

**Máster universitario de Ciencia de Datos**

**Práctica 1**

**Diseño y uso de bases de datos analíticas – análisis y diseño del *data warehouse* (análisis de los requerimientos, análisis de las fuentes de datos, análisis funcional y diseño del modelo multidimensional)**

Autor:

Mario Ubierna San Mamés

|  |
| --- |
| Índice de Contenido |

[Índice de Contenido 3](#_Toc68259567)

[Índice de tablas 4](#_Toc68259568)

[Índice de ilustraciones 5](#_Toc68259569)

[1. Introducción 6](#_Toc68259570)

[2. Contexto 8](#_Toc68259571)

[3. Usuarios potenciales 9](#_Toc68259572)

[4. Fuentes de datos 10](#_Toc68259573)

[5. PRA1 – Análisis y diseño del modelo 12](#_Toc68259574)

[6. Bibliografía 13](#_Toc68259575)

|  |
| --- |
| Índice de tablas |

[Tabla 1 - Fuentes de datos. 11](#_Toc68259126)

|  |
| --- |
| Índice de ilustraciones |

**No se encuentran elementos de tabla de ilustraciones.**

|  |
| --- |
| 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* 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 [1] adquirió de los principales operadores de telecomunicaciones la información sobre cómo se movía la población durante el confinamiento [2]. Estos datos están disponibles como *open data* [3].

|  |
| --- |
| 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 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 [4]. |
| poblacion\_9687bsc.csv | Cifras de la población española por provincia. | [www.ine.es](http://www.ine.es) [5]. |
| rows.xml | Llamadas al 112 por ámbito geográfico y tipología (accidentes de tráfico, civismo, incendios, asistencia sanitaria, seguridad...). | CAT112 [6]. |
| 35167bsc.csv | Movilidad de la población durante el estado de alarma. | [www.ine.es](http://www.ine.es) [3]. |
| statistic\_id1104235\_covid19\_-poblacion-que-evitabalas-aglomeraciones-segunedad-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 [7]. |

Tabla 1 - Fuentes de datos.

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.

|  |
| --- |
| PRA1 – Análisis y diseño del modelo |

## Análisis de los requerimientos

aa

|  |
| --- |
| Bibliografía |

[1] «INE. Instituto Nacional de Estadística», *INE*. https://www.ine.es/ (accedido abr. 02, 2021).

[2] «exp\_movilidad\_covid\_proyecto.pdf». Accedido: abr. 02, 2021. [En línea]. Disponible en: https://www.ine.es/covid/exp\_movilidad\_covid\_proyecto.pdf.

[3] «Evolución de la movilidad por ámbito geográfico durante el estado de alarma por COVID-19», *INE*. https://www.ine.es/covid/covid\_movilidad.htm#tablas\_resultados (accedido abr. 02, 2021).

[4] «Infracciones y Sanciones Impuestas (COVID-19) - Gobierno Vasco - Euskadi.eus». https://www.euskadi.eus/gobierno-vasco/-/infracciones-y-sanciones-impuestas-covid-19/ (accedido abr. 02, 2021).

[5] «Población residente por fecha, sexo y edad(9687)», *INE*. https://www.ine.es/jaxiT3/Tabla.htm?t=9687 (accedido abr. 02, 2021).

[6] «Datasets - European Data Portal». https://www.europeandataportal.eu/data/datasets?locale=en&minScoring=0 (accedido abr. 02, 2021).

[7] «COVID-19: evitación de las aglomeraciones por edad España en 2020», *Statista*. https://es.statista.com/estadisticas/1104235/poblacion-que-evitaba-las-aglomeraciones-debido-al-covid-19-segun-edad-en-espana/ (accedido abr. 02, 2021).