

Caso práctico: Almacén de datos para el análisis del impacto ambiental y el consumo energético derivados de la actividad económica

PRA1–Análisis y diseño del *data warehouse*

Presentación

La PRA1 consiste en el análisis y el diseño multidimensional de un almacén de datos para examinar el impacto ambiental y el consumo energético derivados de la actividad económica.

Para llevar a cabo su desarrollo, se ha publicado la siguiente información:

- **Caso práctico** (22.510_Enunciado_PRA (PID_00283254).pdf). Documento con la información del caso de uso para desarrollar un almacén de datos que permita analizar la información del impacto ambiental y del consumo energético. En el documento se describen el contexto, los usuarios potenciales y las fuentes de datos. También se realiza una descripción del enunciado de la práctica con cada una de las fases para construir el almacén de datos, así como los programas que se utilizarán y la bibliografía.
- **Fuentes de datos** (fuentes.zip). Fichero comprimido con las fuentes de datos disponibles para desarrollar el caso práctico.

Descripción

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 ofrezca soporte al análisis del impacto ambiental y el consumo energético derivados de la actividad económica.

Para ello **se pide** llevar a cabo los siguientes puntos:

- 1) **Análisis de los requerimientos** que incluya las preguntas a las que el sistema debe dar respuesta. Se deben definir un mínimo de cinco preguntas y un máximo de diez.
- 2) **Análisis de las fuentes de datos** proporcionadas.
- 3) **Análisis funcional** que proponga el tipo de arquitectura para la factoría de información que mejor se adecue al proyecto.
- 4) **Diseño del modelo conceptual, lógico y físico** del almacén de datos. Se deben identificar y diseñar las tablas de hechos (*facts*), las dimensiones del análisis (*dimensions*) y los atributos que permitan tener el nivel de granularidad suficiente para implementar los requerimientos.

El documento de la solución de esta actividad (PRA1) debe incluir una descripción y una justificación de todas las acciones realizadas en los puntos anteriormente enumerados.

Criterios de evaluación

La nota final se calculará a partir de la suma de los siguientes apartados:

- 1) Análisis de los requerimientos (10 %)
- 2) Análisis de fuentes de datos (15 %)
- 3) Análisis funcional (15 %)
- 4) Diseño conceptual (20 %)
- 5) Diseño lógico (20 %)
- 6) Diseño físico (20 %)

Guía de muestra

Con el fin de ayudar a alcanzar los objetivos planteados de la PRA1 se desarrolla esta guía. Esta muestra, con un ejemplo, cómo debe llevarse a cabo y cuáles son los pasos que hay que seguir:

- 1) análisis de los requerimientos,
- 2) análisis de las fuentes de datos,
- 3) análisis funcional, y
- 4) diseño completo del modelo.

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 ambiental y del consumo energético 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, del siguiente objetivo:

- Analizar el balance energético mundial de los países productores y consumidores de energía.

A continuación, se indica la información necesaria identificada para analizarlo desde diferentes perspectivas:

- por fecha,
- por país,
- por categoría de energía (producto) o
- por unidad de medida.

El estudiante, en este punto, deberá completar la definición de los requerimientos. Para ello tendrá que identificar otras necesidades y plantear otras preguntas a las que el sistema deberá responder.

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 para cada tipo de formato.

A continuación, se muestra un ejemplo:

- 1) **02002.xlsx**. Encuesta del gasto de la industria en protección ambiental. Serie 2008-2018. Evolución de la inversión en protección ambiental por tipo de equipo e instalación, ámbito medioambiental y sector de actividad económica.

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

Nombre campo	Tipo	Ejemplo
Periodo	Numérico	2018
Sector de actividad económica	Texto	«C. industria manufacturera»
Tipo de equipo e instalación	Texto	«INVERSIÓN EN EQUIPOS E INSTALACIONES INDEPENDIENTES»
Ámbito medioambiental	Texto	«Protección del aire y el clima»
Comunidad autónoma	Texto	«Asturias, Principado de»
Inversión	Numérico	6.280.443

Observaciones:

- Las columnas desde la E hasta la U corresponden a la inversión de cada comunidad autónoma.
- La columna V contiene el agregado de *Total nacional*.
- La unidad de la inversión son euros.

Total de registros: 154 registros.

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
02002.xlsx	154	22	3.388
...
Total			

El estudiante en este punto deberá completar la definición del resto de fuentes proporcionadas.

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 pide 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 algunos de 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
...	...		

El estudiante, en este punto, deberá completar la tabla de los requerimientos funcionales, asignar las prioridades y elegir la arquitectura funcional que considere más adecuada para el caso de estudio de todas las posibles y estudiadas en los módulos teóricos de la asignatura.

También deberá incorporar, en términos de la arquitectura funcional, cuales son elementos considerados para este caso práctico.

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 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 uno de los hechos que se deben considerar es:

- **Mediciones ambientales.** Hace referencia a ciertas mediciones ambientales relevantes para un desarrollo sostenible.

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

Tabla de hechos	Descripción
FACT_EnvironmentalMeasurements	Mediciones ambientales para un desarrollo sostenible

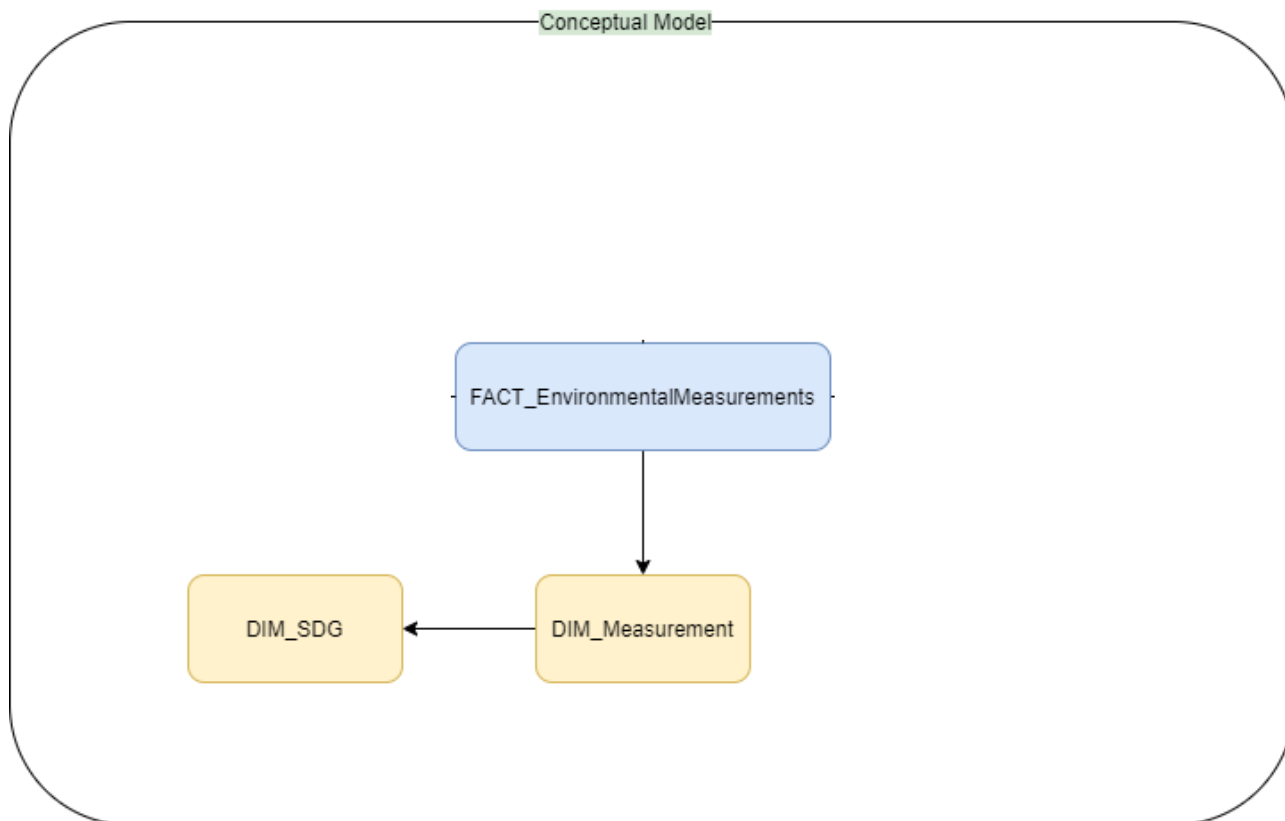
Las **métricas** de la tabla de hechos «FACT_EnvironmentalMeasurements» son las siguientes:

Métricas	Descripción
Value	Valor de la medición

Las métricas de esta tabla de hechos podrán ser analizadas desde las diferentes perspectivas, a partir de las dimensiones. Por ejemplo, las dimensiones «SDG» y «Measurement».

Dimensiones	Descripción
SDG (Sustainable Development Goals)	Objetivos de desarrollo sostenible (ODS) con los que se relaciona la medición
Measurement	Medición a tratar

En las dimensiones con jerarquía se puede optar por un **diseño en copo de nieve**, en el que 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 lógico

Una vez obtenido el modelo conceptual del almacén de datos para el análisis del impacto ambiental, pasamos a realizar el diseño lógico del mismo.

A continuación, se muestra, a modo de ejemplo, una tabla con alguna de las métricas identificadas en el diseño conceptual de la tabla de hechos «FACT_EnvironmentalMeasurements»:

Tabla de hechos	Métricas
FACT_EnvironmentalMeasurements	Value

Después se detallan algunos de los atributos descriptores de las dimensiones de cada hecho:

Dimensiones	Atributos
DIM_SDG	Código, nombre, descripción
DIM_Measurement	Código, nombre de la medición y unidad de medida

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 cual significa obtener una implementación del modelo lógico en términos del sistema gestor de bases de datos escogido.

En este punto se debería reflexionar acerca de las consideraciones previas que hay que tener en cuenta para el diseño físico, como en el siguiente ejemplo:

- El **sistema gestor de bases de datos** con el que vamos a trabajar implementará de una manera concreta los distintos elementos del modelo lógico.

Para ello, vamos a detallar los tipos de datos de cada campo que forma parte de las tablas de hechos y dimensiones.

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 identificadas es el siguiente:

- **DIM_SDG**: contiene los datos de los objetivos de desarrollo sostenible.

Nombre campo	Tipo	Tamaño	Ejemplo
pk_sdg (PK)	Numérico	4	13
sdg_name	Texto	50	«Acción por el clima»
sdg_description	Texto	500	«El cambio climático es un reto global que no respeta las fronteras nacionales»

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

Nombre campo	Tipo	Tamaño	Ejemplo
pk_measurement (PK)	Numérico	4	1
measurement_code	Texto	100	«Protección del aire y el clima»
measurement_name	Texto	200	«Inversión en protección ambiental»
Unit	Texto	25	«euros»
fk_sdg	Numérico	4	13

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 copo de nieve definido.

El modelo físico de las tablas de hechos del almacén de datos para el impacto ambiental y el consumo energético está compuesto, entre otras, de las siguientes tablas:

- **FACT_EnvironmentalMeasurements**: es la tabla física que contendrá la información que permitirá realizar el análisis de los datos de las mediciones ambientales para un desarrollo sostenible. Tendrá los siguientes campos:

Nombre campo	Tipo	Tamaño	Ejemplo
pk_id (PK)	Numérico	4	17
fk_date (FK)	Numérico	4	25
fk_region (FK)	Numérico	4	10
fk_activitysector (FK)	Numérico	4	1
fk_typeequipinstall (FK)	Numérico	4	1
fk_measurement (FK)	Numérico	4	1
value	Numérico	8	6.390.920

Formato y fecha de entrega

La entrega final de esta actividad debe realizarse enviando un único mensaje al buzón de «Registro de AC» del apartado «Evaluación» del aula. Junto con el mensaje se enviará un único archivo en formato Word o PDF con la solución de la PRA1. El nombre del archivo debe ser la composición del nombre de usuario y «_BDA_PRA1». Por ejemplo, si el nombre de usuario es «bantich», el nombre del archivo debe ser «bantich_BDA_PRA1.pdf».

La fecha máxima de entrega es el 9/11/2021 a las 23:59 horas.