Caso práctico: almacén de datos para el análisis del impacto conductual de la COVID-19 sobre la población

Autores: Aitor Cabero Couto
David Díaz Arias
Carles Llorach i Rius
Daniel Miró Pettican
Víctor Ruíz Marqués

Índice

Introducción

El caso «Almacén de datos para el análisis del impacto conductual de la COVID-19 sobre la población» está creado para practicar el diseño y la implementación del almacén de datos como sistema de almacenamiento para el análisis de datos.

El diseño, el desarrollo y la implantación de un sistema de *data warehouse* (DW) en cualquier organización supone llevar a cabo un proyecto que puede durar meses o incluso años, en función del alcance del proyecto, de la naturaleza y del grado de madurez de la organización. También depende de la participación de equipos multidisciplinares que van implementando diferentes proyectos en un proceso de mejora continua del almacén.

El objetivo de este caso no es desarrollar un almacén de datos que dé respuesta a todas las necesidades, sino entender y utilizar las metodologías para desarrollar este tipo de proyectos en un contexto real. Las fases que comprenden los proyectos de esta tipología son las siguientes:

1. **Análisis, diseño e implementación:** consiste en desarrollar e implementar un almacén de datos que permita la gestión de la información disponible.

2. **Carga:** implica diseñar e implementar los procesos de carga de datos necesarios para disponer de información en el almacén de datos implementado en la etapa anterior.

3. **Explotación:** pretende explotar, mediante la generación de informes, los datos previamente cargados en el almacén.

Con el fin de poder desarrollar un proyecto lo más específico posible, el estudiante tendrá que afrontar el reto de construir un almacén de datos que solo describa parte de los servicios que se pueden ofrecer, basándose en los datos tratados en el caso y que formarían parte de un sistema real.

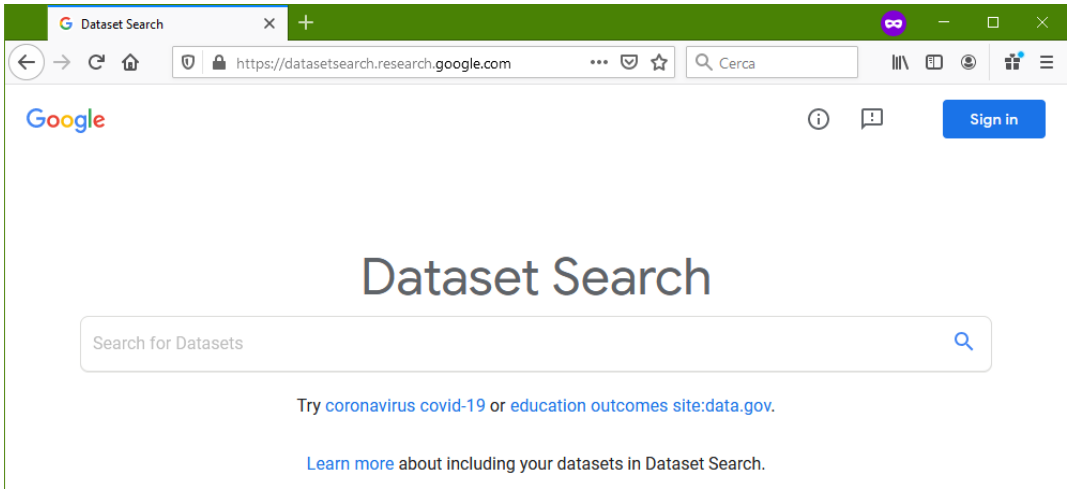
A partir del contexto que se describe a continuación, el estudiante deberá adquirir un conocimiento básico del entorno tecnológico, detectar las necesidades existentes y definir una propuesta adecuada que responda a ellas.

Mediante el desarrollo del caso, el estudiante se va a encontrar con los problemas, las dudas y las dificultades que se plantean en un proyecto de estas características.

Contexto

Nos encontramos ante una explosión de recursos *open data* a nivel global y es necesario comprender cuáles son las posibilidades reales de estos y su capacidad de interrelación con otras fuentes y herramientas disponibles de manera libre.

En marzo de 2020, Google da el impulso definitivo al *open data* al publicar veinticinco millones de *datasets* gratuitos, no solo limitados a datos estructurados en un fichero formateado, sino también a documentos, cartografía o imágenes.



Según el informe del Portal Europeo de Datos, España ocupa la segunda plaza en el ranquin de países europeos con mayor desarrollo del *open data*. Varios organismos locales y autonómicos ya han desarrollado sus propias iniciativas y, desde el ámbito privado, múltiples empresas han publicado directorios de datos.

Asimismo, en diciembre de 2019, un extraño virus, el SARS-CoV-2, aparece en la ciudad china de Wuhan y, tan solo tres meses más tarde, pone en jaque al mundo entero. Las consecuencias son miles de contagiados y fallecidos, hospitales desbordados, supermercados desabastecidos y economías colapsadas.

Se produce así una segunda explosión de datos relacionados con la COVID-19 sin precedentes que aumenta de forma exponencial cada quince días. Por ejemplo, el Instituto Nacional de Estadística (www.ine.es) adquirió de los principales operadores de telecomunicaciones la información sobre cómo se movía la población durante el confinamiento (https://www.ine.es/covid/exp_movilidad_covid_proyecto.pdf). Estos datos están disponibles como *open data*:

https://www.ine.es/covid/covid_movilidad.htm#tablas_resultados

Usuarios potenciales

Como fase inicial del diseño del sistema de análisis de datos COVID-19 identificaremos los requerimientos de los usuarios potenciales. De este modo el sistema los podrá tener en cuenta al dar respuesta a sus necesidades y generar información que les pueda ser útil.

Los usuarios finales que harán uso del sistema son los siguientes:

- Las **administraciones**. Con la información proporcionada por el sistema integrado, los gobiernos y los ayuntamientos dispondrán de la información de soporte para elegir las distintas medidas, controlar el impacto de la movilidad por zonas, registrar las llamadas de emergencia al 112, implementar servicios adicionales innovadores, establecer las medidas reguladoras que estimen oportunas, y mucho más.

- Las **empresas** y **organizaciones**. El sistema integrado de datos les permitirá extraer información útil relativa a las características conductuales de la población en su ámbito territorial. Además, contribuye a mejorar la calidad de sus servicios, dado que tendrán un conocimiento que les permitirá una mejor respuesta ante los cambios. Así podrán realizar comparativas y tomar decisiones comerciales mejor orientadas.

- Los **medios de comunicación**. Con la información del sistema integrado podrían disponer de información oficial para generar contenidos de calidad.

- La **población** en general. Esta puede consultar los datos y valorar la eficacia de las políticas aplicadas, el acierto de las iniciativas comerciales, la constatación de ciertos comportamientos colectivos, etc.

Fuentes de datos

Uno de los objetivos de este caso de estudio es integrar las diversas fuentes de datos (y formatos) proporcionadas para poder realizar diferentes tipos de análisis. En concreto, disponemos de información detallada de la población, la movilidad, las denuncias, las llamadas de emergencia y los datos para evitar aglomeraciones.

La relación de ficheros *open data* que utilizaremos para la carga inicial es la siguiente:

Nombre del fichero	Descripción	Fuente
ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx (específicamente la hoja «Datos_tratados»)	Estadística sobre los expedientes incoados por el artículo 36.6 LOPSC de desobediencia durante el estado de emergencia sanitaria de la COVID-19 en la comunidad de Euskadi	Gobierno vasco https://www.euskadi.eus/gobierno-vasco/-/infracciones-y-sanciones-impuestas-covid-19/
poblacion_9687bsc.csv	Cifras de la población española por provincia	www.ine.es https://www.ine.es/jaxiT3/Tabla.htm?t=9687
rows.xml	Llamadas al 112 por ámbito geográfico y tipología (accidentes de tráfico, civismo, incendios, asistencia sanitaria, seguridad...)	CAT112 https://www.europeandataportal.eu/data/datasets/https-analisi-transparenciacatalunya-cat-api-views-mfqb-sbx4?locale=en
35167bsc.csv	Movilidad de la población durante el estado de alarma https://www.ine.es/experimental/movilidad/experimental_em.htm	www.ine.es https://www.ine.es/covid/covid_movilidad.htm#tablas_resultados
statistic_id1104235_covid-19_-poblacion-que-evitaba-las-aglomeraciones-segun-edad-en-espana-2020.xlsx (específicamente la hoja "Datos_provincias")	Porcentaje de la población que evitaba las aglomeraciones con motivo del coronavirus, por grupo de edad y provincia	Statista https://es.statista.com/estadisticas/1104235/poblacion-que-evitaba-las-aglomeraciones-debido-al-covid-19-segun-edad-en-espana/

Se constata que los datos de las llamadas de emergencia y de la población se recibirán anualmente y, por tanto, serán necesarias las cargas incrementales para su integración en el *data warehouse*. El desarrollo de estos procesos futuros queda fuera del alcance de esta actividad.

Enunciado

1. PRA1: análisis y diseño del *data warehouse*

A partir del análisis del contexto del caso y de las fuentes de datos disponibles, el estudiante deberá diseñar y proponer un almacén de datos que permita y facilite el análisis del impacto conductual de la COVD-19 sobre la población.

A partir de la metodología de diseño de un *data warehouse* propuesta en la asignatura, el estudiante debe llevar a cabo lo siguiente:

- El **análisis de los requerimientos:** como resultado, se generará un documento que describa las preguntas para las que los usuarios potenciales esperan una respuesta del sistema.
- El **análisis de las fuentes de datos:** se deben revisar las fuentes de datos proporcionadas, qué tipo de información contienen, cuál es su formato y qué cantidad representan para la carga inicial.
- El **análisis funcional:** se debe proponer el tipo de arquitectura para la factoría de información que mejor se adecue al proyecto (por ejemplo, si es necesario un *data mart* operacional o una estructura de carga intermedia).
- El **diseño del modelo conceptual, lógico y físico del almacén de datos:** se deben identificar y diseñar las tablas de hechos, las dimensiones y los atributos que describen la información.

Para este apartado, el estudiante debe preparar un documento (solución PRA1) en el que se detalle cada uno de los apartados anteriores.

Se deberá tener en cuenta que para el desarrollo del DW es preciso definir correctamente los hechos (*facts*), las dimensiones del análisis (*dimensions*) y los atributos que nos permitan tener el nivel de granularidad suficiente para la medida y la presentación de los objetivos que se definan en el análisis de los requerimientos.

2. PRA2: carga de los datos

A partir de la solución oficial de la primera práctica (PRA1), el estudiante debe diseñar, implementar y ejecutar los procesos de extracción, transformación y carga de los datos (ETL) de las fuentes de datos proporcionadas.

Veamos las tareas que debe llevar a cabo el estudiante:

- Identificación de los procesos de extracción, transformación y carga de los datos (ETL) hacia el almacén de datos.
- Diseño y desarrollo de los procesos ETL mediante las herramientas de diseño proporcionadas.
- Implementación con los trabajos de los procesos ETL para su carga efectiva planificada.

3. PRA3: explotación de los datos

Tras la carga efectiva de estos en el almacén de datos (PRA2), se debe implementar un cubo multidimensional para explotar la información como apoyo a la toma de decisiones de los usuarios potenciales. La finalidad del diseño del modelo *multidimensional online analytical processing* (MOLAP) será responder a las preguntas definidas en el análisis de requerimientos.

Programas

Para este caso, la UOC proporciona un entorno VDI con todo el software preconfigurado con las siguientes características:

- Sistema operativo: Windows 10
- Base de datos: base de datos remota Microsoft SQL Server 2016 accesible desde cliente mediante SQL Server Management Studio 17
- Herramienta para la creación de cubos OLAP: Visual Studio 2017
- Herramienta de diseño de ETL: Spoon–Pentaho Data Integration 9.0
- Herramienta de creación de informes: PowerBI Desktop

Bibliografía

Material de la asignatura *Data Warehouse* de la UOC:

Kimball, R. (2013). *The Data Warehouse Toolkit* (3.ª ed.). Nueva York: John Wiley & Sons Inc.

Krishnan, K. (2013). *Data Warehousing in the Age of Big Data. The Morgan Kaufmann Series on Business Intelligence*.

Inmon W. H.; Imhoff C.; Sousa R. (1998). *Corporate Information Factory*. EE. UU.: John Wiley & Sons Inc.

Inmon W. H. (1996). *Building the Data Warehouse* (2.ª Ed.). EEUU: John Wiley & Sons Inc.

Inmon, W.H.; Strauss, D.; Neushloss, G. (2008). *DW 2.0: The Architecture for next generation of Data Warehousing*. EE. UU.: Morgan Kaufman Series.

Enlaces a internet:

MSDN Analysis Services tutorial:

<https://docs.microsoft.com/es-es/analysis-services/analysis-services-tutorials-ssas?view=asallproducts-allversions>

Tutorial Pentaho Data Integration:

<http://wiki.pentaho.com/display/EAI/Latest+Pentaho+Data+Integration+%28aka+Kettle%29+Documentation>