Caso práctico: Almacén de datos para el análisis de indicadores de crisis en la eurozona

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 para el análisis de indicadores de crisis en la eurozona.

Índice

1. Análisis de los requisitos 2

2. Análisis de las fuentes de datos 4

Estimación de volumetría 6

3. Análisis funcional 7

4. Diseño del modelo conceptual, lógico y físico del almacén de datos 10

Diseño conceptual 10

Diseño lógico 13

Diseño físico 15

1. Análisis de los requisitos

El análisis de los requisitos 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 de indicadores de crisis en la eurozona 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, entre otros, de los siguientes objetivos:

- Analizar la evolución de indicadores de la eurozona.
- Analizar los precios de las distintas energías (electricidad y gas).

El diseño de un almacén de datos de un proyecto incluye la creación e implementación de un modelo dimensional o multidimensional, el diseño y la implementación de procesos de 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 la evolución de indicadores de la eurozona:
    - por año
    - por año y mes
    - por finalidad de consumo
    - por país
2. Analizar el precio de la energía desde diferentes perspectivas:
    - por año
    - por año y semestre
    - por país
    - por tipo de energía
    - por moneda
    - por franja de consumo
    - por tipo de impuesto
    - por unidad de medida

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 manera específica las preguntas que, como mínimo, el sistema debe ser capaz de responder:

- Relación de precios de consumo por país, ordenados por HICP de menor a mayor.
- Análisis del top cinco de fines de consumo (COICOP) durante 2022, según HICP, ordenados por HICP de mayor a menor.
- Evolución del HICP mes a mes durante 2022, para España, Portugal, Francia e Italia.
- Evolución del precio de la energía por producto y año, con el precio redondeado a dos decimales.
- Tomando el consumo del segundo semestre de 2021, calcular la estimación del precio del gas natural (producto 4100) para el mismo periodo del año 2022, por país y banda de consumo, con TAX = X_TAX (*excluding taxes and levies*) y unidad de medida = KWH, suponiendo un aumento del precio del 125 %. Mostrar tanto el precio real del periodo indicado como la estimación calculada redondeados a dos decimales.
- Tomando el consumo del segundo semestre de 2021, calcular la estimación del precio de la energía eléctrica (producto 6000) para el mismo periodo del año 2022, para España y Alemania por banda de consumo, con TAX = X_TAX (*excluding taxes and levies*) y unidad de medida = KWH, suponiendo un decremento del precio del 80 %. Mostrar tanto el precio real del periodo indicado como la estimación calculada y la diferencia entre el valor real y el estimado, redondeados a dos decimales.

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. Podéis ver a continuación un análisis detallado para cada tipo de formato.

1) **Countries.json.** Contiene los nombres de los países en orden alfabético y los elementos de código ISO 3166-1-alpha-2 en formato JSON. La estructura del fichero es la siguiente**:**

Nombre de campo	Descripción	Tipo	Ejemplo
name	Nombre de país	Texto	'Spain'
code	Código	Texto	'ES'

Total de registros: 250

2) **prc_hicp_mv12r.tsv.** Contiene la información relativa al indicador HICP. Uno de los indicadores que se usa en la eurozona para ver la evolución de la economía.

Nombre de campo	Descripción	Tipo	Ejemplo
freq	Código de frecuencia de cálculo del indicador	Texto	'M'
unit	Código de la unidad de medida del indicador	Texto	'RCH_MV12MAVR'
coicop	Código de la finalidad de consumo *Classification of individual consumption by purpose* (COICOP)	Texto	'CP01'
geo	Código geográfico	Texto	'ES'
period	Año y mes del indicador	Numérico	202206
hicp	Valor del indicador HICP	Numérico	3.8

Total de registros: 14.175

- El campo **period** está en las columnas y contiene el valor de **hicp** para cada mes.

3) **nrg_pc_202_tabular.tsv.** Contiene la información relativa a la evolución del mercado energético europeo de **gas**, por países.

Nombre de campo	Descripción	Tipo	Ejemplo
freq	Código de frecuencia de cálculo del indicador	Texto	'S'
product	Código del producto	Numérico	4100
consom	Código de la franja de consumo	Numérico	4141901
unit	Código de la unidad de medida de la energía	Texto	'KWH
tax	Código del tipo de impuestos	Texto	'I_TAX'
currency	Código de la moneda	Texto	'EUR'
geo	Código geográfico	Texto	'ES'
period	Año y semestre del indicador	Texto	'2022-S1'
pgas	Consumo gas	Numérico	0.2653

Total de registros: 1.800

- El campo **period** está en las columnas y contiene el valor de **pgas** para cada semestre.

4) **nrg_pc_204_tabular.tsv.** Contiene información relativa a la evolución del mercado energético europeo de la **electricidad** por países.

Nombre de campo	Descripción	Tipo	Ejemplo
freq	Código de frecuencia de cálculo del indicador	Texto	'S'
product	Código del producto	Numérico	6000
consom	Código de la franja de consumo	Numérico	4161901
unit	Código de la unidad de medida de la energía	Texto	'KWH
tax	Código del tipo de impuestos	Texto	'I_TAX'
currency	Código de la moneda	Texto	'EUR'
geo	Código geográfico	Texto	'ES'
period	Año y semestre del indicador	Texto	'2022-S1'
pelec	Consumo electricidad	Numérico	0.2653

Total de registros: 1.845

- El campo **period** está en las columnas y contiene el valor de **pelec** para cada semestre.

5) **COICOP.xml.** Contiene información relativa al código y descripción de la finalidad de consumo (COICOP, *Classification of individual consumption by purpose*) en formato XML.

Nombre de campo	Descripción	Tipo	Ejemplo
code	Código de la finalidad de consumo	Texto	'CP01113'
description	Descripción de la finalidad de consumo	Texto	'Bread'

Total de registros: 951

6) **consumption_band.csv.** Contiene información relativa al código y descripción de la franja de consumo de energía en formato CSV.

Nombre de campo	Descripción	Tipo	Ejemplo
code	Código de la franja de consumo	Texto	'4161901'
description	Descripción de la franja de consumo	Texto	'Band DA : Consumption < 1 000 kWh'

Total de registros: 8

Estimación de volumetría

En los proyectos de diseño de 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 inicial de los datos sería la siguiente:

Fichero	Registros	Valores	Datos
Countries.json	250	2	500
prc_hicp_mv12r.tsv	14.175	40	567.000
nrg_pc_202_tabular.tsv	1.800	38	68.400
nrg_pc_204_tabular.tsv	1.845	38	70.110
COICOP.xml	951	2	1.902
consumption_band.csv	8	2	16
Total	**19.029**	**122**	**707.928**

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 requisitos exigibles son aquellos que pide el enunciado, mientras que los deseables son los que complementan la actividad.

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

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

#	Requisito	Prioridad	Exigible/ deseable
1	Se extraerá de manera adecuada la información de las fuentes de datos.	1	E
2	Se creará un almacén de datos.	1	E
3	Se cargará la información para realizar el análisis de indicadores de crisis en la eurozona.	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 requisitos funcionales, como los que se muestran a continuación:

- Análisis de viabilidad y análisis de riesgos.
- Creación de procesos de calidad de datos.
- Creación de *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 de 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 requisitos funcionales sería mucho mayor, como puede ser la administración de seguridad en cuanto a datos y usuarios.

En términos de la arquitectura funcional, existen los siguientes elementos:

- Las fuentes de datos de las que se dispone son las siguientes:
  - un fichero en formato CSV
  - tres ficheros en formato TSV
  - un fichero en formato JSON
  - un fichero en formato XML

- 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 de ETL.

  - *Data mart* para el análisis de los indicadores de crisis en la eurozona. Al centrarnos 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. La siguiente figura 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, al tener una única base de datos, simulamos con los prefijos en los nombres de las tablas intermedia, cuyos objetos se identificarán con un prefijo en los nombres (IN_).

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

Diseño conceptual

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

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

- **Evolución de indicadores de la eurozona**. Hace referencia a la información relevante sobre indicadores de la eurozona.

- **Precio de la energía**. Hace referencia a la información relevante sobre precios de las distintas energías.

El **análisis de la evolución de indicadores de la eurozona** determina el diseño de la primera tabla de hechos, como se puede observar a continuación:

Tabla de hechos	Descripción
FACT_EUROZONE_INDICATORS	Análisis de los indicadores de la eurozona

En la siguiente tabla, se indica la métrica de la tabla de hechos FACT_EUROZONE_INDICATORS.

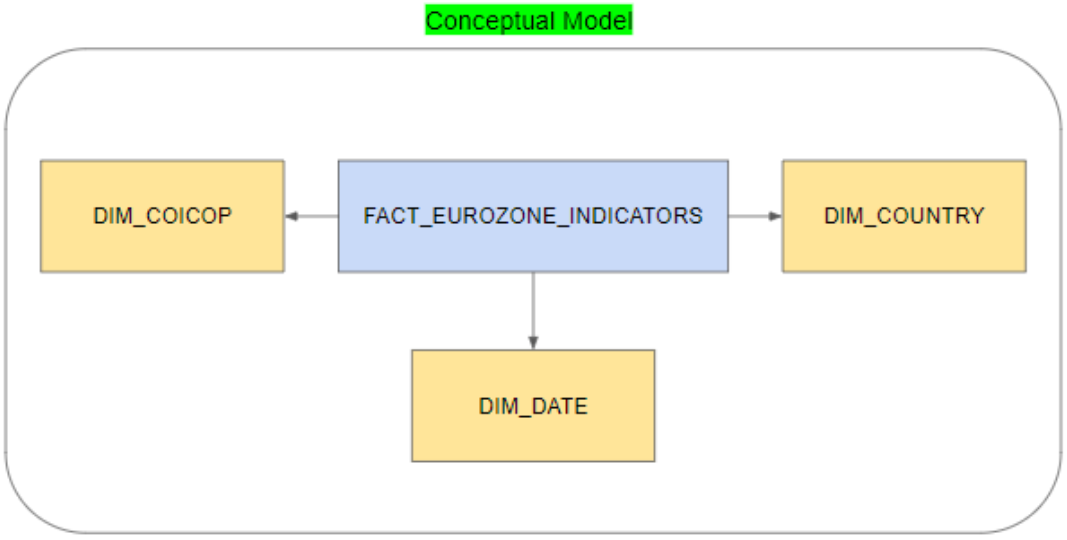
Métricas	Descripción
HICP	*Harmonised index of consumer prices*. Este es un indicador estadístico de precios de consumo de los Estados de la zona del euro, elaborado con los mismos criterios metodológicos en toda la zona, que publica Eurostat y que se utiliza para medir la inflación de los precios de consumo. Está «armonizado» porque todos los países de la Unión Europea siguen la misma metodología. Esto garantiza que los datos de un país puedan compararse con los de otro. Véase descripción en ecb.europa.eu (en

	inglés)

La métrica de esta tabla de hechos podrá analizarse desde las diferentes perspectivas, a partir de las siguientes dimensiones:

Dimensiones	Descripción
Tiempo	Año y mes del registro del indicador
País	País al que hace referencia el indicador
Finalidad de consumo	Finalidad de consumo (*Classification of individual consumption by purpose*, COICOP)

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



Este modelo considera las fuentes de datos siguientes:

- Countries.json
- prc_hicp_mv12r.tsv
- COICOP.xml

Para el **análisis del precio de la energía** se identifica una segunda tabla de hechos; es la siguiente:

Tabla de hechos	Descripción
FACT_ENERGY_PRICE	Análisis del precio de las distintas energías en la eurozona

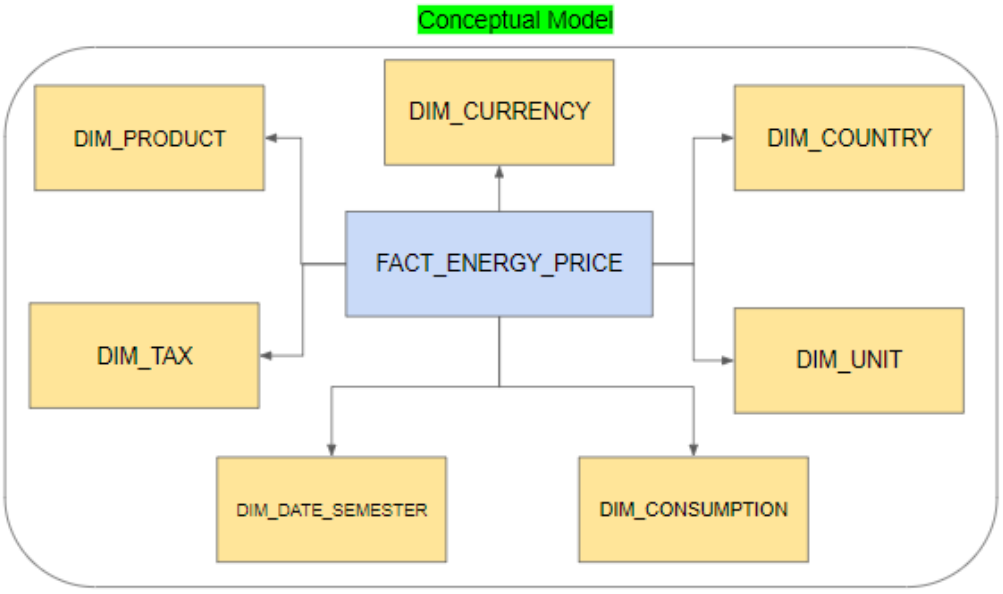
En la siguiente tabla, se indica la métrica de la tabla de hechos FACT_ENERGY_PRICE.

Métricas	Descripción
Precio	Precio energía

La métrica de esta tabla de hechos podrá ser analizada desde las diferentes perspectivas, a partir de las siguientes dimensiones

Dimensiones	Descripción
Tiempo	Fecha de registro del precio de la energía
País	País al que hace referencia el precio de la energía
Tipo de energía (*product*)	Tipo de energía a la que corresponde el precio
Moneda	Moneda en la que se expresa el precio de la energía
Tipo de impuesto (*tax*)	Tipo de impuesto en el que se expresa el precio de la energía
Franja de consumo	Franja de consumo a la que pertenece el precio de la energía
Unidad de medida	Unidad de medida en la que se expresa el precio de la energía

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



Este modelo considera las fuentes de datos siguientes:

- Countries.json
- nrg_pc_202_tabular.tsv.gz
- nrg_pc_204_tabular.tsv.gz
- consumption_band.csv

Diseño lógico

Una vez obtenido el modelo conceptual del almacén de datos del análisis de indicadores de crisis en la eurozona, pasamos a realizar su diseño lógico.

Teniendo en cuenta que vamos a utilizar tecnología relacional y que el modelo de datos va a ser el multidimensional, pasamos a describir el modelo lógico en términos de tablas, atributos y claves primarias y foráneas.

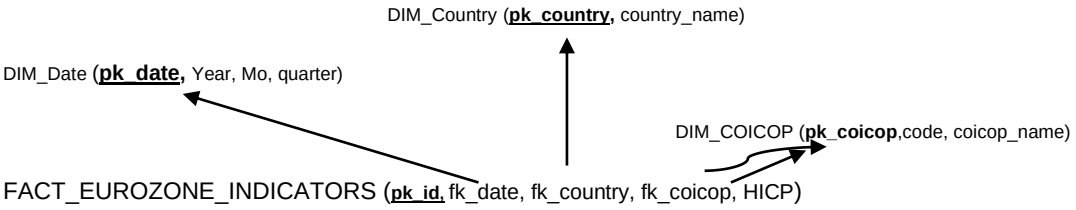
El primer paso corresponde a identificar las métricas de cada hecho que va a representarse mediante una tabla de hechos. En nuestro caso, hemos identificado estos dos hechos y sus métricas:

Tabla de hechos	Métricas
FACT_EUROZONE_INDICATORS	HICP
FACT_ENERY_PRICE	Precio

Después se detallan los atributos de las dimensiones de cada hecho. Específicamente, los atributos de las dimensiones de la tabla de hechos FACT_EUROZONE_INDICATORS se muestran en la siguiente tabla:

Dimensiones	Atributos
DIM_DATE	Code, Year, Mo, quarter
DIM_COUNTRY	Code, country_name
DIM_COICOP	Code, coicop_name

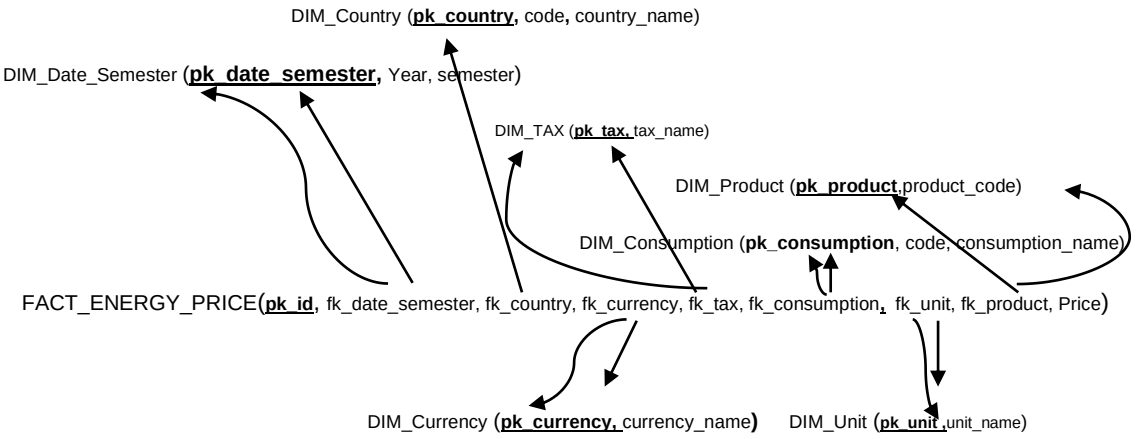
La representación visual del modelo lógico para el análisis de la **evolución de indicadores de la eurozona** sería la siguiente:



Procedemos igual con el hecho FACT_ENERY_PRICE. Se muestran en la siguiente tabla los atributos de las dimensiones de esta tabla de hechos:

Dimensiones	Atributos
DIM_DATE_SEMESTER	Code, Year, semester
DIM_COUNTRY	Code, country_name.
DIM_PRODUCT	Code, product_code
DIM_TAX	Código, tax_name
DIM_CURRENCY	Code, currency_name
DIM_UNIT	Code, unit_name
DIM_CONSUMPTION	Code, consumption_name

La representación visual del **modelo lógico** de la tabla de hechos y sus dimensiones para el **análisis de los precios de las energías en la eurozona** es la siguiente:



Diseño físico

Una vez que se han determinado las tablas de hechos, las dimensiones, las métricas y los atributos que existen en el modelo lógico, podemos pasar a realizar el diseño físico, lo que significa obtener una implementación del modelo lógico en términos del sistema gestor de bases de datos elegido.

Además, para el correcto diseño físico del almacén, se deben tener en cuenta los siguientes aspectos:

- El **sistema gestor de bases de datos** con el que vamos a trabajar implementará de una manera concreta los distintos elementos del modelo lógico.
- El ajuste del **diseño físico** a las particularidades de nuestro sistema gestor de bases de datos, con el fin de obtener buen rendimiento en el procesamiento de consultas.
- La **revisión periódica del diseño físico inicial**, para validar que continúa dando respuesta a las necesidades del cliente.

Puesto que utilizaremos SQL Server, y este es un sistema gestor de bases de datos relacional, en esta etapa deberemos tener en cuenta, entre otras cosas, la implementación de las claves primarias y foráneas en las tablas de hechos y en las de dimensiones.

En este paso, también es necesario tener en cuenta el tamaño adecuado de los atributos (por ejemplo, la longitud de los campos de textos o si los valores numéricos contienen decimales).

Para ello, vamos a detallar los tipos de datos de cada campo que forman parte de las tablas de hechos y 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 para determinar si en el modelo del almacén existirán dimensiones comunes o conformadas, como DIM_COUNTRY, 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 manera, se crea cada una de las tablas del almacén de datos.

Dimensiones

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

- **DIM_Country**: contiene los datos de los países. La dimensión es común en todo el modelo diseñado y permite analizar los hechos desde un punto de vista geográfico.

Nombre de campo	Tipo	Tamaño	Ejemplo
pk_country (PK)	Numérico	8	1
code	Texto	8	ES
country_name	Texto	100	Spain

La dimensión conformada DIM_COUNTRY, se utilizará tanto para analizar los indicadores de la eurozona como para analizar los precios de la energía.

- **DIM_Date**: corresponde a la dimensión temporal para el análisis de la información de indicadores de la eurozona. La dimensión temporal 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.

Nombre de campo	Tipo	Tamaño	Ejemplo
pk_date (PK)	Numérico	8	255467
Year	Numérico	4	2020
Mo	Numérico	2	2
Quarter	Numérico	1	1

- **DIM_Date_Semester**: corresponde a la dimensión temporal para el análisis de la información de precios de la energía. En este modelo las dos tablas de hechos tienen distintos niveles de temporalidad, mensual y semestral, por ello es necesario tener dos tablas de dimensiones temporales.

Nombre de campo	Tipo	Tamaño	Ejemplo
pk_date_semester (PK)	Numérico	8	255467
Year	Numérico	4	2020
Semester	Numérico	1	1

- **DIM_COICOP**: contiene los datos de las diferentes finalidades de consumo en las que se clasifica el indicador HICP.

Nombre de campo	Tipo	Tamaño	Ejemplo
pk_coicop (PK)	Numérico	8	1
code	Texto	25	CP01
coicop_name	Texto	100	Food and non-alcoholic beverages

- **DIM_TAX**: contiene los datos de los diferentes tipos de impuestos.

Nombre de campo	Tipo	Tamaño	Ejemplo
pk_tax (PK)	Numérico	8	1
tax_name	Texto	50	I_TAX

- **DIM_PRODUCT**: contiene los datos de los diferentes tipos de energías.

Nombre de campo	Tipo	Tamaño	Ejemplo
pk_product (PK)	Numérico	8	1
product_code	Numérico	8	4100

- **DIM_CONSUMPTION**: contiene los datos de las distintas franjas de consumo.

Nombre de campo	Tipo	Tamaño	Ejemplo
pk_consumption (PK)	Numérico	8	1
code	Numérico	8	4161901
consumption_name	Texto	100	Band DA : Consumption < 1 000 kWh

- **DIM_CURRENCY**: contiene los datos de las distintas monedas.

Nombre de campo	Tipo	Tamaño	Ejemplo
pk_currency(PK)	Numérico	8	1
currency_name	Texto	10	EUR

- **DIM_UNIT**: contiene los datos de las distintas unidades de medida.

Nombre de campo	Tipo	Tamaño	Ejemplo
pk_unit (PK)	Numérico	8	1
unit_name	Texto	10	KWH

Tablas de hechos

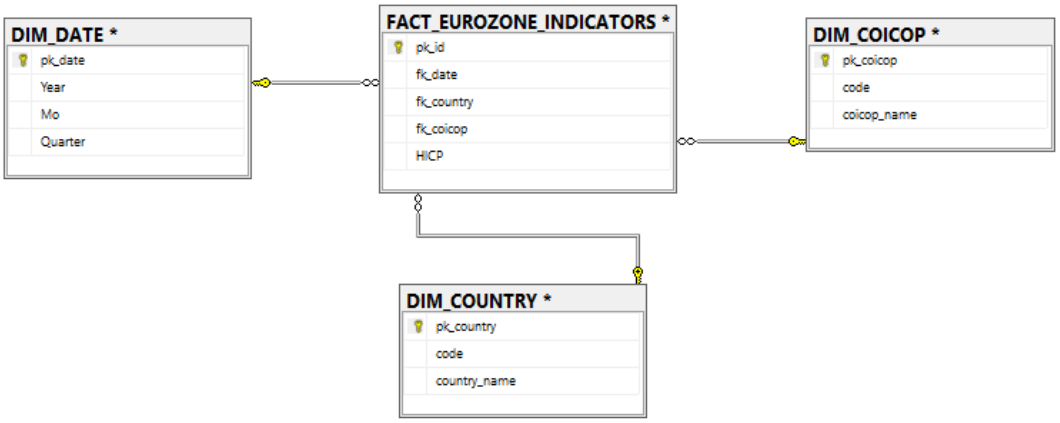
La composición del modelo físico de las tablas de hechos consistirá en la creación de tablas, cuyos campos serán las métricas, los atributos y los atributos referenciales definidos en el modelo conceptual y en el modelo lógico. Para crear los atributos referenciales en las tablas de hechos, se definen como *claves foráneas* las primarias de las dimensiones con las que están relacionadas, siguiendo el diagrama en estrella definido.

El modelo físico de las tablas de hechos del almacén de datos para el análisis de indicadores de crisis en la eurozona está compuesto de las siguientes tablas:

- **FACT_EUROZONE_INDICATORS**: es la tabla física que contendrá la información que permitirá realizar el análisis de indicadores de la eurozona. Tendrá los siguientes campos:

Nombre campo	Tipo	Tamaño	Ejemplo
pk_id (PK)	Numérico	8	1
fk_date(FK)	Numérico	8	255467
fk_country (FK)	Texto	20	ES
fk_coicop (FK)	Texto	8	CP01
HICP	Numérico	12,2	0.25

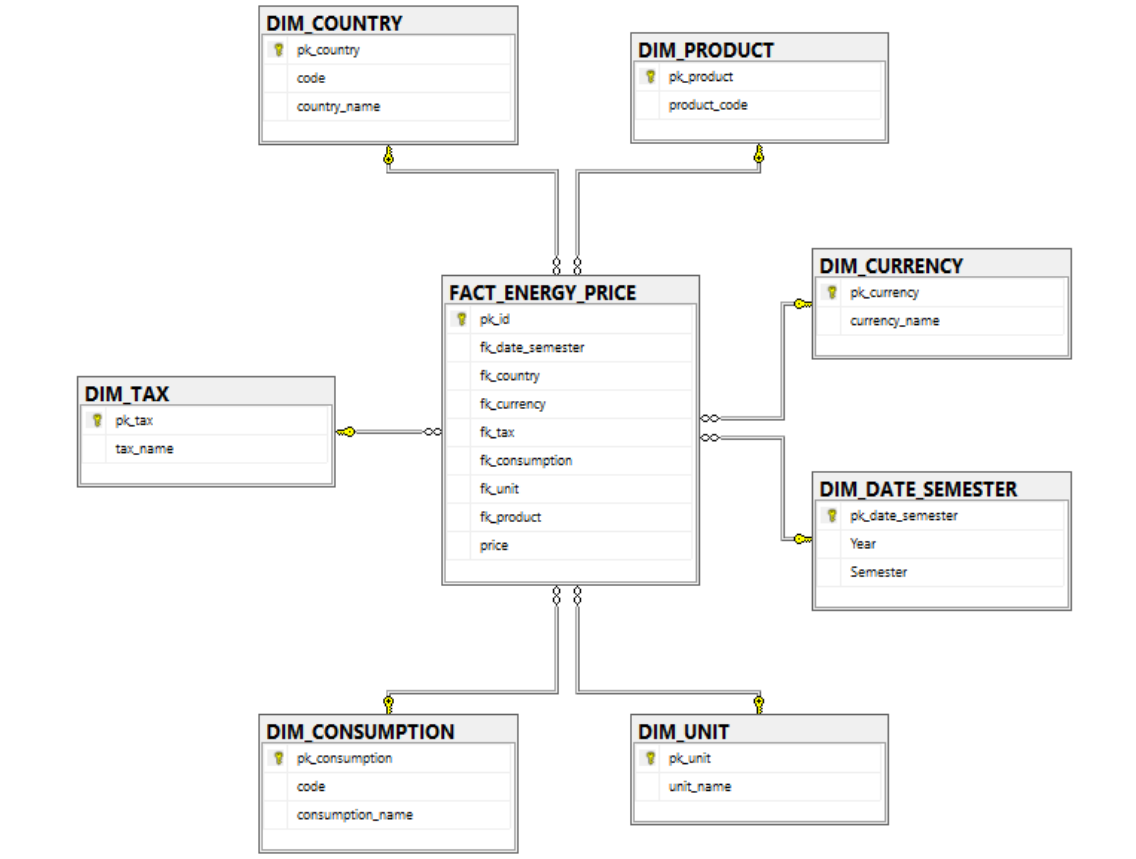
En la siguiente imagen se muestra una posible implementación del **diseño del modelo físico** para la tabla de hechos **FACT_EUROZONE_INDICATORS:**



- **FACT_ENERY_PRICE**: es la tabla física que contendrá la información que permitirá realizar el análisis del precio de las energías en la eurozona. Entre otros, tendrá los siguientes campos:

Nombre de campo	Tipo	Tamaño	Ejemplo
pk_id (PK)	Numérico	8	1
fk_date_semester (FK)	Numérico	8	255467
fk_country (FK)	Texto	8	231
fk_currency (FK)	Numérico	8	1
fk_tax(FK)	Numérico	8	1
fk_consumption (FK)	Numérico	8	4161901
fk_unit (FK)	Numérico	8	1
fk_product (FK)	Numérico	8	4100
Price	Numérico	12,2	0.25

En la siguiente imagen se muestra una posible implementación del **diseño del modelo físico** para la tabla de hechos **FACT_ENERY_PRICE:**


DIM_COUNTRY
pk_country
code
country_name
DIM_PRODUCT
pk_product
product_code
DIM_CURRENCY
pk_currency
currency_name
FACT_ENERGY_PRICE
pk_id
fk_date_semester
fk_country
fk_currency
fk_tax
fk_consumption
fk_unit
fk_product
price
DIM_TAX
pk_tax
tax_name
DIM_DATE_SEMESTER
pk_date_semester
Year
Semester
DIM_CONSUMPTION
pk_consumption
code
consumption_name
DIM_UNIT
pk_unit
unit_name