación para el proyecto Alumbra ya que ayudaría a desarrolladores y analistas a documentar sus taras (documentación automática), evitar realizar trabajo innecesario (diccionario de datos y análisis de la calidad del dato) y sobre todo a prevenir errores y agilizar el trabajo (linaje e impacto).

El mayor inconveniente de la integración de EDC en el proyecto Alumbra es la propia integración. Algo a lo que no se le debe dar gran importancia, ya que una vez los arquitectos de PWC, OBIEE y EDC la integren en el sistema los problemas de integración no volverán a aparecer, al menos en un largo periodo de tiempo (por ejemplo, en una migración dentro de tres años).

Algo a tener en cuenta es que la herramienta también proporciona otras funcionalidades no tan interesantes a corto plazo. Funciones que requieren invertir demasiado tiempo y que por ahora no son útiles para el proyecto, pero que en un futuro pueden aportar un valor positivo para Alumbra.

6.2 Conclusiones personales

El resultado final de mi trabajo ha servido para ayudar a tomar una decisión muy importante dentro del organismo técnico de la *Consellería de Sanitat Universal i Salut Pública*, con lo que me siento muy satisfecho con el estudio realizado.

Trabajar en un proyecto tan importante para la sociedad de la Comunidad Valenciana ha supuesto una motivación extra para encarar las prácticas con mucho entusiasmo e intentar aprender todo lo posible.

He podido ver cómo funciona una empresa de consultoría como es *Capgemini* y como se trabaja en proyectos de 30 personas, donde el trabajo en equipo, la comunicación y una atmosfera de trabajo positiva son clave para el buen desarrollo del proyecto.

Gracias al trabajo realizado en el estudio he podido desarrollar y mejorar competencias transversales cómo el conocimiento de problemas contemporáneos, la comunicación efectiva, la gestión y organización del tiempo y sobre todo el trabajo en equipo, el análisis y resolución de problemas y el uso de instrumental específica.

Mi valoración final de las prácticas es, sin lugar a dudas, positiva. He tenido una buena formación que me ha servido para seguir completando mis estudios, así como para ganar experiencia en el mundo laboral.

6.3 Trabajos futuros

Una vez tomada la decisión de adquirir la herramienta, el jefe de proyecto ha organizado cursos de adaptación para los analistas del proyecto, con el objetivo de que estos enseñen lo aprendido a los desarrolladores a su cargo, y así conseguir que todo el equipo adquiera los conocimientos necesarios para poder sacar provecho a EDC.

A su vez, la negociación entre el organismo técnico responsable de la toma de decisiones por parte de la Generalitat Valenciana sigue negociando las condiciones con la empresa proveedora *Informatica.*

Por último, los arquitectos de OBIEE y PWC han sido notificados para que vayan preparando el entorno de las herramientas para la integración de EDC en el proyecto.

7 Bibliografía

[1] R.Ferrando, «Implantación del programa Orion Logis en el servicio de farmacia: preguntas y respuestas,» SEFAF, Valencia, 2012.

[2] España, *Ley orgánica 15/1999, de 13 de Diciembre, de Protección de Datos de Carácter Personal,* Boletín Oficial del Estado num. 298, de 14 de diciembre, 1999.

[3] *Plan de calidad de los procesos desarrollados con PowerCenter,* Valencia: Sistema de Información de la Generalitat Valenciana, 2010.

[4] J. Nielsen, *Heuristic evaluation of user interfaces,* Seattle: ACM, 1990.

[5] *Memoria de Gestión,* Valencia: Consellería de Sanitat i Salut Pública, 2016.

[6] Informatica, *Enterprise Information Catalog User Guide,* CA, 2016.

[7] Informatica, *Enteprise Infromation Data Catalog Data Sheet,* CA, 2016.

[8] Informatica, *Axon User Guide 5.0 HotFix 1,* CA, 2017.

[9] Informatica, *Axon Data Sheet,* CA, 2017.

[10] Informatica, *Metadata Manager User Guide,* CA, 2015.

[11] Informatia, *Metadata Manager Data Sheet,* CA, 2012.

[12] M. Rittman, *Integración de Oracle Hyperion Essbase - System 9 con Oracle Business Intelligence,* 2008.

[13] R. Kimball, *The Data Warehouse ETL Toolkit,* Indianapolis, 2004.

[14] S. K. Singh, *Database System: Concepts, Design and Applications,* Noida, 2011.