PEC 2

Presentación

La PAC2 consiste en una serie de preguntas con el objetivo de consolidar los conocimientos teóricos de los módulos 3, 4 y 5 de la asignatura.

Objetivos

Teniendo en cuenta el contenido de los módulos como objetivos más específicos cabe señalar:

- Conocer todos los elementos que intervienen en la gestión del dato en la FIC.
- Comprender los componentes del modelo multidimensional.
- Diferenciar claramente entre el diseño conceptual, lógico y físico.
- Entender las etapas del ciclo de vida del almacén de datos.
- Comprender la importancia de un adecuado diseño de un proyecto de almacén de datos, antes de su desarrollo y puesta en funcionamiento.

Contenido.

Esta parte de la PEC está compuesta por 8 preguntas que tienen por objetivo comprobar la correcta comprensión de los módulos 3, 4 y 5 de la asignatura a partir de las respuestas del estudiante.

Recursos

Módulo 3: Los Datos en la FIC
Módulo 4: Diseño Multidimensional y Explotación de Datos
Módulo 5: Administración de Sistemas de Data Warehouse

Soluciones oficiales de PRA1 y PRA2 del caso de uso “Almacén de datos para el análisis de la cobertura de inmunización”.

Criterios de evaluación.

Todas las respuestas se deben justificar. No demostrar en la justificación la correcta comprensión teórica implica un 0% en dicha pregunta. Se valorarán positivamente todas las justificaciones/argumentaciones que no hagan referencia directa (o copia) a los recursos proporcionados por la Universidad. Las respuestas deben demostrar la asimilación de los contenidos por parte del estudiante.

La nota final estará formada por:

Pregunta 1 (12,5%) + Pregunta 2 (12,5%) + Pregunta 3 (7,5%) + Pregunta 4 (15%) + Pregunta 5 (15%) + Pregunta 6 (7,5%) + Pregunta 7 (15%) + Pregunta 8 (15%)

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 PEC. 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 18/06/2020 a las 23:59 h.

Pregunta 1 (12,5%):

En el proceso de definición del modelo de datos de la PRAC 1 nos hemos encontrado con dos ficheros, Coverage y Wuenic, que contenían información muy similar respecto a la cobertura de inmunización, pero de distintas fuentes de datos. Hemos podido ver que la información facilitada por la OMS y UNICEF respectivamente no siempre coincide, por lo que deberemos tomar una decisión a la hora de analizar los datos. Por ejemplo:

- Suponemos que una fuente (OMS) es más importante o fiable que la otra, por lo tanto, descartaremos la otra fuente de datos.
- Suponemos que ambas fuentes tienen una fiabilidad similar, por lo que calcularemos un nuevo dato de cobertura a partir de ambas
- Queremos demostrar que el sistema sanitario funciona perfectamente, y por lo tanto nos quedamos con el dato más positivo en cada caso. O por el contrario, consideramos necesario forzar un incremento en los esfuerzos del sistema sanitario, y por lo tanto nos quedamos en cada caso con el valor más pesimista
- Nos quedamos con ambas columnas.

Podemos aplicar los mismos ejemplos en base a la situación actual producida por el COVID-19, donde muchos países han establecido diferentes métodos de recuento de infectados y de víctimas, compitiendo en algunos casos por demostrar que país está gestionando mejor la situación.

Razonad, a nivel ético, las implicaciones que pueden tener la manipulación o preselección de datos utilizados para la toma de decisiones o presentación de conclusiones, así como el uso de fuentes de datos no contrastadas, y justificad cual creéis que sería la mejor solución ante situaciones como pueden ser las presentadas anteriormente (no tiene por qué ser ninguna de las planteadas como ejemplo)

Pregunta 2 (12,5%):

Con la aparición del IoT (Internet de las cosas) todo dispositivo electrónico puede convertirse en un emisor de datos, ya sea directamente desde redes wifi o con soluciones 3G. Esto implica que toda empresa, independientemente de su línea de negocio y de su tamaño, puede disponer de un entorno Big Data con el que trabajar, y como científicos de datos muchas veces tendréis que definir o bien aconsejar que hacer con la inmensa cantidad de datos que pueden obtenerse.

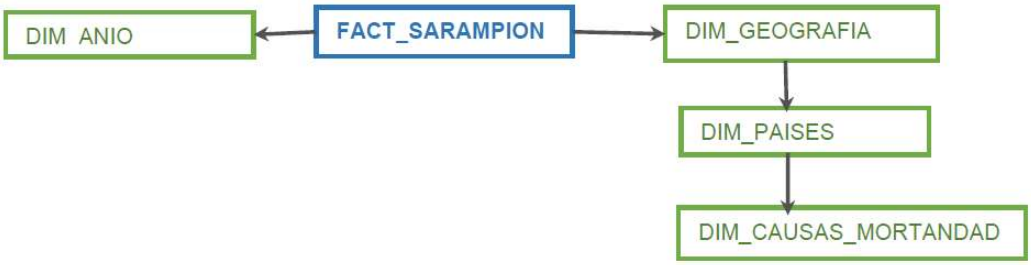
a) Presentad un ejemplo de dispositivos que pueden generar información a través del concepto IoT

b) Plantead las necesidades que pueden detectarse a la hora de gestionar las etapas principales de la calidad de los datos.

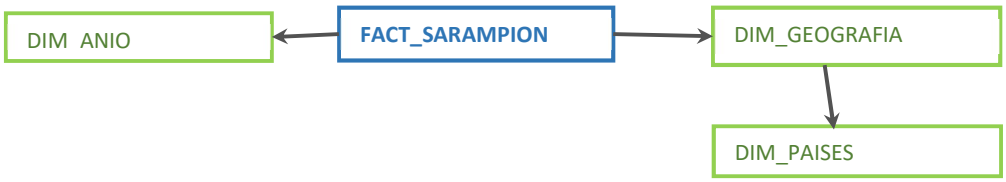
Pregunta 3 (7,5%):

Seleccione la opción correcta en relación al diseño conceptual del almacén de cobertura de inmunización:

a) En la solución oficial del caso práctico, el diagrama del diseño conceptual de la tabla de hechos FACT_SARAMPION es:



Pero si se normaliza la información de las causas principales de mortandad en la dimensión DIM_PAISES, el diagrama del diseño conceptual sería:



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

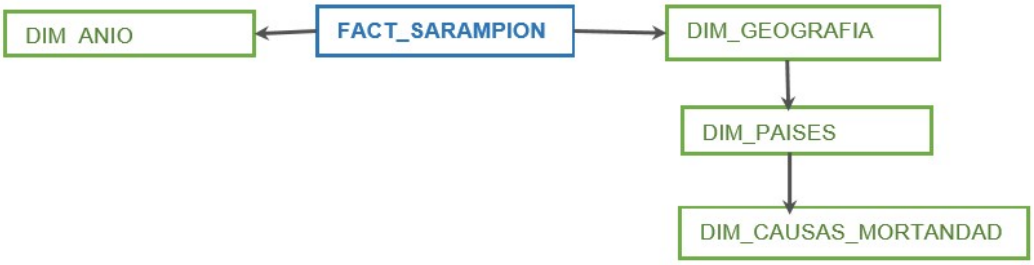
c) La representación gráfica del diseño conceptual es un diagrama de estrella o copo de nieve.

d) Todas las anteriores son correctas.

e) Ninguna de las anteriores son correctas.

Pregunta 4 (15%):

Justifica brevemente si la representación gráfica del diseño conceptual de la FACT_SARAMPION es un diagrama de estrella o de copo de nieve.



Pregunta 5 (15%):

Dibuja el diagrama del modelo lógico correspondiente a la siguiente tabla, con los atributos descriptores de la tabla de hechos FACT_SARAMPION:

Dimensiones	**Atributos descriptores**
DIM_ANIO	COD_ANIO
DIM_GEOGRAFIA	COD_GEOGRAFIA
DIM_PAISES	COD_PAIS
DIM_CAUSAS_MORTANDAD	COD_CAUSA

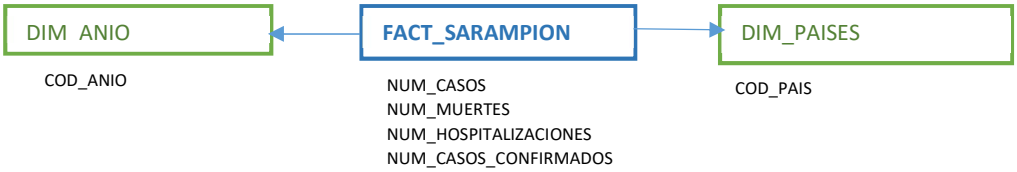
Pregunta 6 (7,5%):

Seleccione la afirmación correcta en relación al diseño del modelo del almacén de cobertura de inmunización. Justifica aquellas afirmaciones que consideres que no son correctas.

a) La dimensión DIM_VACUNA es una dimensión conformada en el diseño del modelo.

b) En la solución oficial del caso práctico, el diagrama del diseño lógico de la tabla de hechos FACT_SARAMPION del modelo del almacén de cobertura de inmunización es:

Pero si se normaliza la información de las causas principales de mortandad en la dimensión DIM_PAISES, el diagrama del diseño lógico sería:



c) En la solución oficial del caso práctico, la tabla física de FACT_COBERTURA es:

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

Por lo tanto, si se crea en la tabla una clave primaria (*primary key*) sería compuesta, y estaría formada por tres campos (sk_dim_anio, sk_dim_geografia y sk_dim_vacuna), que a su vez serán claves foráneas (*foreign key*) a las dimensiones conformadas del modelo.

d) Todas las anteriores son correctas.

e) Ninguna de las anteriores son correctas.

Pregunta 7 (15%)

Cuando se diseña e implementa un Data Warehouse se debe decidir, inevitablemente, su arquitectura tecnológica, cada una de ellas puede tener una mayor o menor adaptación al proyecto a desarrollar. Debemos conocer sus características y diferencias para tomar la decisión más adecuada. Dos de estas Arquitecturas, y a su vez las más implementadas y aceptadas, son la Enterprise Bus Architecture (modelo Kimball) y la Corporate Information Factory (modelo Inmon 1.0).

Haz un análisis pormenorizado de ambas arquitecturas: descríbelas, compáralas y explica brevemente tu opinión sobre las fortalezas y debilidades de cada una de ellas. (A parte de los apuntes de la asignatura puedes encontrar muchísima información en la web).

Pregunta 8 (15%)

Otra decisión de Infraestructura que se deberá tomar es el software utilizado para el desarrollo e implementación del Data Warehouse. En la práctica realizada en este semestre se han utilizado:

- Herramienta de ETL: Pentaho Data Integration.
- Aplicaciones BI: Herramientas Microsoft, Analysis Services (con Visual Studio) y opcionalmente Power BI.

Busca una alternativa a cada uno de estos dos conjuntos de software, compara las características técnicas, analiza y detalla si cumplen los requerimientos expuestos en el módulo 5.