

**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](#_Toc68611706)

[Índice de tablas 4](#_Toc68611707)

[Índice de ilustraciones 5](#_Toc68611708)

[1. Introducción 6](#_Toc68611709)

[2. Contexto 8](#_Toc68611710)

[3. Usuarios potenciales 9](#_Toc68611711)

[4. Fuentes de datos 10](#_Toc68611712)

[5. PRA1 – Análisis y diseño del modelo 12](#_Toc68611713)

[5.1. Análisis de los requerimientos 12](#_Toc68611714)

[5.2. Análisis de fuentes de datos 13](#_Toc68611715)

[5.2.1. Ficheros CSV 13](#_Toc68611716)

[5.2.2. Ficheros Excel 15](#_Toc68611717)

[5.2.3. Ficheros XML 16](#_Toc68611718)

[5.2.4. Estimación de la volumetría 17](#_Toc68611719)

[5.3. Análisis funcional 18](#_Toc68611720)

[6. Bibliografía 22](#_Toc68611721)

|  |
| --- |
| Índice de tablas |

[Tabla 1 - Fuentes de datos. 11](#_Toc68611722)

[Tabla 2 - Campos del fichero 35167bsc.csv 14](#_Toc68611723)

[Tabla 3 - Campos del fichero poblacion\_9687bsc.csv 14](#_Toc68611724)

[Tabla 4 - Campos del fichero ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx 15](#_Toc68611725)

[Tabla 5 - Campos del fichero statistic\_id1104235\_covid-19\_-poblacion-que-evitaba-las-aglomeraciones-segun-edad-en-espana-2020.xlsx 16](#_Toc68611726)

[Tabla 6 - Campos del fichero rows.xml 17](#_Toc68611727)

[Tabla 7 - Volumetría de los datos. 18](#_Toc68611728)

[Tabla 8 - Requerimientos de la factoría de información. 19](#_Toc68611729)

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

[Ilustración 1 - Arquitectura de la FIC. 20](#_Toc68611730)

|  |
| --- |
| 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 - 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

El análisis de los requerimientos tiene como objetivo el saber cuáles son las necesidades u objetivos que tiene una empresa/organización respecto al análisis de información. Cabe destacar que esta información tiene que estar integrada en un sistema que permita dar respuestas a los analistas, con el fin de mejorar la toma de decisiones.

Respecto al contexto de este problema, partimos de que a finales del 2019 apareció una nueva mutación de la familia de coronavirus, llamado *SARS-CoV-2*, el cual ha dejado un gran número de fallecidos en todo el mundo. Sin contar los hospitales desbordados, una cuarentena global, economías paralizadas…

Una vez que hemos analizado el contexto, podemos definir las necesidades que se nos presentan:

* Conocer la evolución de las diferentes infracciones/denuncias por desobediencia durante el estado de emergencia en la comunidad del País Vasco.
* Analizar la evolución de las llamadas al 112 producidas en la comunidad de Cataluña.
* Se tiene que poder realizar al análisis del punto anterior desde diferentes perspectivas: ámbito geográfico y tipología.
* Movilidad de la población española durante el estado de alarma, en términos de características como la edad y la provincia.

Una vez que tenemos claro las necesidades que se tienen que cubrir, podemos plantearnos diversas preguntas, algunas de ella son:

* Evolución de las infracciones durante el estado de emergencia en el país vasco.
* Top de las fechas con más infracciones cometidas en el país vasco por provincia.
* Población por provincia.
* Top de provincias con más población.
* Evolución del número de llamadas al 112 por ámbito geográfico y tipología.
* Fechas en las que se ha llamado más al 112 según su ámbito geográfico y tipología.
* Evolución de la media a nivel nacional del porcentaje movilidad de la población durante el estado de alarma.
* Mayor porcentaje de movilidad de la población por provincia durante el estado de alarma.
* Media del porcentaje de población a nivel nacional que evitaba aglomeraciones por edad.
* Top provincias cuya población evitaba el mayor número de aglomeraciones por edad.

## Análisis de fuentes de datos

Respecto al análisis de fuentes de datos, tenemos como objetivo en revisar las fuentes de datos proporcionadas e identificar el tipo de información que contienen, su formato y volumetría.

### Ficheros CSV

El formato de los ficheros CSV es el siguiente:

* Primera línea con los campos del dataset.
* El separador de campos es el punto y como “;”.
* El separador decimal es la coma “,”.

**35167bsc.csv**: recoge la movilidad de la población durante el estado de alarma. Este fichero presenta los siguientes campos:

|  |  |  |
| --- | --- | --- |
| **Nombre del campo** | **Tipo** | **Ejemplo** |
| Zonas de movilidad | Text | Almería |
| Periodo | Date | 20/6/2020 |
| Total | Number | 15,15 |

Tabla - Campos del fichero 35167bsc.csv

Podemos observar, que dicho dataset tiene un registro por cada provincia y por cada fecha durante todo el estado de alarma.

Por otro lado, tenemos 4.732 observaciones y 3 atributos.

**poblacion\_9687bsc:** contiene las cifras de la población española por provincia.

|  |  |  |
| --- | --- | --- |
| **Nombre del campo** | **Tipo** | **Ejemplo** |
| Edad Simple | Text | Total |
| Provincias | Text | 02 Albacete |
| Sexo | Text | Ambos sexos |
| Periodo | Text | 1 de enero de 2020 |
| Total | Number | 389.830 |

Tabla - Campos del fichero poblacion\_9687bsc.csv

Podemos observar, que el campo “Edad simple” siempre tiene el mismo valor “Total”, el campo “Sexo” también tiene siempre el mismo valor “Ambos sexos” y el campo “Periodo” siempre tiene el valor “1 de enero de 2020”.

Por otro lado, tenemos 52 observaciones y 5 atributos.

### Ficheros Excel

**ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx:** 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.

Observamos que hay diferentes hoja, pero para el estudio de esta práctica solo vamos a tener en cuenta la hoja “Datos\_tratados”.

|  |  |  |
| --- | --- | --- |
| **Nombre del campo** | **Tipo** | **Ejemplo** |
| TT.HH | Text | ARABA |
| ERTZAINTZA IDENTIFICADOS | Number | 10.717 |
| ERTZAINTZA DETENIDOS | Number | 40 |
| ERTZAINTZA DENUNCIAS INTERPUESTAS | Number | 2.586 |
| ERTZAINTZA VEHÍCULOS INTERCEPTADOS | Number | 15.182 |
| PP.LL IDENTIFICADOS | Number | 21.599 |
| PP.LL DETENIDOS | Number | 29 |
| PP.LL DENUNCIAS INTERPUESTAS | Number | 2.748 |
| PP.LL VEHÍCULOS INTERCEPTADOS | Number | 19.250 |
| FECHA FINAL | Date | 18/06/2020 |

Tabla - Campos del fichero ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx

Podemos observar que el campo “TT.HH” se repiten siempre tres valores (“ARABA”, “BIZKAIA”, “GIPUZKOA”).

Por otro lado, tenemos 218 observaciones y 10 atributos.

**statistic\_id1104235\_covid-19\_-poblacion-que-evitaba-las-aglomeraciones-segun-edad-en-espana-2020.xlsx:** porcentaje de la población que evitaba las aglomeraciones con motivo del coronavirus, por grupo de edad y provincia.

Observamos que hay diferentes hojas, pero para el estudio de esta práctica vamos a tener en cuenta solo la hoja “Datos\_provincias”.

|  |  |  |
| --- | --- | --- |
| **Nombre del campo** | **Tipo** | **Ejemplo** |
| Provincia | Text | Álava (Araba) (País Vasco) |
| 14 - 24 | Number | 36,99 |
| 25 - 34 | Number | 52,46 |
| 35 - 44 | Number | 42,70 |
| 45 - 54 | Number | 46,66 |
| 55 - 64 | Number | 35,73 |
| > 64 | Number | 51,13 |

Tabla - Campos del fichero statistic\_id1104235\_covid-19\_-poblacion-que-evitaba-las-aglomeraciones-segun-edad-en-espana-2020.xlsx

Podemos observar, que los atributos están divididos por el rango de edad, además, en el atributo provincia encontramos también la comunidad autónoma a la que pertenece.

Por otro lado, tenemos 49 observaciones y 7 atributos.

### Ficheros XML

**rows:** Llamadas al 112 por ámbito geográfico y tipología (accidentes de tráfico, civismo, incendios, asistencia sanitaria, seguridad…)

|  |  |  |
| --- | --- | --- |
| **Nombre del campo** | **Tipo** | **Ejemplo** |
| Any | Number | 2014 |
| Mes | Number | 1 |
| Provincia | Text | BARCELONA |
| Comarca | Text | ALT PENEDES |
| Municipi | Text | AVINYONET DEL PENEDES |
| Tipus | Text | Seguretat |
| trucades | Number | 10 |

Tabla - Campos del fichero rows.xml

Tenemos 340307 observaciones y 7 atributos.

### Estimación de la volumetría

La estimación respecto al número de datos que debe contener el almacén de datos es la siguiente:

|  |  |
| --- | --- |
| **Fuente de datos** | **Datos** |
| 35167bsc.csv | (4.732 observaciones +  1 registro de columnas) \*  3 atributos =  14.199 datos |
| poblacion\_9687bsc.csv | (52 observaciones +  1 registro de columnas) \*  5 atributos =  265 datos |
| ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx | (218 observaciones +  1 registro de columnas) \*  10 atributos =  2.190 datos |
| statistic\_id1104235\_covid-19\_-poblacion-que-evitaba-las-aglomeraciones-segun-edad-en-espana-2020.xlsx | (49 observaciones +  1 registro de columnas) \*  7 atributos =  350 datos |
| rows.xml | 340.307 observaciones \*  7 atributos =  2.382.149 datos |
| **TOTAL** | **2.399.153 datos** |

Tabla - Volumetría de los datos.

## Análisis funcional

Cuando hablamos del análisis funcional nos referimos al tipo de arquitectura que debe de tener la factoría de información. Para poder realizar un buen análisis y diseño tenemos que definir los requisitos funcionales, a los cuales hay que establecer una prioridad “E” si es exigible o “D” si es deseable.

Para poder saber qué requisitos son los que tenemos que hacer frente en la construcción de la factoría de información, debemos analizar el enunciado de la práctica, junto con su contexto y los usuarios potenciales.

Por otro lado, para determinar cómo es de prioritario un requisito definimos una prioridad del 1 al 3, siendo 1 completamente prioritario y 3 no prioritario. En la siguiente tabla, se describe cada uno de los requisitos junto con sus prioridades:

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Requerimiento** | **Prioridad** | **Exigible/Deseable** |
| 1 | Extraer de forma adecuada la información de las fuentes de datos (considerando solo la información relevante). | 1 | E |
| 2 | Crear un almacén de datos. | 1 | E |
| 3 | Cargar la información respecto al impacto conductual de la *COVID-19* sobre la población. | 1 | E |
| 4 | Crear un modelo OLAP para así poder realizar las diferentes consultas multidimensionales de los usuarios potenciales. | 2 | E |
| 5 | Crear cada uno de los diferentes informes para cada tipo de usuario. | 2 | E |
| 6 | Desarrollar las cargas incrementales para la integración en el *data warehouse*. | 3 | D |

Tabla - Requerimientos de la factoría de información.

Estos son los requisitos que hemos podido concluir según el enunciado del problema y su contexto, pero cabe destacar que podría haber más requisitos, como por ejemplo: creación de diferentes almacenes de datos departamentales, que todas las cargas de datos se realicen de forma automática, transformar los datos…

Una vez que tenemos claro los requisitos que necesitamos, en nuestro almacén de datos podemos definir la arquitectura del mismo:

* Respecto a las fuentes de datos, como bien se han explicado en el apartado anterior, vamos a tener ficheros csv, hojas de Excel y fichero xml.
* En cuanto a la necesidad de un *staging area*, en nuestro caso no es totalmente necesario, pero sí que es verdad que al tener diversas fuentes de datos y cada una tiene una estructura diferente, se podría hacer uso de un *staging area* para así definir una estructura intermedia entre las fuentes de datos y el almacén de datos. Además, podríamos en esta fase desarrollar procesos para mejorar la calidad de los datos.
* Respecto al almacén de datos en sí, realmente para este caso de estudio con un almacén de datos departamental enfocado al área de análisis del impacto de la *COVID-19* sería suficiente, ya que no hay más áreas de estudio (negocio, marketing… Todo está enfocado al mismo área de análisis), por lo que un almacén corporativo es innecesario (aunque si se quiere añadir se podría hacer). Finalmente, el almacén de datos operacional no tendría sentido en nuestra situación, ya que los datos se van a cargar anualmente y no hay un proceso de interacción continua con los usuarios/procesos, pero al igual que el almacén corporativo si se quiere añadir se podría hacer.
* Para concluir con la arquitectura de la factoría de información, necesitaríamos de un modelo multidimensional para poder responder a las preguntas de los analistas, y obtener informes con los datos que quieran. Es por ello, que necesitamos hacer uso de *MOLAP* para poder crear y consultar sobre los cubos multidimensionales.

En la siguiente imagen podemos aprecias cómo quedaría la arquitectura funcional de nuestro almacén de datos:

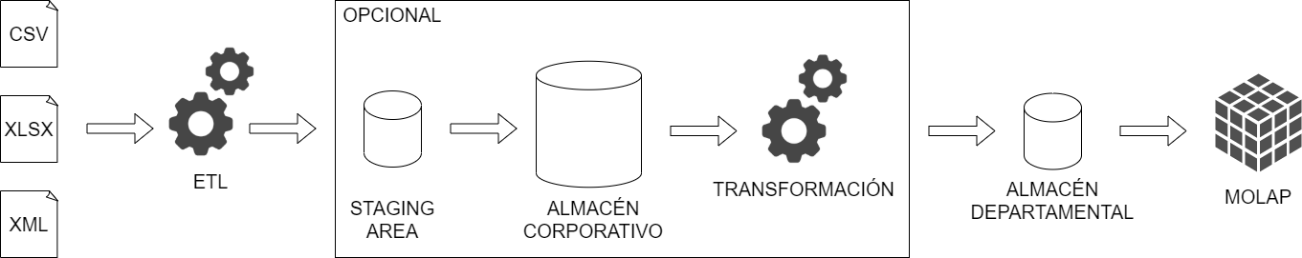


Ilustración - Arquitectura de la FIC.

Como podemos apreciar en la anterior ilustración, tenemos un apartado opcional correspondiente al *staging area* y al almacén corporativo. Esto quiere decir que se podría incluir en la arquitectura de la factoría de información, sin embargo, dado el caso de estudio que se nos ha proporcionado no podemos saber a ciencia cierta el alcance del proyecto.

Es por ello que cuánto más sencillo sea el diseño que se ajuste a nuestro caso, más fácil es de ampliar luego las características de nuestro almacén de datos. Cabe destacar, que crear *el stating area* sería interesante, ya que al tener diferentes fuentes de datos nos permitiría dar una misma estructura de carga inmediata a los mismos.

Para terminar, tal y como podemos apreciar solo hay un almacén departamental, se ha considerado que esta era la mejor opción, ya que no queda bien definido según el enunciado el cómo está formada la organización y cuál es el alcance del proyecto de forma detallada. Además, aunque los ficheros fuente proporcionan información diferente, todos están enfocados en la misma área temática, es decir, el análisis del impacto del *COVID-19*.

|  |
| --- |
| 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).