PEC 2

Presentación

La PEC2 consiste en una serie de preguntas con el objetivo de consolidar los conocimientos teóricos del módulo 4 - "Diseño multidimensional y explotación de datos" de la asignatura.

Objetivos

Teniendo en cuenta el contenido del módulo como objetivos más específicos cabe señalar:

- Comprender los componentes del modelo multidimensional (estructuras de datos, operaciones y restricciones de integridad).
- Diferenciar claramente entre el diseño conceptual, lógico y físico.
- Entender cuál es el diseño multidimensional y los problemas de diseño que presenta (en el ámbito conceptual, así como en el lógico y físico).
- Comprender la importancia de un adecuado diseño de un proyecto de almacén de datos, antes de su desarrollo y puesta en funcionamiento.
- Comprender los mecanismos de almacenamiento e indexación asociados a las herramientas multidimensionales (MOLAP, ROLAP, etc.).

Contenido.

La PEC2 consta de dos partes diferenciadas: una parte teórica y una parte práctica.

- La primera parte, la teórica, la compone por 4 preguntas que tienen por objetivo comprobar la correcta comprensión del módulo 4 por parte del estudiante.
- La segunda parte la constituye una única pregunta con varios apartados. Su objetivo es introducir al estudiante en la creación de procesos ETL- que deberá de llevar a cabo en la PRA2- utilizando la herramienta Spoon (PDI), instalada en el entorno VDI. Se recomienda consultar el documento **“Guía y consejos para el desarrollo de ETL”**.

Recursos

Módulo 4. Diseño multidimensional y explotación de datos
Fe de erratas. Material DBD Analíticas. (03/11/2020)
Guía Estudio del Módulo 4: Diseño Multidimensional y Explotación de Datos
Guía y consejos para el desarrollo de ETL
Caso práctico: Sistema integrado de Egresados universitarios (completo)
Enunciado Práctica (muestra)
Enunciado PRA2 (muestra)
Solución PRA2 (muestra)

Criterios de evaluación.

La nota final estará formada por:

Pregunta 1 (15%) + Pregunta 2 (15%) + Pregunta 3 (15%) + Pregunta 4 (15%) + Pregunta 5 (40%)

Formato y fecha de entrega

La entrega se realizará enviando un único mensaje al buzón de entrega de actividades del aula. Dicho mensaje llevará adjunto un único documento en formato word o pdf con la solución de la PEC2. En el documento se debe indicar obligatoriamente el nombre completo del estudiante y los estudios que está cursando. El nombre del archivo debe ser la composición del nombre de usuario y "_DW_PEC2" (por ejemplo: si el nombre de usuario es "bantich", el nombre del archivo debe ser "bantich_DW_PEC2.pdf" o "bantich_DW_PEC2.doc").

Es responsabilidad única del estudiante asegurarse que entrega el documento que pretende en el lugar que la Universidad habilita con este objetivo.

La fecha máxima de entrega es 28/04/2021 a las 23:59 h.

Pregunta 1 (15%):

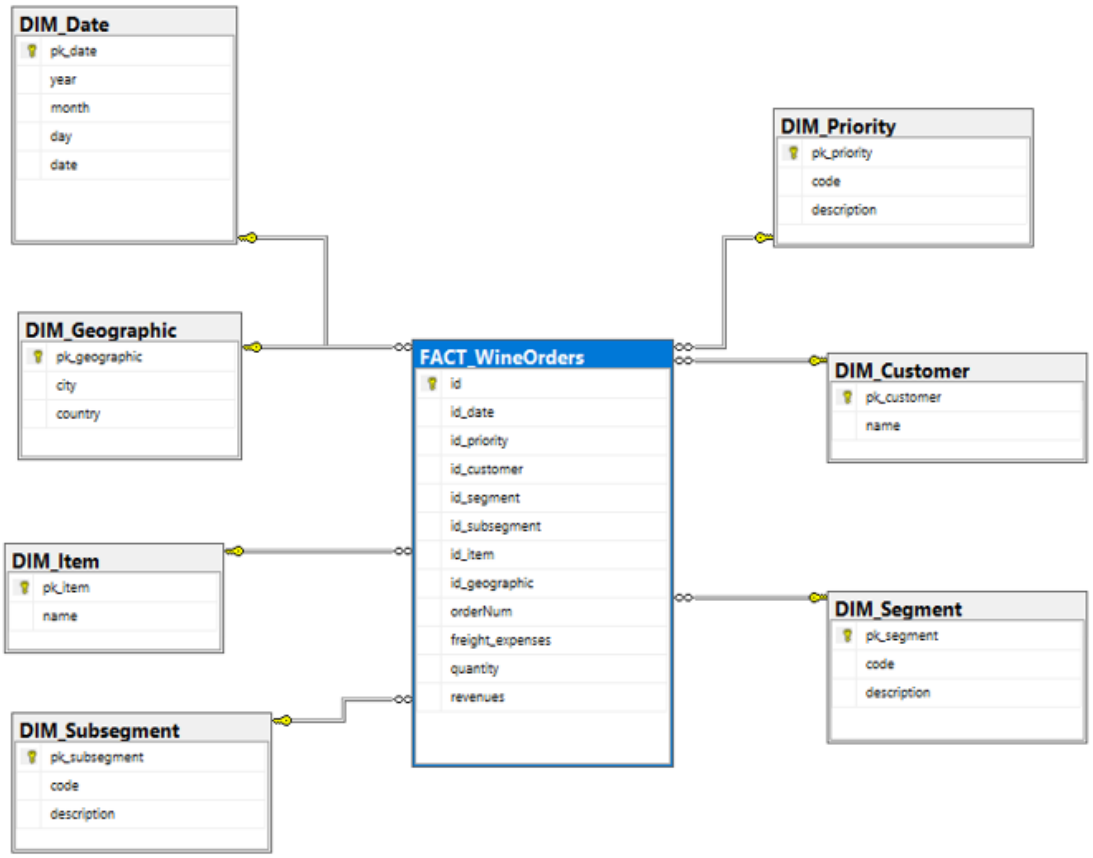
Disponemos de la tabla de hechos *FACT_Composition* que almacena información sobre las composiciones de alimentos, nutriente a nutriente, formando un esquema con forma de estrella con sus dimensiones.

Dibujad el diagrama del modelo lógico correspondiente a la tabla de hechos *FACT_COMPOSITION*, con los siguientes atributos descriptores de las dimensiones detalladas a continuación:

Dimensiones	Atributos descriptores
DIM_Country	codigo, nombre_pais, nombre_region
DIM_Date	año, mes
DIM_Food	codigo, familia, nombre, especie
DIM_Type	codigo, nombre
DIM_Processing	codigo, nombre
DIM_Component	codigo, nombre, unidad, grupo

Pregunta 2 (15%):

Disponemos del diseño físico siguiente:



Este modelo está basado en información sobre pedidos de venta de botellas de vino.

Nos indican el significado de los siguientes campos:

- **orderNum**: Número de pedido (numérico).
- **freight_expenses**: Gastos de envío (numérico).
- **quantity**: Cantidad solicitada en el pedido (numérico).
- **revenues**: Importe de ese pedido (numérico).

A partir de esta información, **indicad si las siguientes afirmaciones son correctas o no, justificando tus respuestas**.

a) El campo *[orderNum]* se podría definir como una dimensión degenerada.
b) El diagrama físico no es correcto porque en la tabla de hechos faltaría añadir también como PKs los campos que se relacionan con las dimensiones: *[id_date], [id_priority], [id_customer], [id_segment], [id_subsegment], [id_item] y [id_geographic]*.
c) Se dispone de 7 dimensiones (*DIM_Date, DIM_Priority, DIM_Customer, DIM_Segment, DIM_Subsegment, DIM_Item y DIM_Geographic*) y de 4 medidas (*orderNum, freight_expenses, quantity y revenues*).
d) El diagrama físico es correcto porque el campo *[id]* de la tabla de hechos corresponde con una clave subrogada y éstas siempre se deben definir en las tablas de hechos.
e) Ninguna de las anteriores es correcta.

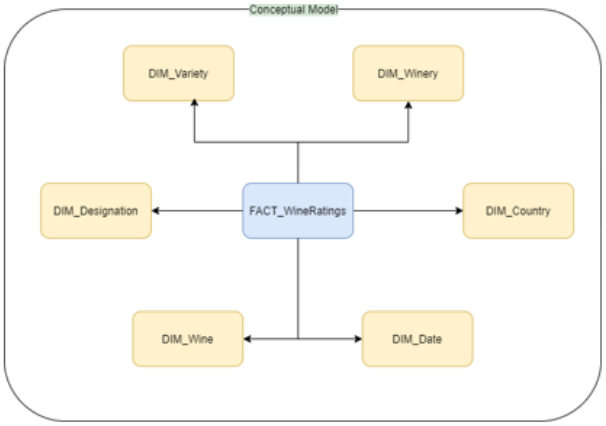
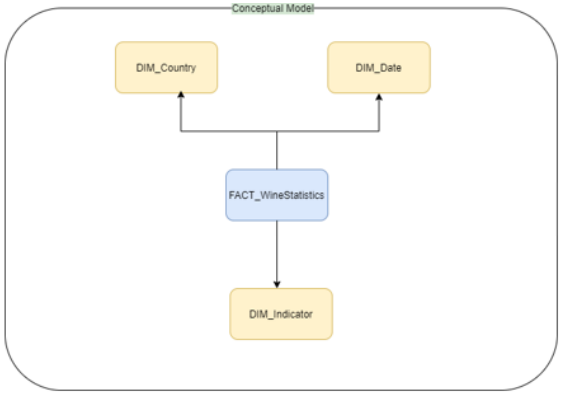
Pregunta 3 (15%):

Disponemos de dos modelos conceptuales diseñados a partir de un conjunto de fuentes con información correspondiente al sector vitivinícola.

Ambos modelos se refieren a un único *data mart*, dado que principalmente se basan en una única área temática. La información que se almacena el *data mart* está compuesta de las siguientes tablas de hechos:

- **FACT_WineRatings**: Datos de calificaciones y reseñas de vino.
- **FACT_WineStatistics**: Datos con diferentes estadísticas relacionadas con el sector vitivinícola.

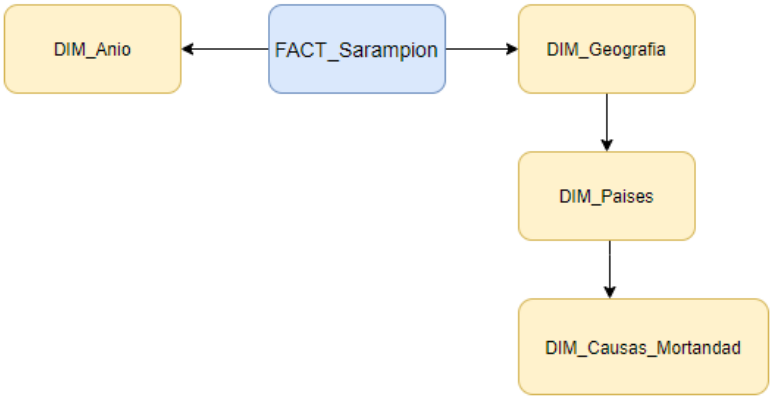
A partir de estos dos modelos:

FACT_WineRatings	FACT_WineStatistics
	

Justificad brevemente si puede existir alguna dimensión conformada en el modelo planteado.

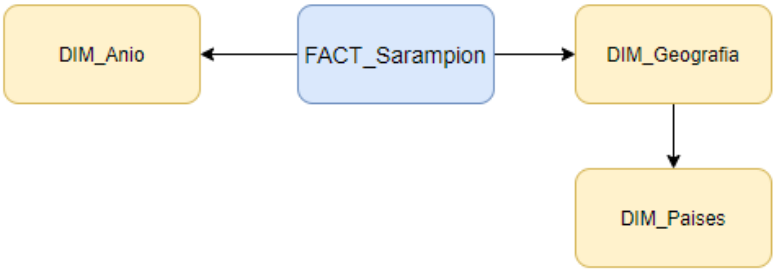
Pregunta 4 (15%):

Disponemos de la tabla de hechos FACT_SARAMPION, basada en información sobre la cobertura de inmunización. Su modelo de datos es el siguiente:



De acuerdo con este modelo, **indicad si las siguientes afirmaciones son correctas o no, justificando brevemente todas las respuestas**.

a) Si en el diagrama del modelo conceptual se desnormaliza la información de las causas principales de mortandad en la dimensión DIM_PAISES, el diagrama del diseño conceptual de la tabla de hechos FACT_SARAMPION sería:



b) El diseño conceptual presenta el mayor nivel de abstracción ya que es el más alejado a la representación física del modelo.

c) La representación gráfica de su correspondiente diseño físico es un diagrama en copo de nieve.

d) Todas las anteriores son correctas**.**

Pregunta 5 (40%):

A partir del fichero "ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx" hoja "Datos_tratados", se debe diseñar, implementar y ejecutar los procesos de extracción, transformación y carga para la Transformación IN_DENUNCIAS_INFRACCIONES, siguiendo y completando las siguientes cuestiones:

a) Completad y ejecutad el siguiente comando SQL para la creación de la tabla intermedia donde se almacenará los datos del origen "ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx" hoja "Datos_tratados".

Sustituid las XXXX, por el valor apropiado:

IN_DENUNCIAS_INFRACCIONES

```
XXXXX TABLE [dbo].[STG_Denuncias_Infracciones](
 [provincia] [varchar](100) NULL,
 [identificados_ertzaintza] [float] NULL,
 [XXXXXX_XXXXXXXX] [float] NULL,
 [XXXXXX_XXXXXXXX] [float] NULL,
 [XXXXXX_XXXXXXXX] [XXXX] NULL,
 [identificados_ppll] [XXXX] NULL,
 [XXXXXX_XXXX] [XXXX] NULL,
 [XXXXXX_XXXX] [XXXX] NULL,
 [vehic_intercept_ppll] [XXXX] NULL,
 [XXXXXXX] [datetime] NULL
    ) ON [PRIMARY]
    GO
```

b) Lectura de los ficheros xlsx. Completad la siguiente información del paso “File Input”, <marcada en azul>:

Nombre: < Nombre del paso >

Componente: Microsoft Excel input

Descripción: Permite cargar datos de entrada provenientes de un fichero Excel.

< Completad Descripción del paso especifico >

Parámetros:

Files / File or directory:< ruta del fichero >

Sheets (Hojas) mediante “get sheet names”, hoja “Datos_tratados”, con fila inicial “start row” = 5.

< Incluid captura de pantalla pestaña Sheets completada>

Fields (campos) Mediante el botón «Get fields from header row...» se obtienen todos los campos del fichero, así como el tipo, formato y longitud del dato.

< Incluid captura de pantalla pestaña Fields>

Preview: botón «*Preview rows*» (Previsualizar filas).

< Incluid captura de pantalla Preview rows >

c) Asegurad la homogeneidad de los datos mediante la normalización de los valores de los campos tipo «*String*». Convirtiendo a mayúsculas y eliminando los espacios en blanco al inicio y al final de cada cadena.

Completad la siguiente información del paso “*String Operation*”, <marcada en azul>:

Nombre: < Nombre del paso >

Componente: < Completad nombre componente spoon >

Descripción: < Completad Descripción del paso spoon >

Asegurad la homogeneidad de los datos mediante la normalización de los valores de los campos tipo «String». Convirtiendo a mayúsculas y eliminando los espacios en blanco al inicio y al final de cada cadena.

Parámetros:

<incluid captura de pantalla con todos los parámetros configurados>

d) Ordenación ascendente de todos los campos según su colocación en la tabla Staging.

Completad la siguiente información del paso “Row Order”, <marcada en azul>:

Nombre: < Nombre del paso >

Componente: < Completar nombre componente spoon >

Descripción: < Completar Descripción del paso spoon >

Ordenación ascendente de todos los campos según su colocación en la tabla Staging.

Parámetros:

<incluid captura de pantalla con todos los parámetros configurados>

e) Cargad la información transformada en la tabla de base de datos.

Completar la siguiente información del paso “Table Output”, <marcada en azul>:

Nombre: < Nombre del paso >

Componente: Table Output

Descripción: < Completad Descripción del paso spoon >

< Completad Descripción del paso especifico >

Parámetros:

Connection: < Completad >

Target table: < Completad >

Truncate table : < Completad >

f) Capturad la pantalla de la transformación completa, incluyendo la pestaña informativa de ejecución “*step metrics*”.

g) Realizad una Consulta en la Base de datos, que devuelva el número de registros de la tabla cargada. ¿Coincide con el número de registros procesados en cada paso, mostrados en “*step metrics*”?.

h) Realizad la consulta en la Base de Datos y capturad el resultado del Top 10 de registros sin ordenar, ¿coinciden con los 10 primeros registros ordenados ascendentemente de todos los campos según su colocación en la tabla *Staging*?.