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

Solución 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 el análisis del impacto conductual de la COVID-19 sobre la población.

1. Análisis de los requerimientos

El análisis de los requerimientos se basa en identificar las necesidades específicas que tiene una organización particular respecto al análisis de la información. Normalmente, en esta fase se debe ser previsor y pensar más allá de las necesidades actuales para poder cubrir las futuras.

La necesidad principal de la organización encargada del análisis del impacto conductual es disponer de la información integrada para su análisis y su posterior difusión mediante las herramientas de inteligencia de negocio. Estas ayudarán a facilitar la toma de decisiones a todos los usuarios potenciales para garantizar el cumplimiento de los siguientes objetivos:

- Realizar el seguimiento de ciertas mediciones obtenidas durante el estado de alarma, como la movilidad y los datos de la población, los expedientes sancionadores aplicados y el grado de consciencia de los ciudadanos para evitar las aglomeraciones con motivo del coronavirus.

- Estudiar la gestión de llamadas al teléfono de urgencias 112 de Cataluña (CAT112) y analizar el impacto conductual que ha tenido la COVID-19 en estas.

El diseño de un almacén de datos de un proyecto incluye la creación y la implementación de un modelo dimensional o multidimensional, el diseño y la implementación de procesos ETL y del modelo OLAP y, por último, el diseño de las consultas establecidas en el enunciado.

A continuación, se indica la información necesaria identificada:

1) Analizar las mediciones desde diferentes perspectivas:

- por fecha,
- por ámbito geográfico,
- por grupo de edad,
- por unidad de medida.

2) Analizar la gestión de llamadas al teléfono de urgencias 112 de Cataluña desde diferentes perspectivas:

- por fecha,
- por ámbito geográfico,
- por tipología del incidente.

Si se tiene en cuenta toda esta información, el sistema podrá responder a múltiples preguntas y de esta manera conseguirá cubrir las necesidades de los usuarios potenciales.

A continuación se indican de forma específica las preguntas que, como mínimo, el sistema debe ser capaz de responder:

- Análisis de las provincias con mayor porcentaje de movilidad según los datos móviles.
- Análisis del porcentaje de la población que evitaba las aglomeraciones según la comunidad autónoma.
- Análisis del promedio de sanciones por habitante.
- Evolución de las llamadas al teléfono de urgencias 112 en Cataluña por tipología de la llamada.
- Análisis de las llamadas de urgencia frente al porcentaje de la población que evitaba las aglomeraciones entre los meses comprendidos entre marzo y junio del 2020 en Cataluña desglosado por provincias.
- Determinación del día de la semana con menor número de denuncias.
- Análisis de las diez fechas (*top ten*) con mayor número de llamadas de urgencia al 112 con la tipología de tránsito registrada, tanto en la época de la COVID-19 como antes de ella.

2. Análisis de las fuentes de datos

En este apartado se deben revisar las fuentes de datos proporcionadas, qué tipo de información contienen, cuál es su formato y qué datos deben ser cargados. Véase a continuación un análisis detallado por cada tipo de formato:

1) **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.

   A continuación, se analizan los campos del fichero .xlsx (Excel):

Nombre del campo	Tipo	Ejemplo
TT. HH.	Texto	ARABA
IDENTIFICADOS (ERTZAINTZA)	Numérico	10.717
DETENIDOS (ERTZAINTZA)	Numérico	40
DENUNCIAS INTERPUESTAS (ERTZAINTZA)	Numérico	2.586
VEHÍCULOS INTERCEPTADOS (ERTZAINTZA)	Numérico	15.182
IDENTIFICADOS (PP. LL.)	Numérico	21.599
DETENIDOS (PP. LL.)	Numérico	29
DENUNCIAS INTERPUESTAS (PP. LL.)	Numérico	2.748
VEHÍCULOS INTERCEPTADOS (PP. LL.)	Numérico	19.250
FECHA FINAL	Fecha	18/06/2020

   Observaciones:

   - Los apartados «identificados», «detenidos» y «denuncias interpuestas» hacen referencia a personas.
   - Solo se debe tratar la hoja «**datos_tratados**».
   - Los datos son los **acumulados** de las actuaciones específicas por la COVID-19 desde el 16 de marzo hasta la fecha indicada.
   - El intervalo de la columna «fecha» es el siguiente: inicio 7/4/2020 y fin 18/6/2020.

   Terminología:

   - **Ertzaintza**: policía autonómica del País Vasco.
   - **PP. LL.**: policías locales.

   Total de registros: 219 registros.

2) **poblacion_9687bsc.csv:** cifras de la población española por provincia a 1 de enero de 2020.

   Este fichero plano desde el origen viene con las siguientes características:

   - Formato: CSV.

- Primera línea con etiquetas de los campos.
- Separador de campos: punto y coma (;).

Nombre del campo	Tipo	Ejemplo
Edad simple	Texto	Total
Provincias	Texto	02 Albacete
Sexo	Texto	Ambos sexos
Periodo	Texto	1 de enero de 2020
Total	Numérico	389.830

Observaciones:

- El campo «edad simple» siempre tiene el mismo valor («Total»).
- El campo «sexo» siempre tiene el mismo valor («Ambos sexos»).
- El campo «periodo» siempre tiene el mismo valor («1 de enero de 2020»).

Total de registros: 52 registros.

3) **rows.xml**: contiene los datos de las llamadas operativas gestionadas por el CAT112 (lugar de entrada de las llamadas al teléfono de emergencias 112 de Cataluña). Se detalla el ámbito geográfico (provincia, comarca y municipio) y la tipología del incidente (accidente de tráfico, civismo, incendios, asistencia sanitaria, seguridad...). Se consideran *llamadas operativas* aquellas que son una emergencia real y que generan una activación de los cuerpos de emergencia. Así, se detallan las llamadas operativas de los incidentes en los que el ámbito de gestión es el ámbito geográfico de Cataluña.

A continuación, se analizan los campos del fichero .xml:

Nombre del campo	Traducción	Tipo	Ejemplo
Any	Año	Numérico	2020
Mes	Mes	Numérico	10
Provincia	Provincia	Texto	Tarragona
Comarca	Comarca	Texto	Montsià
Municipi	Municipio	Texto	Sant Carles de la Ràpita
Tipus	Tipos	Texto	Seguridad
Trucades	Llamadas	Numérico	30

Total de registros: 340.307 registros.

4) **35167bsc.csv:** contiene los datos de la movilidad (‘movimiento de personas’) de la población durante el estado de alarma por comunidades autónomas y por provincias.

Este fichero plano desde el origen viene con las siguientes características:

- Formato: CSV.
- Primera línea con etiquetas de los campos.

- Separador de campos: punto y coma (;).

Nombre del campo	Tipo	Ejemplo
Zonas de movilidad	Texto	Zaragoza
Periodo	Fecha	30/3/2020
Total	Numérico	10,82

Observaciones:

- La unidad del campo «total» está expresada en porcentaje (%).
- Como aclaración sobre qué significan los porcentajes de movilidad del estudio ofrecido por el INE, la metodología utilizada se explica en el siguiente enlace: <https://www.ine.es/covid/exp_movilidad_covid_proyecto.pdf>.

Total de registros: 4.732 registros.

5) **statistic_id1104235_covid-19_-poblacion-que-evitaba-las-aglomeraciones-segun-edad-en-espana-2020.xlsx:** contiene los datos sobre el porcentaje de la población que evitaba las aglomeraciones con motivo del coronavirus. Están organizados por grupos de edad y por provincias en el periodo comprendido entre el 1 de marzo y el 30 de septiembre de 2020.

A continuación, se analizan los campos del fichero .xlsx (Excel):

Nombre del campo	Tipo	Ejemplo
Ámbito geográfico	Texto	Álava (Araba) (País Vasco)
Grupo de edad 14-24	Numérico	36,99
Grupo de edad 25-34	Numérico	52,46
Grupo de edad 35-44	Numérico	42,70
Grupo de edad 45-54	Numérico	46,66
Grupo de edad 55-64	Numérico	35,73
Grupo de edad > 64	Numérico	51,13

Observaciones:

- Solo se debe tratar la hoja «Datos_provincias».
- Las hojas «Ficha técnica», «Datos» y «Proyeccion_Datos_provincias» no se deben tratar.

Total de registros: 50 registros.

Estimación de la volumetría

En los proyectos de diseño de la factoría de información corporativa existe una primera fase en la que se realiza una carga inicial y, *a posteriori*, una segunda fase para realizar las cargas incrementales de los datos nuevos que van llegando.

Una posible estimación del volumen de datos del almacén para la carga de los datos inicial sería la siguiente:

Fichero	Registros	Valores	Datos
ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx	219	10	2.190
poblacion_9687bsc.csv	52	5	260
rows.xml	340.307	7	2.382.149
35167bsc.csv	4.732	3	14.196
statistic_id1104235_covid-19_-poblacion-que-evitaba-las-aglomeraciones-segun-edad-en-espana-2020.xlsx	50	7	350
Total	345.360	32	2.399.145

3. Análisis funcional

A continuación, se propone el tipo de arquitectura para la factoría de información que mejor se adecua al proyecto. Para ello, se consideran los requisitos funcionales y se establece la prioridad entre exigible (E) o deseable (D). En el contexto de esta actividad, los requerimientos exigibles son aquellos que demanda el enunciado, y los deseables, los que complementan la actividad.

Además, en términos de la escala de prioridades, se asigna una prioridad del 1 al 3, siendo 1 completamente prioritario para la actividad y 3 no prioritario.

A continuación, se describen los requerimientos funcionales para el diseño de una factoría de información para la organización, teniendo en cuenta las consideraciones del enunciado:

#	Requerimiento	Prioridad	Exigible/ deseable
1	Se extraerá de forma adecuada la información de las fuentes de datos.	1	E
2	Se creará un almacén de datos.	1	E
3	Se cargará la información sobre el impacto conductual de la COVID-19 sobre la población en el almacén de datos.	1	E
4	Se creará un modelo OLAP para consultas multidimensionales de los usuarios.	2	E
5	Se crearán los informes estáticos solicitados.	2	E
6	Se redactará un manual de carga de datos inicial e incremental.	3	D

Cabe comentar que, en un caso genérico real, se pueden encontrar también otros requerimientos funcionales, como los que se muestran a continuación:

- Análisis de viabilidad y análisis de riesgos.
- Creación de procesos de calidad de datos.
- Creación de un *data marts* (si se analizan otras áreas).
- Creación de procesos de cargas incrementales.
- Creación de un repositorio de metadatos de gestión del almacén de datos, así como de los procesos ETL, que permita realizar la trazabilidad a lo largo del ciclo de vida de los datos.

Asimismo, dado que estos sistemas frecuentemente forman parte de la implementación de un sistema de inteligencia de negocio, la lista de requerimientos funcionales sería mucho mayor, como puede ser la administración de seguridad en cuanto a datos y a usuarios.

En términos de la arquitectura funcional existen los siguientes elementos:

- Las fuentes de datos de las que se dispone son las siguientes:
  - Dos ficheros Excel (xlsx).
  - Un fichero en formato XML.

- Dos ficheros planos (csv).

- La arquitectura de la factoría de información puede estar formada por varios elementos alojados en la misma máquina:

  - *Staging area* (opcional): en el caso de tener múltiples fuentes (ficheros, bases de datos, servicios RSS, etc.) es conveniente cargarlas para consolidar la información en una estructura de carga intermedia que puede ser creada en la misma base de datos.

    Esta área del DW también puede servir para entender, simplificar y consolidar los procesos ETL.

  - *Data mart* del impacto conductual de la COVID-19 sobre la población: como nos centramos en una única área temática es más correcto considerar que se está creando un *data mart* en lugar de un almacén de datos corporativo.

  - MOLAP: a partir de la información del *data mart* se creará un cubo multidimensional.

Según lo comentado anteriormente, se podría elegir entre dos diseños para la arquitectura funcional. Por un lado, tenemos una arquitectura funcional que usa un área intermedia (*staging area*) y se crearía dentro de la misma base de datos, cuyos objetos se identificarán con un prefijo en los nombres. El siguiente gráfico resume los elementos de la arquitectura necesarios para esta actividad:

Por otro lado, también sería correcto utilizar una arquitectura sin área intermedia (*staging area*) que identifique las tablas intermedias en el *data mart* con un prefijo en el nombre, como, por ejemplo, IN_nombre_tabla_intermedia.

En esta solución se propone un diseño que utiliza un área intermedia (*staging area*) que se crearía dentro de la misma base de datos, cuyos objetos se identificarán con un prefijo en los nombres (STG_).

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

4.1. Diseño conceptual

Para el correcto desarrollo del DW es preciso definir los hechos (*facts*), las dimensiones del análisis (*dimensions*), las métricas y los atributos que permitan tener el nivel de granularidad suficiente para la presentación de los objetivos. Estos se han definido en el análisis de los requerimientos y de las fuentes de datos.

Del análisis de las fuentes de datos y de los requerimientos iniciales, se puede determinar que los hechos que se deben considerar son los siguientes:

- **Mediciones.** Hace referencia a ciertas mediciones obtenidas durante el estado de alarma, como la movilidad y los datos de la población, los expedientes sancionadores aplicados y el grado de consciencia de los ciudadanos para evitar las aglomeraciones con motivo del coronavirus.

- **Llamadas112.** Hace referencia a la tipología de las llamadas de urgencia al 112 de Cataluña (CAT112).

El **análisis de las mediciones** determina el diseño de la primera tabla de hechos, como se puede observar a continuación:

Tabla de hechos	Descripción
FACT_Mediciones	Mediciones obtenidas durante el estado de alarma

Las **métricas** de la tabla de hechos FACT_Mediciones son las siguientes:

Métricas	Descripción
Valor	Valor de la medición

Las métricas de esta tabla de hechos podrán ser analizadas desde diferentes perspectivas, a partir de las siguientes dimensiones:

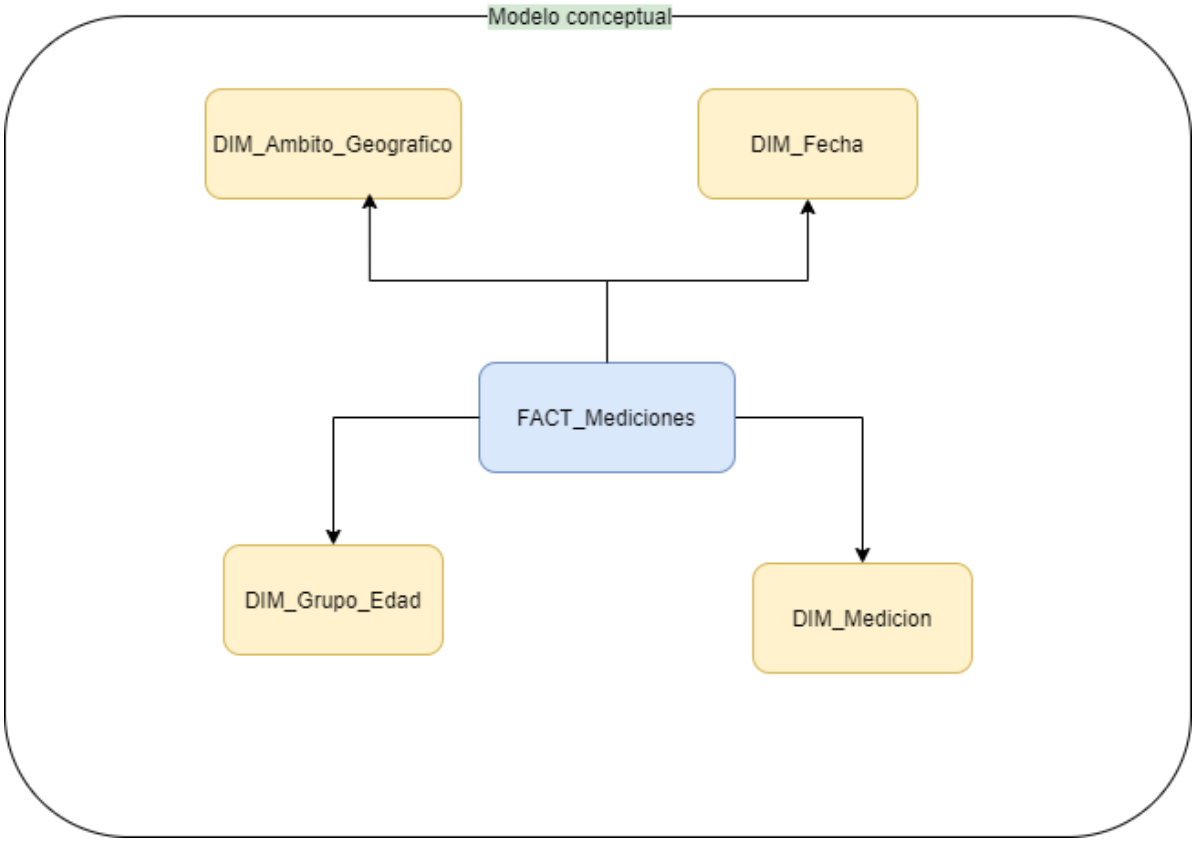
Dimensiones	Descripción
Fecha	Fecha en la que se realiza la medición.
Ámbito geográfico	Ámbito geográfico donde se localiza la medición.
Grupo edad	Distribución de los grupos de edad de la medición.
Medición	Medición que tratar.

En las dimensiones con jerarquía se puede optar por las siguientes alternativas:

- **Diseño en copo de nieve**: se aplican las técnicas de normalización de las bases de datos para optimizar el espacio y eliminar las redundancias. Esto requiere la creación de nuevas tablas y relaciones, lo que empeora el rendimiento.

- **Diseño en estrella**: no se normalizan las bases de datos y se mantienen las jerarquías dentro de una misma tabla.

El diseño conceptual para esta tabla de hechos (Fact_Mediciones) y sus dimensiones con un **diseño en estrella** es el siguiente:



Para el análisis de la **gestión de las llamadas de urgencia al 112 de Cataluña** se identifica una segunda tabla de hechos, como la siguiente:

Tabla de hechos	Descripción
FACT_Llamadas112	Gestión de las llamadas de urgencia al 112 de Cataluña (CAT112)

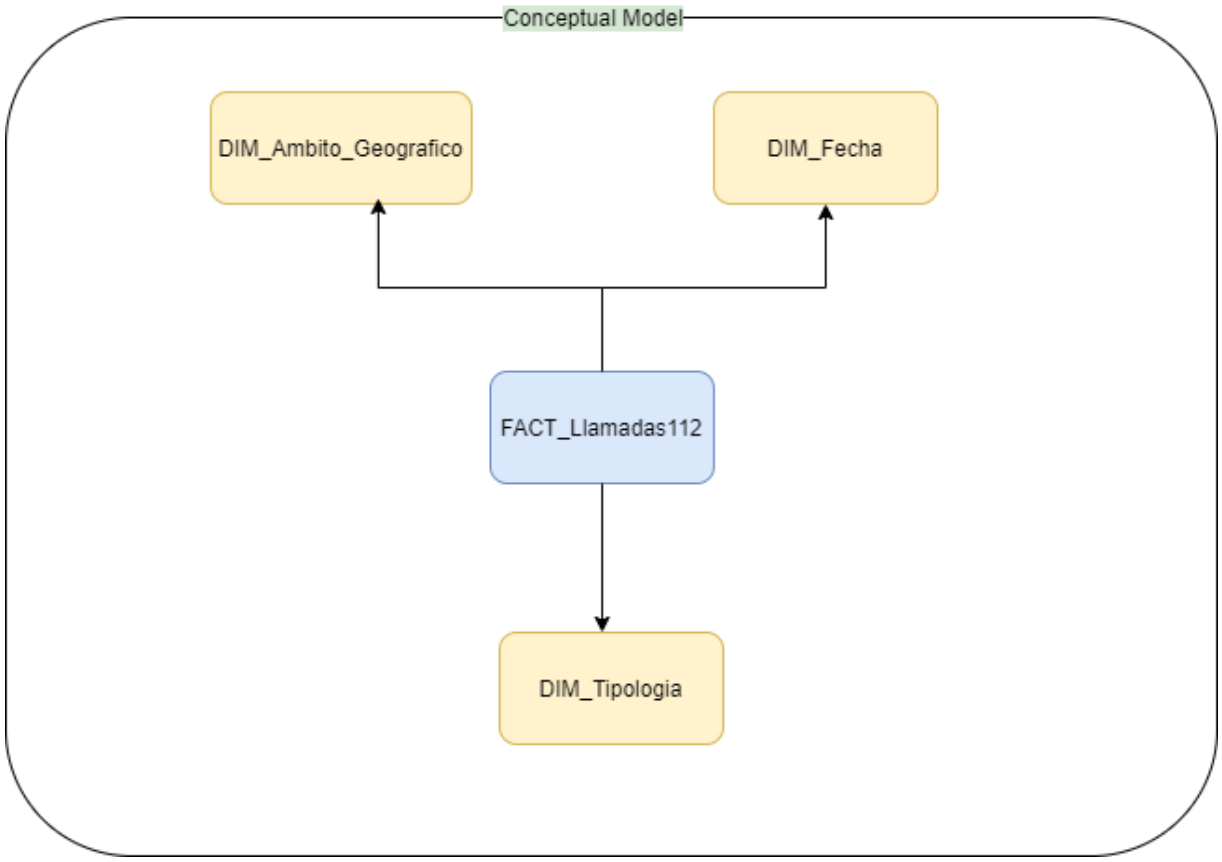
Esta tabla de hechos almacenará las siguientes **métricas**:

Métricas	Descripción
Llamada	Número de llamadas operativas de incidencias

Esta métrica puede ser analizada desde diferentes perspectivas a través de las siguientes dimensiones:

Dimensiones	Descripción
Fecha	Año o mes en que se producen las llamadas sobre la incidencia.
Ámbito geográfico	Ámbito geográfico donde se producen las llamadas sobre la incidencia.
Tipología	Tipología de las llamadas sobre la incidencia.

El diseño conceptual para esta tabla de hechos (FACT_Llamadas112) y sus dimensiones con un **diseño en estrella** es el siguiente:



4.2. Diseño lógico

Una vez obtenido el modelo conceptual del almacén de datos para el análisis del impacto conductual, se detallan las métricas de cada una de las tablas de hechos y sus atributos para diseñar el modelo lógico.

A continuación, se muestra una tabla con las métricas que contiene cada tabla de hechos y que compone el modelo lógico del almacén de datos:

Tabla de hechos	Métricas
FACT_Mediciones	Valor
FACT_Llamadas112	Llamadas

A continuación, se procederá a detallar los atributos que contiene cada tabla de hechos. Estos, junto con las métricas, permitirán realizar los diferentes análisis de los requerimientos planteados.

En la siguiente tabla, se muestran los atributos de las dimensiones de la tabla de hechos FACT_Mediciones:

Dimensiones	Atributos
DIM_Fecha	Año, mes y día
DIM_Ambito_Geografico	Código, provincia y comunidad autónoma
DIM_Grupo_Edad	Nombre e intervalo
DIM_Medicion	Nombre de la medición y unidad de medida

A continuación, se detallan ejemplos para cuatro de las mediciones (aunque haya más mediciones, solo se incorporan estas a modo de ejemplo):

Nombre del fichero	Nombre de la medición	Unidad de medida
Acumulado-denuncias-infracciones.xls	Denuncias interpuestas	Denuncias
poblacion_9687bsc.csv	Población española por provincia	Habitantes
35167bsc.csv	Movilidad de la población durante el estado de alarma	%
statistic_id1104235_covid-19_-poblacion-que-evitaba-las-aglomeraciones-segun-edad-en-espana-2020.xlsx	Porcentaje de la población que evitaba las aglomeraciones	%

Los atributos de las dimensiones de la tabla de hechos FACT_Llamadas112 son los siguientes:

Dimensiones	Atributos
DIM_Fecha	Año, mes y día
DIM_Ambito_Geografico	Código, provincia, comarca y municipio
DIM_Tipologia	Nombre de la tipología

Aunque hay fuentes de datos que solo indican el año, otras el mes y otras el día, conviene pedagógicamente tratar las fechas completas y trasladar los datos al día 1 de enero de cada año tratado, en caso de no disponer de los niveles del mes y del día.

Así, para cargar la dimensión «DIM_Fecha» podemos hacerlo solo con las fechas que nos aparecen en las fuentes de datos o con todas las fechas desde el 01/01/2000 hasta el 01/01/2021.

4.3. Diseño físico

Para el correcto diseño físico del almacén se deben tener en cuenta los siguientes aspectos:

- El **tipo de base de datos** con el que se trabaja, puesto que cada una de ellas tiene su particularidad.
- El **diseño físico**, que debe estar orientado a generar un buen rendimiento en el procesamiento de las consultas.
- La **definición de los procesos de administración** del DW.
- La **revisión periódica del diseño físico inicial**, para validar que continúa dando respuesta a las necesidades del cliente.

Una vez que se han determinado las tablas de hechos, las dimensiones, las métricas y los atributos que existen en el modelo, se pueden determinar las claves foráneas que deben definirse en el modelo físico. En este paso, también es necesario tener en cuenta el tamaño adecuado de los atributos (por ejemplo, qué longitud tiene una cadena o si los valores numéricos contienen decimales). También es relevante acordarse de crear correctamente las claves primarias en las dimensiones.

Dado que el modelo de almacén está compuesto por más de una tabla de hechos (*facts*), también se deben revisar las dimensiones que se han definido en el diseño conceptual y en el lógico de cada *fact* y aplicar una visión conjunta del modelo. Esto permitirá definir las dimensiones comunes, como la «fecha» y el «ámbito geográfico», y así simplificar el modelo final y conseguir un rendimiento óptimo en la ejecución de los análisis.

Como es lógico, primero, se crean las tablas de dimensiones y, posteriormente, las tablas de hechos, ya que contienen atributos referenciales a aquellas. De esta forma, se crea cada una de las tablas del almacén de datos.

4.4. Dimensiones

Las dimensiones del modelo podrán estar referenciadas en las tablas de hechos utilizando sus claves primarias o, en inglés, *primary key* (PK). El modelo físico de las dimensiones es el siguiente:

- **DIM_Fecha**: corresponde a la dimensión temporal del almacén. Se dispone de datos anuales, mensuales y diarios, según la fuente tratada. La dimensión temporal es común en todo el modelo diseñado y permite analizar los hechos desde un punto de vista temporal, como el análisis de tendencias o los evolutivos. Este tipo de análisis no se puede realizar si el modelo no cuenta con una dimensión de tiempo. Se puede ampliar con más información como, por ejemplo, los días de la semana, los festivos, los semestres, los cuatrimestres, etc.

Nombre del campo	Tipo	Tamaño	Ejemplo
pk_fecha (PK)	Numérico	4	25
año	Numérico	4	2020
mes	Numérico	2	6
día	Numérico	2	20
fecha	Fecha	10	20/06/2020

- **DIM_Ambito_Geografico**: contiene los datos de los ámbitos geográficos.

Nombre del campo	Tipo	Tamaño	Ejemplo
pk_ambito_geografico (PK)	Numérico	4	105
provincia_codigo	Texto	2	43
provincia_nombre	Texto	100	Tarragona
comunidad_autonoma	Texto	100	Cataluña
comarca	Texto	100	Baix Ebre
municipio	Texto	100	L'Ampolla

- **DIM_Grupo_Edad**: contiene los datos de los grupos de edad.

Nombre del campo	Tipo	Tamaño	Ejemplo
pk_grupo_edad (PK)	Numérico	4	1
nombre	Texto	20	Grupo edad 14-24
intervalo	Texto	10	14-24

- **DIM_Medicion**: contiene los datos de la medición.

Nombre del campo	Tipo	Tamaño	Ejemplo
pk_medicion (PK)	Numérico	4	1
nombre	Texto	100	Porcentaje de la población que evitaba las aglomeraciones.
unidad_medida	Texto	20	%

- **DIM_Tipologia**: contiene los datos de la tipología de las llamadas.

Nombre campo	Tipo	Tamaño	Ejemplo
pk_tipologia (PK)	Numérico	4	1
nombre	Texto	100	Seguridad

4.5. Tablas de hechos

La composición del modelo físico de las tablas de hechos consistirá en la creación de tablas cuyos campos serán las métricas, los atributos y los atributos referenciales definidos en el modelo conceptual y en el modelo lógico. Para crear los atributos referenciales en las tablas de hechos, se

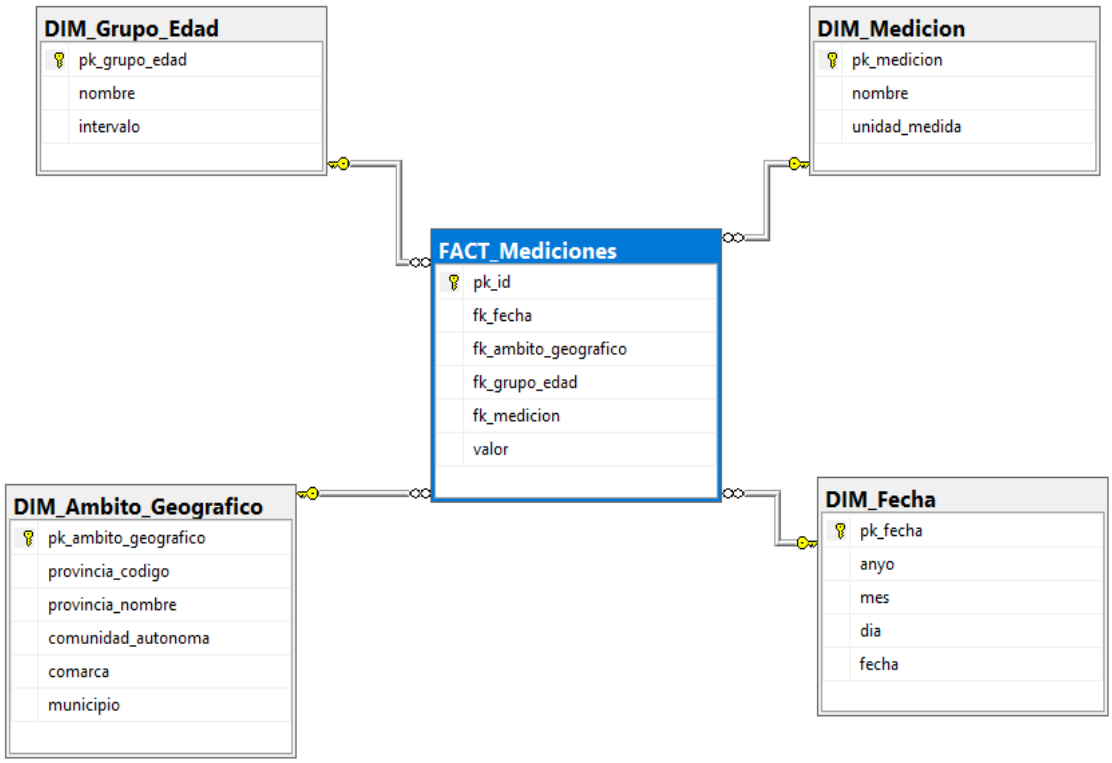
definen como claves foráneas a las primarias de las dimensiones con las que están relacionadas, siguiendo el diagrama de estrella definido.

El modelo físico de las tablas de hechos del almacén de datos para el análisis del impacto conductual está compuesto de las siguientes tablas:

- **FACT_Mediciones**: es la tabla física que contendrá la información que permitirá realizar el análisis de los datos de las mediciones obtenidas durante el estado de alarma. Tendrá los siguientes campos:

Nombre del campo	Tipo	Tamaño	Ejemplo
pk_id (PK)	Numérico	4	17
fk_fecha (FK)	Numérico	4	25
fk_ambito_geografico (FK)	Numérico	4	10
fk_grupo_edad (FK)	Numérico	4	1
fk_medicion (FK)	Numérico	4	1
valor	Numérico decimal	8	389.830

En la siguiente imagen se muestra el **diseño del modelo físico**[1] para la tabla de hechos **FACT_Mediciones**:



[1] El diseño se ha realizado utilizando la herramienta Microsoft SQL Server Management Studio.

- **FACT_Llamadas112**: es la tabla física que contendrá la información que permitirá realizar el análisis de la gestión de las llamadas de urgencia al 112 de Cataluña. Tendrá los siguientes campos:

Nombre del campo	Tipo	Tamaño	Ejemplo
pk_fk_fecha (PK y FK)	Numérico	4	20
pk_fk_ambito_geografico (PK y FK)	Numérico	4	10
pk_fk_tipologia (PK y FK)	Numérico	4	1
llamadas	Numérico	8	14

El **diseño del modelo físico** para la tabla de hechos **FACT_Llamadas112** se muestra a continuación: