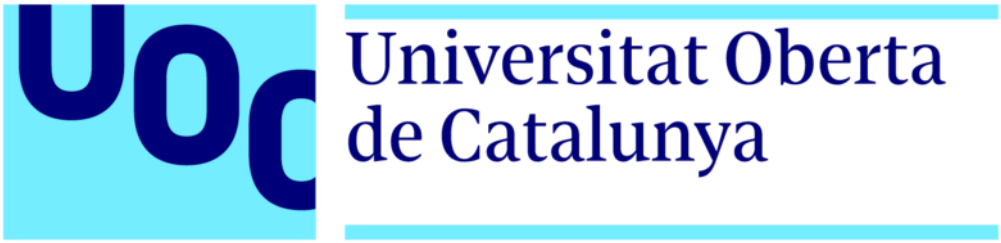

Máster universitario de Ciencia de Datos

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 tablas

Índice de ilustraciones

1. 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.

2. 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].

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.

4. 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 [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 [3].
statistic_id1104235_covid19_-poblacion-que-evitabalas-aglomeraciones-segunedad-en-espana-2020.xlsx	Porcentaje de la población que evitaba las aglomeraciones con motivo del coronavirus, por grupo de edad y provincia	Statista [7].

(específicamente la hoja “Datos_provincias”)		

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.

5. PRA1 – Análisis y diseño del modelo

5.1. 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.
- En qué fechas se han cometido más infracciones en el País Vasco.
- Número de habitantes por provincia.
- Evolución del número de llamadas al 112 por ámbito geográfico y tipología.
- Fechas en las que se ha llamado más al 112.
- Evolución del porcentaje movilidad de la población durante el estado de alarma.
- Media porcentaje por mes de movilidad de la población por provincia durante el estado de alarma.
- Porcentaje de población a nivel nacional que evitaba aglomeraciones por edad.
- Máximo porcentaje de la población que evitaba las aglomeraciones.

5.2. Análisis de fuentes de datos

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

5.2.1. 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 coma ";".
- 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 2 - 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 3 - 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.

5.2.2. 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 hojas, 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 4 - 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 5 - 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.

5.2.3. 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 6 - Campos del fichero rows.xml

Tenemos 340307 observaciones y 7 atributos.

5.2.4. 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-	(49 observaciones +

aglomeraciones-segun-edad-en-espana-2020.xlsx	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 7 - Volumetría de los datos.

5.3. 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 8 - 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:

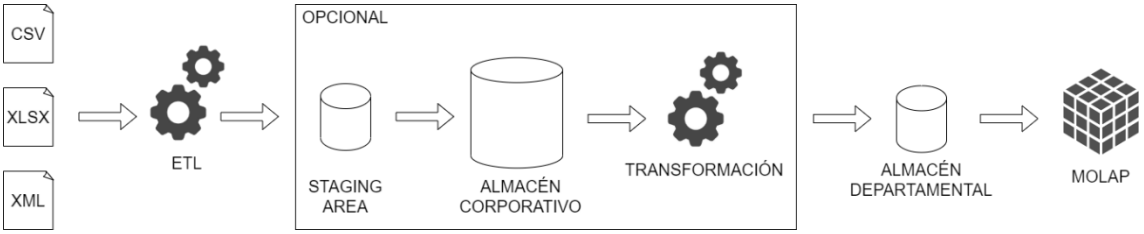

Ilustración 1 - Arquitectura de la FIC.

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*.

5.4. Diseño del modelo conceptual, lógico y físico del almacén de datos.

Para poder realizar un buen diseño del modelo multidimensional, tenemos que definir correctamente el modelo conceptual, lógico y físico de nuestro *data warehouse*.

Esto lo realizamos definiendo los hechos que tenemos en nuestro caso de estudio, para ello observamos las necesidades que tienen los usuarios potenciales, las fuentes de datos y los requisitos que hemos definidos en el análisis de los requerimientos. Tras todo ello hemos identificado los siguientes hechos:

- Las infracciones/desobediencias que se produjeron durante el estado de emergencia sanitaria en el País Vasco.
- La gestión de las llamadas al 112 en Cataluña.
- La población determinada por su provincia.
- La gestión de la movilidad de la población.

Todos estos hechos están basados en una serie de necesidades:

- Análisis de las llamadas al 112, según su tipología y ámbito geográfico.
- El análisis temporal de las infracciones/desobediencias en el País Vasco, según la provincia y el tipo de policía.
- Análisis de la población por provincia.
- Evolución del porcentaje de movilidad durante el estado de alarma, según la provincia.
- Análisis del porcentaje de población que evitaba aglomeraciones, según la provincia.

Diseño conceptual

Una vez que ya tenemos claro las necesidades que tiene nuestro sistema, vamos a desarrollar cada modelo para satisfacer las mismas. Mencionar que las estrellas que se van a definir a continuación, posteriormente se podría unir unas con otras para realizar determinadas operaciones, pero tal y como se ha definido en la práctica, se busca el diseño más simple y que mejor represente las necesidades y requerimientos definidos.

La gestión de las llamadas al 112 determina la primera tabla de hechos:

Hecho	Descripción
Fact_Llamadas112	Gestión de las llamadas al 112.

Tabla 9 - Tabla de Hechos para la gestión de llamadas al 112.

Esta tabla de hechos, además de tener la métrica número de llamadas al 112 según su tipología y ámbito geográfico, se podrá tener acceso al recuento de llamadas diarias.

En la siguiente tabla observamos las dimensiones para satisfacer esta necesidad:

Dimensión	Descripción
Dim_Tiempo	Análisis de las llamadas respecto a la fecha en las que se han producido.
Dim_Tipologia	Tipo/causa de la llamada al 112.
Dim_Lugar	Ámbito geográfico desde el que se han producido las llamadas.

Tabla 10 - Tabla de Dimensiones para la gestión de llamadas al 112.

Una vez que ya hemos definido las dimensiones y el hecho, realizamos el diseño conceptual de nuestra estrella:

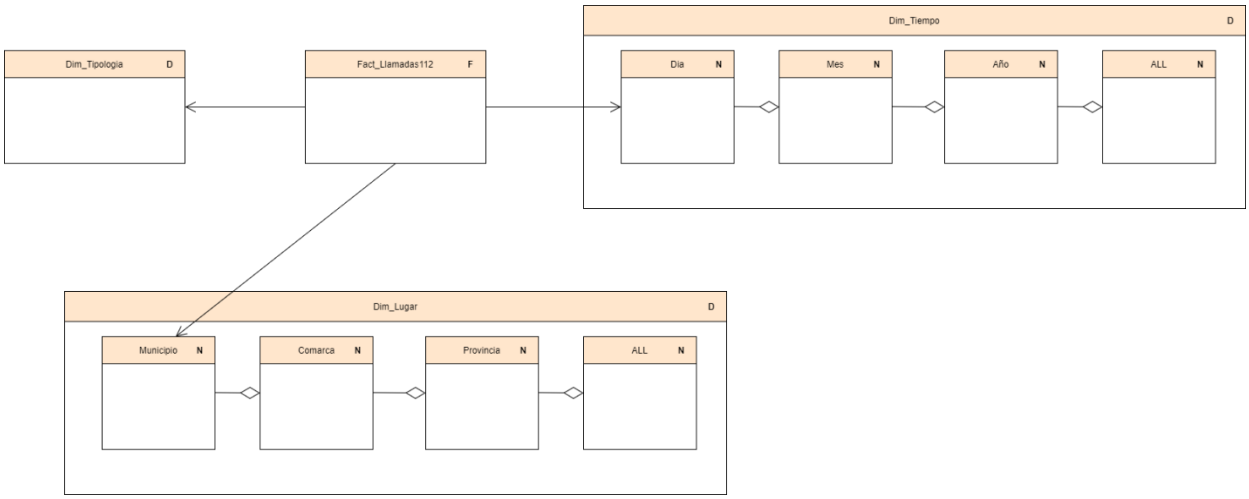

Ilustración 2 - Diseño conceptual de las llamadas al 112.

Tal y como podemos apreciar en la anterior ilustración, tanto la dimensión tiempo como la dimensión lugar presentan una jerarquía de agregación. Respecto al nivel “ALL” estamos haciendo referencia a la agrupación de todas las instancias de la dimensión al mismo tiempo (esto será así a lo largo de toda la práctica).

Otro aspecto a mencionar es que la dimensión tiempo presenta un nivel día, cuando en la fuente de datos relativa a las llamadas al 112 solo viene el nivel mes y año, sin embargo siguiendo los pasos de la teoría proporcionada, varias estrellas comparten esta misma dimensión, por lo que tenemos la misma dimensión pero cambiando la jerarquía de agregación. En este caso, el nivel día se podría tener en cuenta si indicamos que todas las fechas tienen el mismo día, o podríamos ignorar dicho nivel pasando directamente al mes. El resto de estrellas que comparten la misma dimensión de tiempo, sí que tienen todos los niveles de la jerarquía de agregación, es decir, el día, el mes y el año.

La gestión de las infracciones es el segundo hecho, y nos determina la siguiente tabla:

Hecho	Descripción
Fact_Infracciones_Evolucion	Recoge el número de desobediencias durante el estado de alarma.

Tabla 11 - Tabla de Hechos para la gestión de las infracciones.

En la siguiente tabla observamos las dimensiones para satisfacer dicha necesidad:

Dimensión	Descripción
Dim_Tiempo	Análisis de las desobediencias según la fecha en las que se produjeron.
Dim_Lugar	Análisis de las desobediencias según la provincia en las que se produjeron.
Dim_Policia	Análisis de las desobediencias según la policía que las notificó.

Tabla 12 - Tabla de Dimensiones para la gestión de las infracciones.

Una vez determinados el hecho y las dimensiones, observamos el siguiente diseño conceptual:

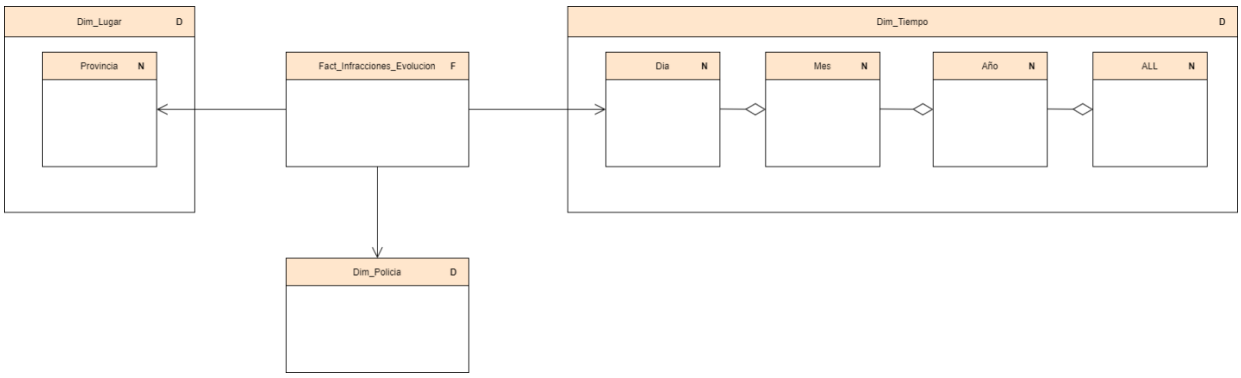

Ilustración 3 - Diseño conceptual de las infracciones.

Tal y como podemos apreciar en la anterior ilustración, mantenemos las dimensiones de tiempo y lugar como en el hecho anterior. Sin embargo, la dimensión lugar aunque es la misma dimensión ha cambiado su jerarquía de agregación, debido a que no existen los niveles municipio y comarca. Por otro lado, vemos que desapareció

el nivel “ALL” dentro de la dimensión lugar, esto se debe a que dicho nivel solamente nos está indicando la agrupación de todas las instancias de la dimensión al mismo tiempo, pero en este caso con el nivel provincia conseguimos lo mismo.

Al tener solo el nivel provincia podríamos considerar que ese nivel sea directamente la dimensión, pero para mantener un diseño de dimensiones hemos considerado que lo mejor sea seguir con el mismo diseño de las dimensiones comunes, ya que al final y al cabo son las mismas dimensiones.

El análisis de la población por provincia determina la tercera tabla de hechos, cabe mencionar que aunque este hecho por sí mismo no proporcione información sobre el impacto del *COVID-19*, relacionando esta estrella con otras nos pueden dar datos interesantes, sobre todo diferentes ratios respecto a la población por provincia:

Hecho	Descripción
Fact_Poblacion	Recoge el número de la población española por provincia

Tabla 13 - Tabla de Hechos para la población.

Esta estrella es muy simple ya que solo presentamos la siguiente dimensión:

Dimensión	Descripción
Dim_Lugar	Número de habitantes por provincia.

Tabla 14 - Tabla de Dimensiones para la población.

Por lo tanto, tendríamos el siguiente diseño conceptual:

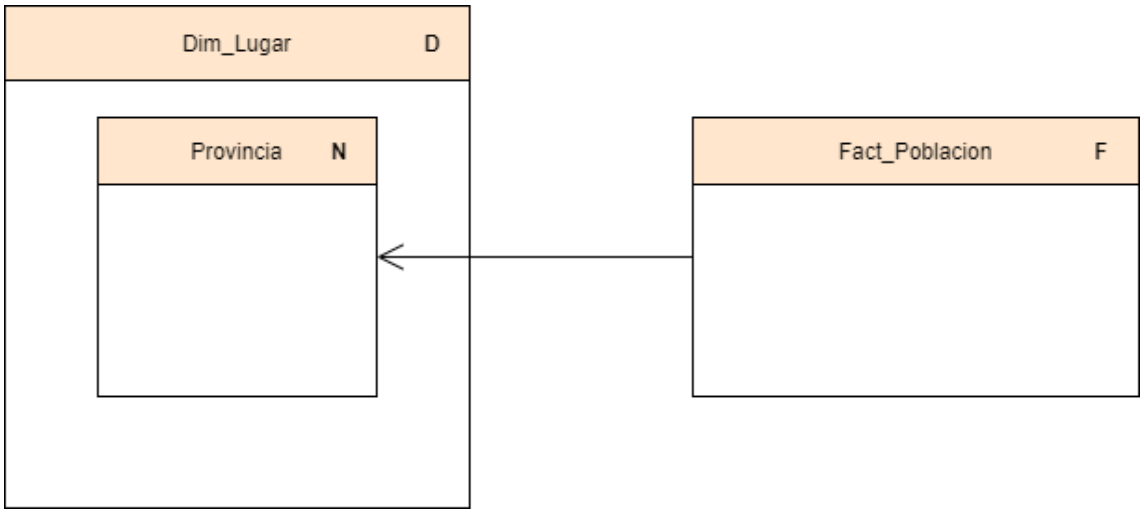

Ilustración 4 - Diseño conceptual del análisis de la población.

Respecto a la evolución del porcentaje de movilidad durante el estado de alarma obtenemos la siguiente tabla de hechos:

Hecho	Descripción
Fact_Movilidad	Evolución de la movilidad a partir del porcentaje de población que se movilizó.

Tabla 15 - Tabla de Hechos para la movilidad.

Una vez que tenemos claro el hecho, definimos las dimensiones:

Dimensión	Descripción
Dim_Tiempo	Porcentaje de población que se movilizó dada una fecha.
Dim_Lugar	Porcentaje de población que se movilizó dada su provincia.

Tabla 16 - Tabla de Dimensiones para la movilidad.

Como resultado de dicho análisis representamos el siguiente diseño conceptual:

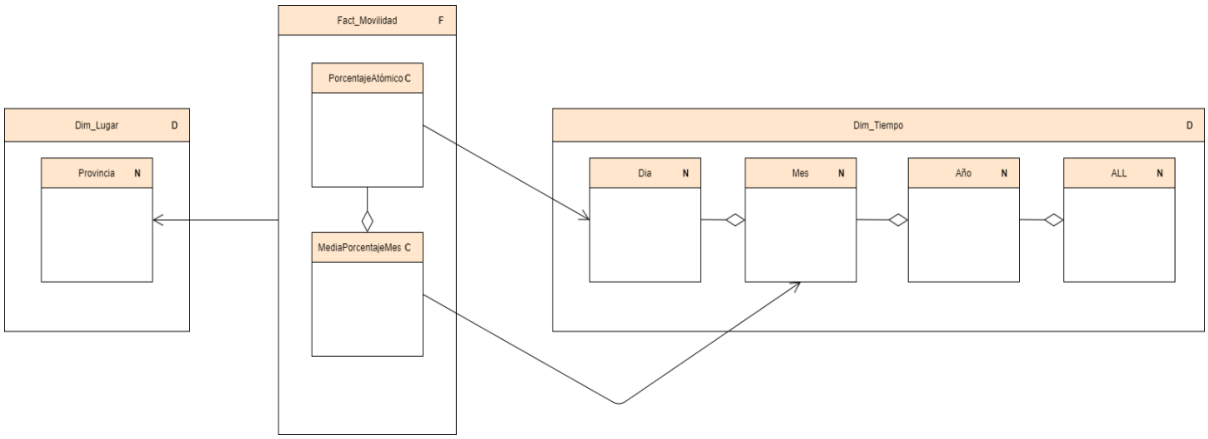

Ilustración 5 - Diseño conceptual de la evolución de la movilidad.

Como podemos apreciar en la anterior ilustración, el hecho ha cambiado un poco respecto a los anteriores, en este caso tenemos dos Celdas, esto es debido a que no es lo mismo el porcentaje por día que la media del porcentaje por mes, es decir, tenemos diferentes granularidades.

Hasta ahora no hemos tenido ese "problema", ya que para el hecho de las llamadas al 112 no teníamos el atributo día, por lo tanto podíamos interpretar que las llamadas todas tienen el mismo día o directamente los cálculos se hacen a partir del mes. Por otro lado, el hecho de las infracciones sí que tenía como atributo el día pero todas las métricas tanto derivadas como no se calculan a partir del día no del mes, es decir, son un recuento diario, aunque con ello podemos tener las perspectiva mensual o anual.

Por lo tanto, el enfoque de este hecho es completamente diferente a los demás con el objetivo de cumplir los requisitos definidos en el primer apartado de la práctica.

Finalmente, nos queda diseñar la última estrella, la cual nos indica el porcentaje de población que evitaba las aglomeraciones, según el grupo de edad y provincia:

Hecho	Descripción
Fact_Evitacion_Aglomeracion	Porcentaje de población que evitaba aglomeraciones.

Tabla 17 - Tabla de Hechos de la evitación de las aglomeraciones.

Las dimensiones que hemos identificado son las siguientes:

Dimensión	Descripción

Dim_Lugar	Porcentaje de población que evitaba las aglomeraciones según la provincia.
Dim_Grupo_Edad	Porcentaje de población que evitaba las aglomeraciones según el grupo de edad.

Tabla 18 - Tabla de Dimensiones de la evitación de las aglomeraciones.

Como resultado del anterior análisis, obtenemos el siguiente diseño conceptual:

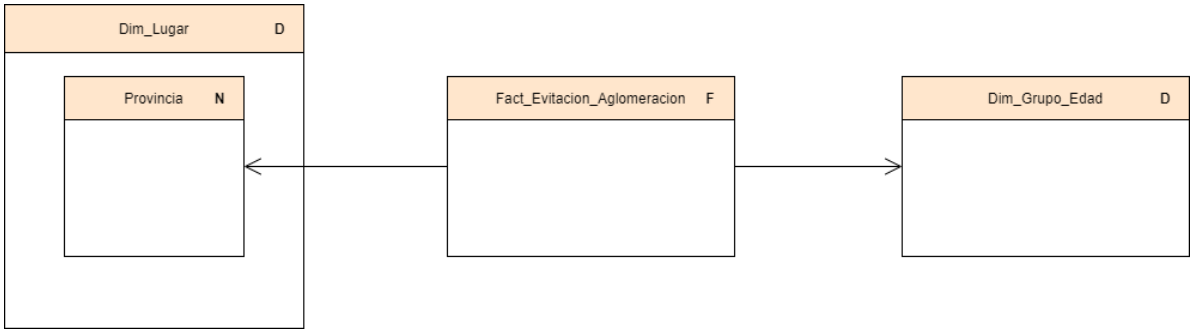

Ilustración 6 - Diseño conceptual de la evitación de las aglomeraciones.

Diseño lógico

Una vez que hemos definidos el modelo conceptual, tenemos que identificar cada una de las métricas que tienen nuestros hechos y los atributos que tienen nuestras dimensiones.

Respecto a la gestión de las llamadas al 112, presentamos las siguientes métricas:

Hecho	Métrica	Descripción
Fact_Llamadas112	N_Llamadas	Número de llamadas que se han realizado.
Fact_Llamadas112	N_Suma_Llamadas_Mensuales	Número que hace el recuento de las llamadas mensuales.

Tabla 19- Tabla de Métricas para las llamadas al 112.

Respecto a los atributos que hay dentro de cada dimensión son:

Dimensión	Atributos descriptores

Dim_Tiempo	N_Dia, N_Mes, N_Año, D_DiaMesAño, S_MesAño
Dim_Tipologia	S_Tipo
Dim_Lugar	S_Municipio, S_Comarca, S_Provincia

Tabla 20 - Tabla de Atributos para las llamadas al 112.

Cabe destacar, que vamos a tener atributos descriptores para cada uno de los niveles que hay dentro de una dimensión.

Por otro lado, como regla general para definir tanto los atributos como las métricas, podemos ver que empiezan por:

- D: si es de tipo *Date*.
- S: si es de tipo *String*.
- N: si es de tipo *Number.*

Finalmente, para este hecho nos quedaría el siguiente diseño lógico:

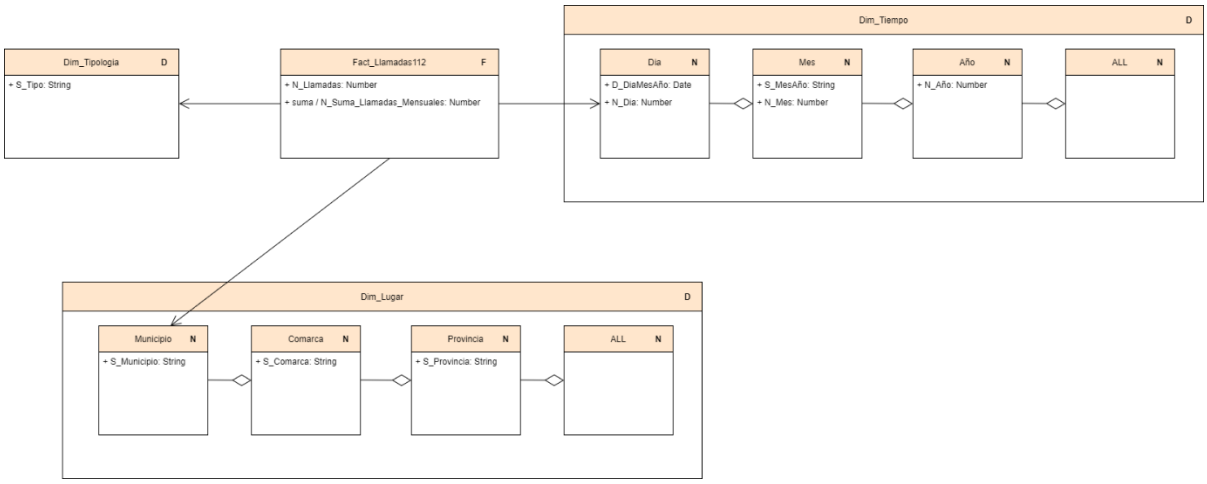

Ilustración 7 - Diseño lógico de las llamadas al 112.

Respecto al nivel “ALL” estamos haciendo referencia a la agrupación de todas las instancias de la dimensión al mismo tiempo (esto será así a lo largo de toda la práctica), es por ello que no tiene atributos descriptores por sí mismo.

Respecto a la gestión de las infracciones presentamos las siguientes métricas:

Hecho	Métrica	Descripción
Fact_Infracciones_Evolucion	N_Identificados	Número de identificados.
Fact_Infracciones_Evolucion	N_Detenidos	Número de detenidos.
Fact_Infracciones_Evolucion	N_Denuncias	Número de denuncias.
Fact_Infracciones_Evolucion	N_Vehiculos	Número de vehículos interceptados.
Fact_Infracciones_Evolucion	N_Suma_Identificados	Recuento del número de identificados diarios.
Fact_Infracciones_Evolucion	N_Suma_Detenidos	Recuento del número de detenidos diarios.
Fact_Infracciones_Evolucion	N_Suma_Denuncias	Recuento del número de denuncias diarias.
Fact_Infracciones_Evolucion	N_Suma_Vehiculos	Recuento del número de vehículos interceptados diarios.

Tabla 21 - Tabla de Métricas para las infracciones.

Respecto a los atributos que hay dentro de cada dimensión:

Dimensión	Atributos descriptores
Dim_Tiempo	N_Dia, N_Mes, N_Año, D_DiaMesAño, S_MesAño
Dim_Lugar	S_Provincia
Dim_Policia	S_TipoPolicia

Tabla 22 - Tabla de Atributos para las infracciones.

Por lo tanto, una vez analizadas las métricas y los atributos nos quedaría el siguiente diseño lógico:

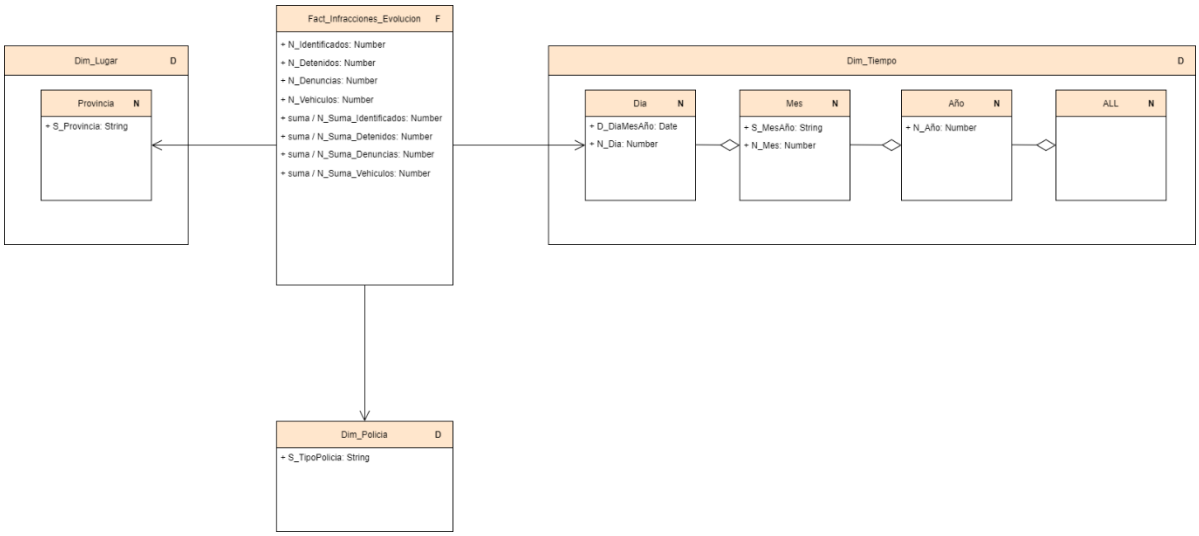

Ilustración 8 - Diseño lógico de las infracciones.

Respecto al análisis de la población según su provincia, identificamos las siguientes métricas:

Hecho	Métrica	Descripción
Fact_Poblacion	N_Poblacion	Número de la población.

Tabla 23 - Tabla de Métricas para la población.

Por otro lado, los atributos que vamos a tener en las dimensiones van a ser los siguientes:

Dimensión	Atributos descriptores
Dim_Lugar	S_Provincia

Tabla 24 - Tabla de Atributos para la población.

Una vez identificadas las métricas y lo hechos, representamos el diseño lógico:

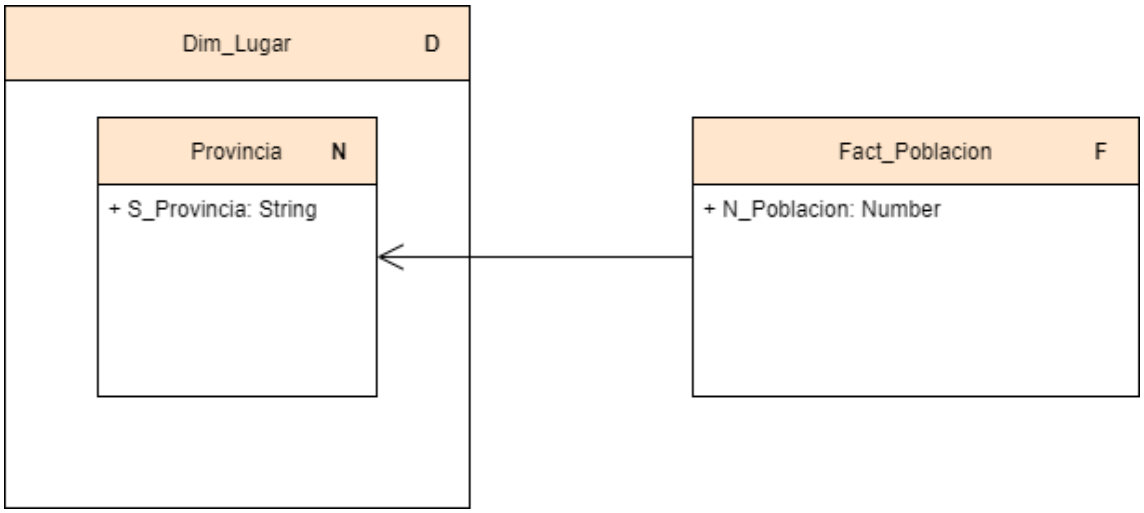

Ilustración 9 - Diseño lógico de la población.

Respecto a la evolución de la movilidad en la población española, con el objetivo de cumplir los requerimientos identificamos las siguiente métricas:

Hecho	Métrica	Descripción
Fact_Movilidad	PorcentajeAtomico.N_Porcentaje	Porcentaje de la movilidad de la población diario.
Fact_Movilidad	MediaPorcentajeMes.N_Media_Porcentaje	Porcentaje de la movilidad de la población mensual.

Tabla 25 - Tabla de Métricas para la movilidad.

Una vez identificadas las métricas, definimos los atributos para cada dimensión:

Dimensión	Atributos descriptores
Dim_Tiempo	N_Dia, N_Mes, N_Año, D_DiaMesAño, S_MesAño
Dim_Lugar	S_Provincia

Tabla 26 - Tabla de Atributos para la movilidad.

Por lo tanto, el diseño lógico sería el siguiente:

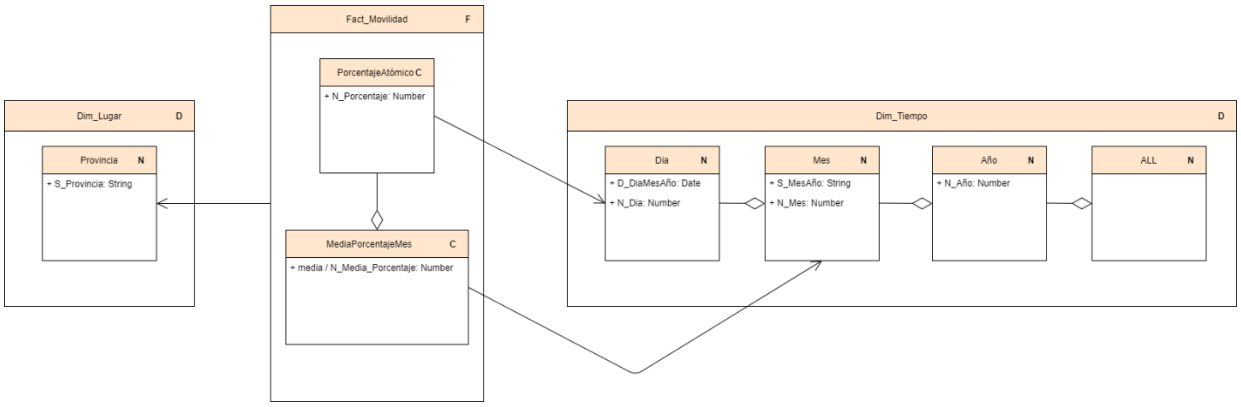

Ilustración 10 - Diseño lógico de la evolución de la movilidad.

Tal y como podemos observar en la anterior ilustración, este hecho tiene dos celdas ya que la granularidad es diferente, esto ya fue explicado en el diseño conceptual.

Finalmente, nos queda diseñar a nivel lógico la última estrella, la cual nos indica el porcentaje de población que evitaba las aglomeraciones, según el grupo de edad y provincia.

Hemos identificado las siguientes métricas, con el fin de cumplir las necesidades y requerimientos:

Hecho	Métrica	Descripción
Fact_Evitacion_Aglomeracion	N_Porcentaje	Porcentaje de la población que evitaba las aglomeración.
Fact_Evitacion_Aglomeracion	N_Max_Porcentaje	Máximo porcentaje de la población que evitaba las aglomeraciones.

Tabla 27 - Tabla de Métricas para el porcentaje de población que evitaba las aglomeraciones.

Por otro lado, los atributos que hemos identificado son:

Dimensión	Atributos descriptores
Dim_Lugar	S_Provincia
Dim_Grupo_Edad	S_GrupoEdad

Tabla 28 - Tabla de Atributos para el porcentaje de población que evitaba las aglomeraciones.

A la vista de lo anterior, hemos diseñado el siguiente modelo:

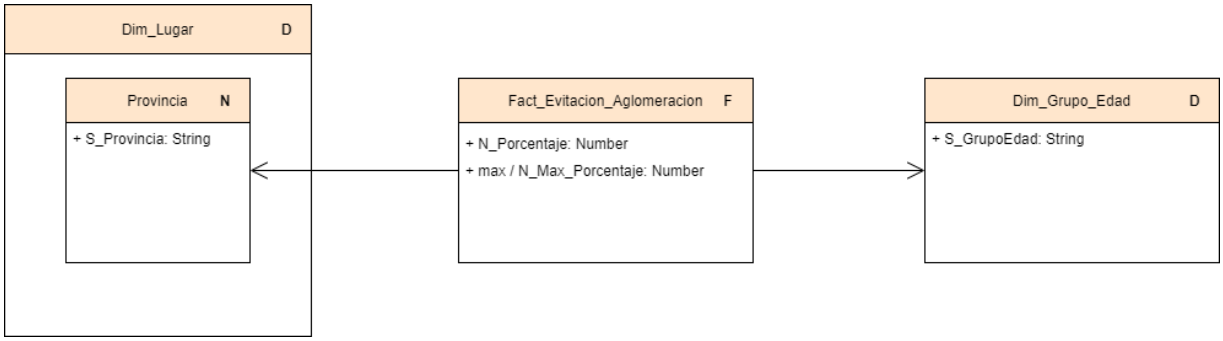

Ilustración 11 - Diseño lógico de la evitación de las aglomeraciones.

Diseño físico

Una vez que ya está claro el diseño conceptual y lógico de nuestro almacén de datos, nos falta el último paso para asegurarnos de que nuestro *data warehouse* nos proporciona el mejor rendimiento posible.

En este paso tenemos que definir las diferentes claves primarias de las dimensiones y las claves foráneas que a su vez serán claves primarias de los hechos. También debemos de tener en cuenta el tipo de datos junto con su tamaño.

Primero vamos a realizar el diseño físico de las dimensiones y posteriormente el de los hechos.

Dimensiones

Dim_Tipología: corresponde a la dimensión de las tipologías correspondientes a las llamadas al 112 en la comunidad de Cataluña.

Nombre Campo	Tipo	Tamaño	Ejemplo
PK_Tipologia (PK)	Numérico	4	1
S_Tipo	Texto	50	“Seguretat”

Tabla 29 - Diseño físico de Dim_Tipologia.

Dim_Policia: corresponde a la dimensión del tipo de policía a la hora de analizar las infracciones/desobediencias durante el estado de emergencia en el País Vasco.

Nombre Campo	Tipo	Tamaño	Ejemplo
PK_Policia (PK)	Numérico	1	1
S_TipoPolicia	Texto	15	“ERTZAINTZA”

Tabla 30 - Diseño físico de Dim_Policia.

Dim_Grupo_Edad: corresponde a la dimensión del rango de edad para analizar el porcentaje de personas que evitaron una aglomeración por motivo del coronavirus.

Nombre Campo	Tipo	Tamaño	Ejemplo
PK_GrupoEdad (PK)	Numérico	1	1
S_GrupoEdad	Texto	10	“14 - 24”

Tabla 31 - Diseño físico de Dim_GrupoEdad.

Dim_Tiempo: corresponde a la dimensión tiempo de nuestro almacén de datos, es una dimensión común en la gran mayoría de estrellas.

Nombre Campo	Tipo	Tamaño	Ejemplo
PK_Fecha (PK)	Numérico	8	20200101
D_DiaMesAño	Date	10	31/12/2020
N_Dia	Numérico	2	31
S_MesAño	Texto	20	“Diciembre 2020”
N_Mes	Numérico	2	12
N_Año	Numérico	4	2020

Tabla 32 - Diseño físico de Dim_Tiempo.

Dim_Lugar_ corresponde a la dimensión del lugar de nuestro almacén de datos, es una dimensión común a todas las estrellas.

Nombre Campo	Tipo	Tamaño	Ejemplo
PK_Lugar (PK)	Numérico	5	11111
S_Provincia	Texto	50	“Barcelona”
S_Comarca	Texto	100	“Alt Penedes”
S_Municipio	Texto	150	“Avinyonte del Penedes”

Tabla 33 - Diseño físico de Dim_Lugar.

Hechos

Una vez definida el diseño para las dimensiones, hacemos lo mismo respecto a los hechos.

Fact_Llamadas112: hecho que nos permite obtener el número de llamadas y el recuento de las llamadas mensuales, según su tipología, fecha y lugar.

Nombre Campo	Tipo	Tamaño	Ejemplo
PK_FK_Tipologia (PK)	Numérico	4	1
PK_FK_Fecha (PK)	Numérico	8	20200101
PK_FK_Lugar (PK)	Numérico	5	11111
N_Llamadas	Numérico	6	100000
N_Suma_Llamadas_Mensuales	Numérico	7	200000

Tabla 34 - Diseño físico de Fact_Llamadas112

Por lo tanto, quedaría finalmente este diseño:

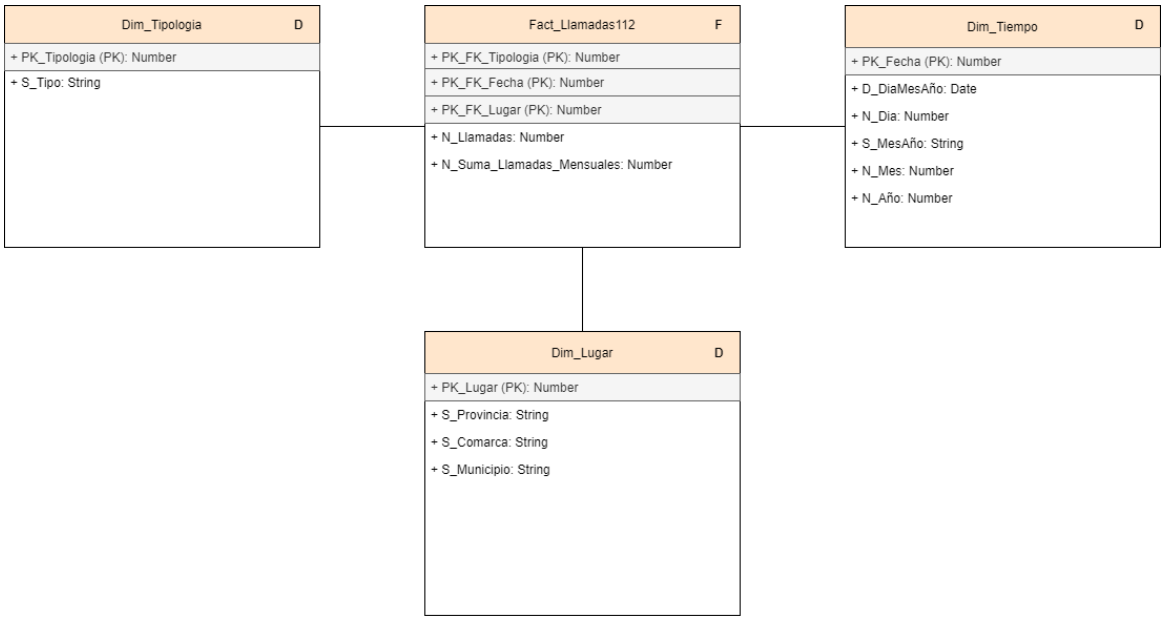

Ilustración 12 - Diseño físico de las llamadas al 112.

Fact_Infracciones_Evolucion: hecho que nos permite medir el número de identificados, detenidos, denuncias, vehículos, y el recuento de los atributos anteriores diariamente.

Nombre Campo	Tipo	Tamaño	Ejemplo
PK_FK_Policia (PK)	Numérico	1	1
PK_FK_Fecha (PK)	Numérico	8	20200101
PK_FK_Lugar (PK)	Numérico	5	11111
N_Identificados	Numérico	6	1000
N_Detenidos	Numérico	6	1000
N_Denuncias	Numérico	6	1000
N_Vehiculos	Numérico	6	1000
N_Suma_Identificados	Numérico	7	10000
N_Suma_Detenidos	Numérico	7	10000

N_Suma_Denuncias	Numérico	7	10000
N_Suma_Vehiculos	Numérico	7	10000

Tabla 35 - Diseño físico de Fact_Infracciones_Evolucion.

El diseño que finalmente nos queda es el siguiente:

Ilustración 13 - Diseño físico de las infracciones.

Fact_Poblacion: hecho que nos permite obtener el número de habitantes que hay por provincia.

Nombre Campo	Tipo	Tamaño	Ejemplo
PK_FK_Lugar (PK)	Numérico	5	11111
N_Poblacion	Numérico	8	6000000

Tabla 36 - Diseño físico del Fact_Poblacion.

El diseño sería el siguiente:

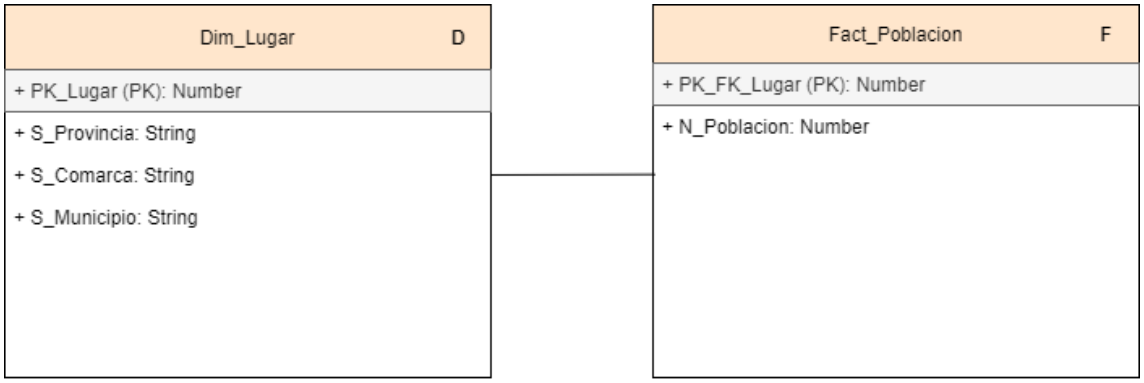

Ilustración 14 - Diseño físico del análisis de la población.

Fact_Movilidad: hecho que nos permite obtener el porcentaje de la población que se ha desplazado durante el estado de alarma, junto con la media del porcentaje mensual.

Nombre Campo	Tipo	Tamaño	Ejemplo
PK_FK_Fecha (PK)	Numérico	8	20200101
PK_FK_Lugar (PK)	Numérico	5	11111
N_Porcentaje	Decimal	(4,2)	99.99
N_Media_Porcentaje	Decimal	(4,2)	50.00

Tabla 37 - Diseño físico de Fact_Movilidad.

El diseño resultante es el siguiente:

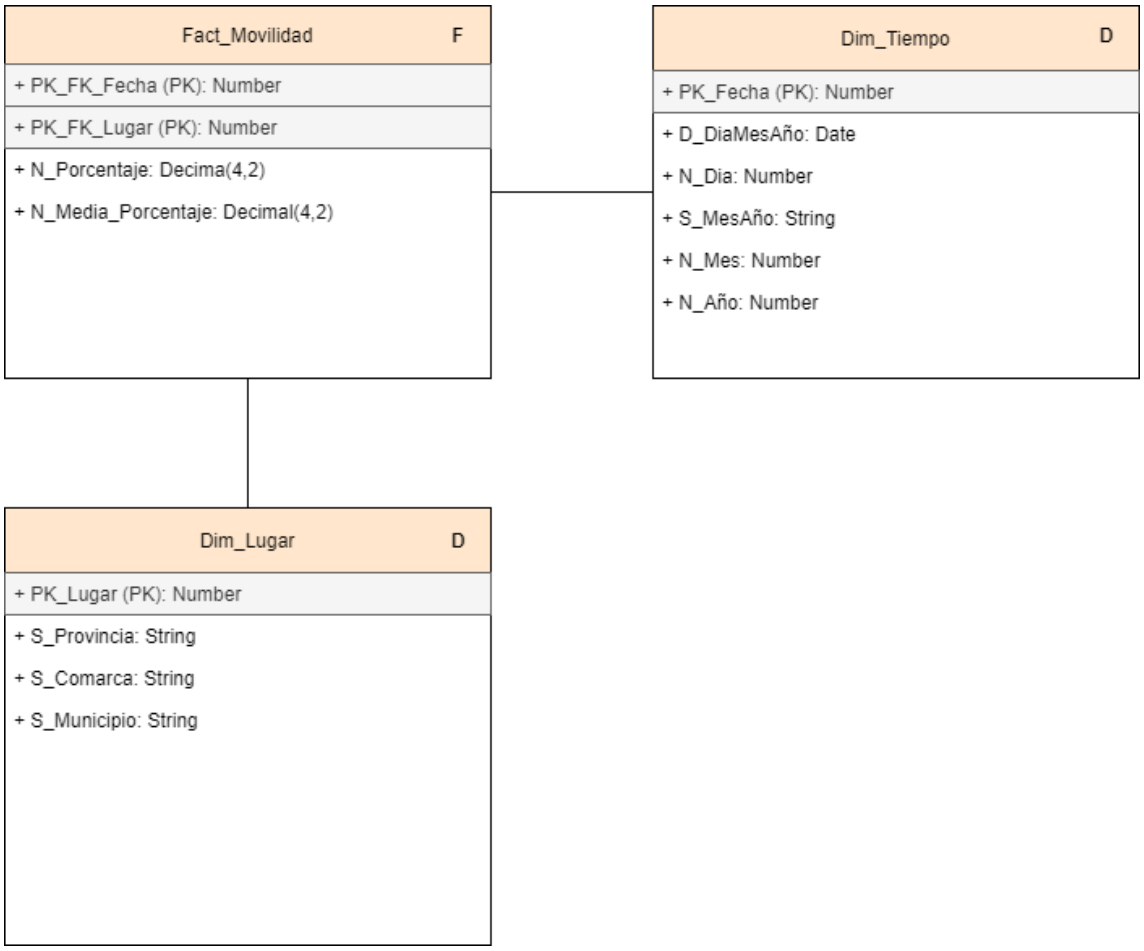

Ilustración 15 - Diseño físico de la evolución de la movilidad.

Fact_Evitacion_Aglomeracion: hecho que nos permite obtener el porcentaje de la población que ha evitado una aglomeración según la edad y la provincia, además obtenemos el máximo porcentaje de entre todos los rangos por provincia.

Nombre Campo	Tipo	Tamaño	Ejemplo
PK_FK_Lugar (PK)	Numérico	5	11111
PK_FK_GrupoEdad (PK)	Numérico	1	1
N_Porcentaje	Decimal	(4,2)	99.99
N_Max_Porcentaje	Decima	(4,2)	99.99

Tabla 38 - Diseño físico de Fact_Evitacion_Aglomeracion.

El diseño resultante sería el siguiente:

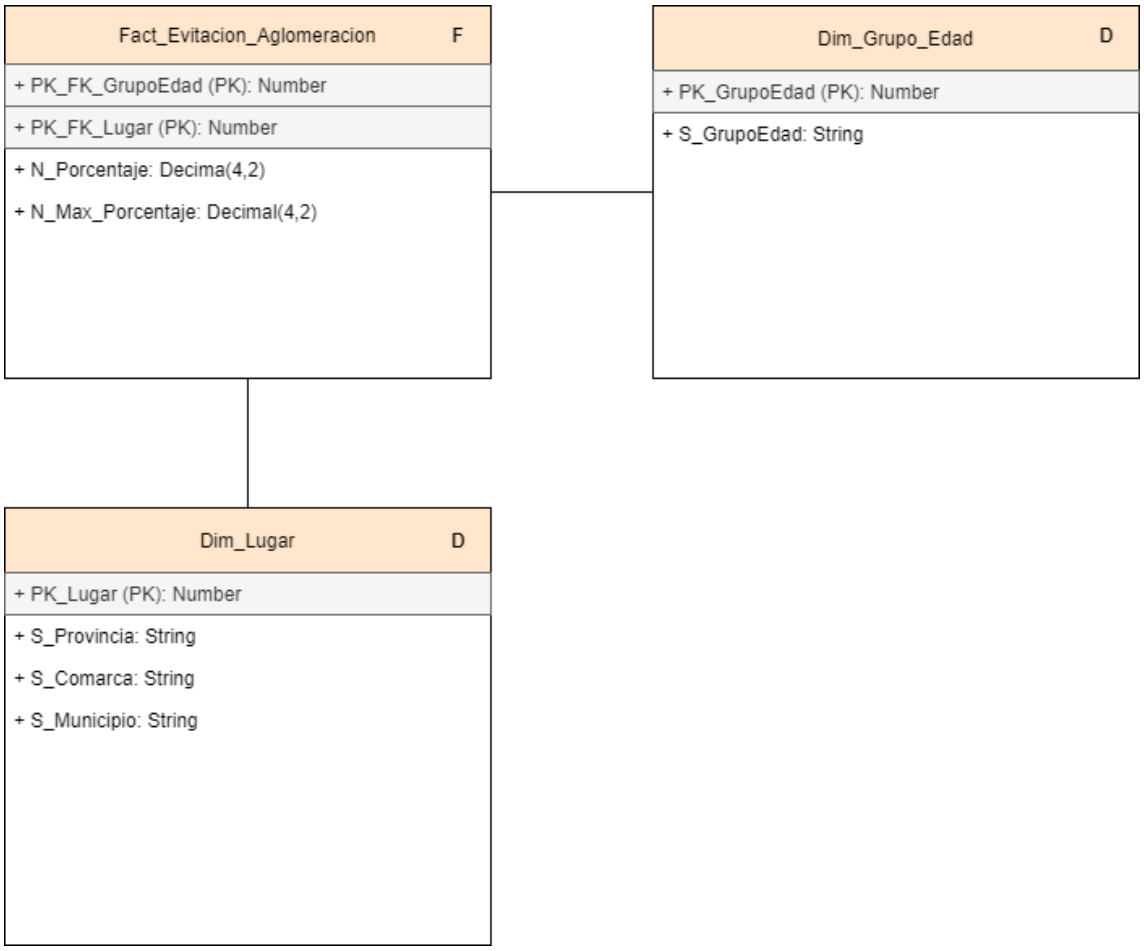

Ilustración 16 - Diseño físico de la evitación de las aglomeraciones.

6. 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).