Sistemas de Información para el Análisis de Grandes Volúmenes de Datos

PRUEBA FINAL

Ejercicio 1 (30 puntos)

La administración de una carrera universitaria desea realizar un análisis multidimensional acerca de las actividades de los estudiantes. Para este análisis se tiene como fuente de datos a las siguientes tablas de una base de datos relacional:

ASIGNATURAS (id-asignatura, nom-asignatura, id-materia, créditos)

MATERIAS (id-materia, área-tematica)

ESTUDIANTES (ci-estudiante, nom-estudiante, fecha-nac, sexo, fecha-ing, telefono, dirección, ciudad, depto)

DICTADO-CURSOS (id-asignatura, fecha, cant-estudiantes, docente-resp)

ACTIVIDAD-CURSO (ci-estudiante, id-asignatura, fecha, nota)

ACTIVIDAD-EXAMEN (ci-estudiante, id-asignatura, fecha, nota)

Los requerimientos de análisis son los siguientes:

1) Se desea analizar el dictado de cursos a través del tiempo (contemplando semestre y año), agrupando las asignaturas según su peso en créditos (alto, medio, bajo) y según la materia y área temática a las que corresponden. Sobre estos dictados interesa analizar la cantidad de estudiantes distintos que hubo en cada caso y la cantidad de asignaturas distintas que se dictaron. En este análisis no se quiere individualizar a cada estudiante, por motivos de privacidad con respecto a los analistas que utilizarán el cubo.

2) Se desea analizar las actividades de los estudiantes a través del tiempo. En este caso sí pueden visualizarse los estudiantes individuales. Interesa agrupar a los estudiantes según su rango etario, su sexo, la ciudad y departamento de donde provienen, y la antigüedad en la carrera (dividida en rangos). Sobre las asignaturas interesa la misma información que en el requerimiento anterior. Los indicadores que se quiere analizar en este caso son: la cantidad de cursos realizados (de la misma o distintas asignaturas), la cantidad de exámenes realizados (de la misma o distintas asignaturas), y las notas obtenidas.

Se pide:

Realizar el modelo conceptual en CMDM para estos requerimientos. Incluir para cada medida la indicación de si tiene o no problemas de aditividad, y en el caso de tener problemas, si es posible aplicar una operación de roll-up que lo resuelva, qué operación aplicaría.

Ejercicio 2 (25 puntos)

Dado el diseño conceptual realizado en el ejercicio anterior, contestar las siguientes preguntas relativas al diseño lógico relacional del DW correspondiente:

1) ¿Identifica alguna medida de las que incluyó en el modelo conceptual que no sea necesario incluir en la tabla de hechos? ¿Cuál/es?
2) Elija una de las dimensiones propuestas en el modelo conceptual y proponga el esquema relacional correspondiente a esa dimensión.
3) Suponga que la consulta de la **cantidad de estudiantes participando en cursos por año** se realiza con muy alta frecuencia y se desea tener un tiempo de respuesta lo más corto posible en esa consulta. ¿Qué se podría hacer en el diseño relacional para lograr una mayor eficiencia en esa consulta?

Ejercicio 3 (20 puntos)

Nombrar 3 tipos de herramientas de front-end y describir cuál es el objetivo principal de cada una, marcando sus diferencias.

Ejercicio 4 (25 puntos)

Para cada una de las siguientes preguntas elija la opción que le parezca mejor, agregando una breve justificación.

Pregunta 1)

¿En cuál arquitectura de las siguientes se le da baja prioridad a la consistencia, calidad y confiabilidad de los datos, frente a otras cualidades?

a) Data Warehouse
b) Data Lake
c) Data Lakehouse

Pregunta 2)

¿Qué significa el “data swamp” como riesgo al utilizar Data Lakes?

a) Es el problema de tener baja performance en las consultas a los datos.
b) Es el problema de no conocer la estructura de los datos por ser no-estructurados, y que sea difícil consultarlos.
c) Es el problema de que los datos queden olvidados en el data lake y se acumulen, dejando de ser utilizados porque no se conocen.

Pregunta 3)

¿En un DW, qué problema se genera cuando cambian los datos de una dimensión, tal que un elemento deja de agrupar en cierto valor para pasar a agrupar en otro, en el siguiente nivel de la jerarquía?

a) Se generan valores nulos.
b) Se modifica el histórico de los hechos.
c) Se generan problemas de aditividad por doble conteo.

Pregunta 4)

En un mismo proceso de ETL implementado en Pentaho Data Integration, ¿es posible cargar datos en un almacenamiento y luego utilizar éstos como entrada de otra transformación?

a) Sí, es posible.
b) No, no es posible, una vez que se carga en un almacenamiento destino, no se pueden utilizar más en ese proceso.
c) Es posible, pero solamente si el almacenamiento es un archivo (no una base de datos).