Caso práctico: almacén de datos para el análisis de la cobertura de inmunización

PRA1-Ánalisis y diseño del *data warehouse*

Presentación

Para desarrollar la parte práctica, en el aula de la asignatura se ha publicado la siguiente información:

- Caso Práctico (PID_00267618_Enunciado_PRA_IB.pdf). Documento del caso de uso para el desarrollo de un almacén de datos que permita analizar la información de la cobertura de inmunización. En el documento se describen: el contexto, los usuarios potenciales y las fuentes de datos del caso práctico. También se realiza una descripción del enunciado de la práctica con cada una de las fases para la construcción del 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.

La PRA1 consiste en el análisis y el diseño multidimensional de un almacén de datos para el análisis de la cobertura de inmunización.

Se pide:

Tal como se indica en la actividad del aula y en el enunciado del caso práctico, esta primera parte de la práctica consiste en:

- Análisis de los requerimientos que describa las preguntas a las que el sistema debe de dar respuesta.
- Análisis de las fuentes de datos proporcionadas.
- Análisis funcional en el que se proponga el tipo de arquitectura para la factoría de la información que mejor se adecue al proyecto.
- Diseño del modelo conceptual, lógico y físico del almacén de datos: se deben identificar, diseñar e implementar las tablas de hecho (*facts*), las dimensiones de análisis (*dimensions*) y los atributos que permitan tener el nivel de granularidad suficiente para implementar los requerimientos.

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 funcionamiento del almacén de datos para el análisis de la cobertura de inmunización.

La solución de la PRA1 debe incluir:

- Cada uno de los puntos solicitados.
- Descripción y justificación de todas las acciones realizadas.

Guía de muestra

Con el fin de ayudar a alcanzar los objetivos planteados de la PRA1 se desarrolla esta guía. La guía servirá de muestra para saber cómo realizar las tareas anteriormente descritas; es decir, según el contexto del caso de uso, cómo se debe realizar el análisis de requerimientos y de fuentes de datos, el análisis funcional y el diseño completo del modelo.

Para la entrega de la primera práctica, el estudiante, siguiendo la guía de muestra, deberá realizar los análisis solicitados que permitan completar el análisis y el diseño del almacén de datos de la cobertura de inmunización.

I. Análisis de requerimientos

El análisis de requerimientos se basa en identificar las necesidades que tiene una organización particular con respecto al análisis de la información.

La necesidad principal de la Organización Mundial de la Salud (OMS) es disponer de información integrada para analizarla y difundirla con el objetivo de que ayude a todos los usuarios potenciales a tomar decisiones que garanticen el cumplimiento de los objetivos en lo relativo a la cobertura de vacunación definido en el Plan de Acción Mundial sobre Vacunas (GVAP).

Según el contexto del caso práctico, una de las necesidades identificadas que nuestro sistema deberá cubrir es la siguiente:

- Conocer el grado de consecución de los objetivos en el compromiso definido en el GVAP.

Una forma de definir los requerimientos es plantear las preguntas a las que nuestro sistema deberá responder, entre otras:

- ¿Qué cobertura de inmunización alcanzan los países miembros a lo largo del tiempo?
- ¿Qué países miembros han conseguido cumplir el objetivo 1 de la meta 2020 del GVAP?

El estudiante, para resolver este punto, deberá completar la definición de requerimientos, para ello deberá identificar otras necesidades y plantear otras preguntas que deberá responder el sistema.

II. Análisis de 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 cargarse.

A continuación, se realiza el análisis de dos de las fuentes de datos proporcionadas:

- Fichero coverage_estimates_series.xls: contiene las series cronológicas de coberturas de inmunización por región de la OMS, Estados miembros y vacunas desde 1966 hasta 2018.

Nombre campo	Tipo	Ejemplo
WHO_REGION	Texto	“EMR”
ISO_code	Texto	“AFG”
Cname	Texto	“Afghanistan”
Continent	Texto	“Asia”
Vaccine	Texto	“DTP1”
Year	Numérico	2012
Percent_covrage	Numérico	99
Asterisc	Numérico	*

Máximo de registros: 53.646 registros.

De los datos sobre la cobertura de inmunización contenida en el Excel, se observa que:

1. No todos los países miembros informan sobre la cobertura de inmunización de todas las vacunas. Por ejemplo, de España no hay información sobre la cobertura de la vacuna BCG.
2. El carácter asterisco (“*”) en el campo Asterisc significa una cobertura del 100%.

- Fichero REGION.json: contiene información de las regiones de la OMS en formato .json.

  La estructura del fichero .json es la siguiente:

Nombre etiqueta	Ejemplo
Código región	"AFR"
Nombre	"Africa"

Estimación de volumetría

La estimación del volumen de en el almacén de datos para la carga de los dos ficheros anteriormente analizados sería la siguiente:

Fuente de datos	Valores a almacenar	Total registros
coverage_estimates_series 1 fichero anual 194 países miembros (6 regiones OMS) 45 vacunas	Años: 53 datos	1 fichero × 194 países × 45 vacunas × 53 datos = 462.690
REGION.json		58
	TOTAL	462.748

El estudiante, para resolver este punto, deberá completar la definición del resto de fuentes proporcionadas.

III. Análisis funcional

En el análisis funcional se debe proponer el tipo de arquitectura para la factoría de información que mejor se adecue al caso de estudio. 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 son aquellos que complementan la actividad.

Por otro lado, en términos de la escala de prioridades, asignamos una prioridad de 1 a 3, en la que 1 es completamente prioritario para la actividad y 3 no es prioritario para la actividad.

En la tabla se describen algunos de los requerimientos funcionales teniendo en cuenta las consideraciones del caso práctico que estamos desarrollando:

	Requerimiento	Prioridad	Exigible / deseable
1	Se extraerá de manera adecuada la información de las fuentes de datos (se considerará solo la información relevante).	1	E
2	Se creará un almacén de datos.	1	E

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

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

Para el correcto desarrollo del almacén de datos es preciso definir los hechos (*facts*), las dimensiones de análisis (*dimensions*), las métricas y los atributos que nos permitan dar respuesta a las preguntas que se han definido en el análisis de requerimientos.

Diseño conceptual

Del análisis de la fuente de datos coverage_estimates_series, que contiene datos sobre inmunización de cada año, se determina que uno de los hechos que hay que analizar es la **cobertura de inmunización**.

Teniendo en cuenta los requerimientos identificados, se analizará el hecho y la cobertura de inmunización para resolver la necesidad de los usuarios de **analizar la cobertura de inmunización a lo largo del tiempo**.

El análisis de la cobertura de inmunización determina el diseño de la siguiente tabla de hechos:

Tabla de hechos	Descripción
fact_cobertura	Datos de la cobertura de inmunización

Una de las métricas de la tabla de hechos fact_cobertura es la que se incluye en la siguiente tabla:

Métricas	Descripción
Cobertura	Cobertura de inmunización

Las métricas que se identifiquen para la tabla de hechos fact_cobertura podrán ser analizadas desde las diferentes perspectivas por medio de las siguientes dimensiones:

Dimensiones	Descripción
Geografía	Distribución geográfica de la OMS a la que corresponde la información de la cobertura de inmunización.
Vacunas	Vacunas incluidas en el calendario de vacunación.

A partir de las dimensiones y de la tabla de hechos identificados se construye el modelo conceptual, en el que tanto las dimensiones como los hechos son entidades independientes que forman parte de nuestro modelo de estrella.

El diseño conceptual con forma de estrella para la tabla de hechos y las dimensiones identificadas en la guía de muestra es:

Diseño lógico

Una vez obtenido el modelo conceptual del almacén de datos para el análisis de la cobertura de inmunización, el siguiente paso es definir el modelo lógico e identificar las métricas y los atributos. Los atributos, junto con las métricas, nos permitirán responder a los requerimientos.

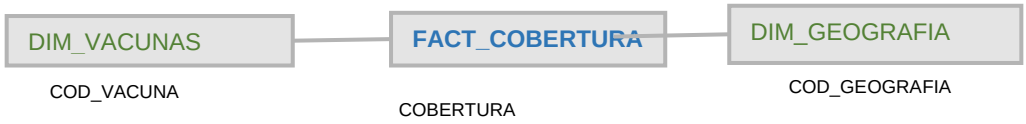
A continuación, se muestra una tabla con la métrica identificada en el diseño conceptual de la tabla de hecho FACT_COBERTURA.

Tabla de hechos	Métricas
FACT_COBERTURA	COBERTURA

En la siguiente tabla, se muestran los atributos descriptores con las referencias a las dimensiones de la tabla de hechos FACT_COBERTURA:

Dimensiones	Atributos descriptores
DIM_GEOGRAFIA	COD_GEOGRAFIA
DIM_VACUNAS	COD_VACUNA

En la siguiente imagen se muestra el diseño con forma de estrella del modelo lógico propuesto para la tabla de hechos FACT_COBERTURA.



Diseño físico

Una vez determinadas qué tablas de hechos, dimensiones, métricas y atributos existen en nuestro modelo, podemos determinar las claves primarias de las dimensiones y 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 de las tablas del modelo (por ejemplo, qué longitud tiene una cadena o si los numéricos contienen decimales).

Dimensiones

- DIM_GEOGRAFIA: contiene los datos de los 194 países miembros de las 6 regiones de la Organización Mundial de la Salud

Nombre campo	Tipo	Tamaño	Ejemplo
SK_DIM_GEOGRAFIA (CP/PK)	**Numérico**	**3**	**1**
COD_REGION	Texto	4	“AFR”
DESC_REGION	Texto	30	“AFRICA”
SK_DIM_PAIS	Numérico	3	5
DESC_PAIS	Texto	300	“ANGOLA”

- DIM_VACUNA: contiene datos relativos a las vacunas que el Plan de Acción Mundial sobre Vacunas quiere supervisar. Añadimos el campo GVAP para identificar las vacunas que tienen marcados objetivos de consecución de cobertura en el GVAP.

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_VACUNA (CP/PK)	**Numérico**	**3**	**1**
TIPO_VACUNA	Texto	50	“BCG”
NOMBRE_VACUNA	Texto	300	“Bacille Calmette-Guérin”
GVAP	Texto	1	1

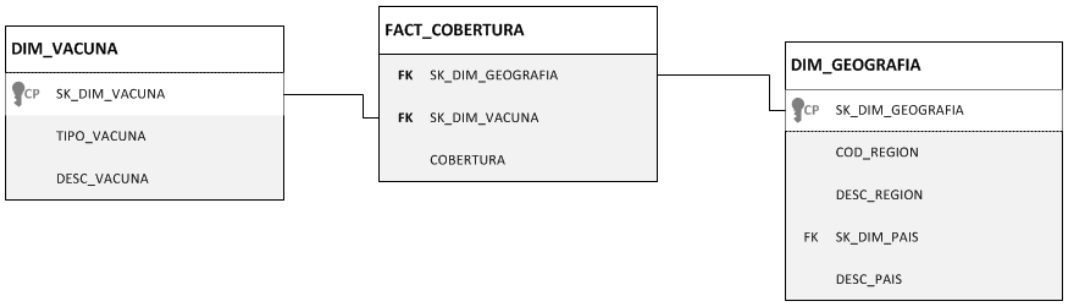
Tablas de hechos

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

- FACT_COBERTURA: es la tabla física que contendrá la información que permitirá realizar el análisis de la cobertura de inmunización desde diferentes perspectivas. Tendrá los siguientes campos:

Nombre campo	Tipo	Tamaño	Ejemplo
SK_DIM_ANIO	Numérico	4	2017
SK_DIM_GEOGRAFIA	Numérico	3	1
SK_DIM_VACUNA	Numérico	3	1
COBERTURA	Numérico	2	92
NUM_PV	Numérico	8	1023
OBJETIVO	Numérico	8	1033

Según lo definido en la guía de muestra, en la siguiente imagen se muestra el diseño del modelo físico para la tabla de hechos FACT_COBERTURA.



El estudiante, para realizar la entrega de la PRA1, tomando como modelo la guía de muestra, deberá completar el análisis y diseño que permita construir un almacén de datos para el análisis de la cobertura de inmunización capaz de cubrir las necesidades de los usuarios potenciales en el contexto del caso de estudio.

Formato y fecha de entrega

La entrega de esta actividad debe realizarse enviando un único mensaje al buzón Registro de AC del apartado Evaluación del aula. Adjunto en el mensaje se enviará un único archivo en formato Word o PDF con la solución de la práctica. El nombre del archivo debe ser la composición del nombre de usuario con el sufijo: "_BDA_PRA1.doc" (por ejemplo: si el nombre de usuario es "jmgarcia", entonces el archivo se debe llamar "jmgarcia_BDA_PRA1.doc ").

La fecha máxima de entrega es el 07/05/2020 a las 23.59 horas.