Universidad Nacional Autónoma de México

Facultad de Ciencias, 2024 - 1

Almacenes y Minería de Datos

Tarea 03. Procesamiento OLAP.

PostgresandoesoSQLazos

Integrantes

Adrian Aguilera Moreno	421005200
Marco Antonio Rivera Silva	318183583
Sebastían Alejandro Gutiérrez Medina	318287021
Israel Hernández Dorantes	318206604
Alejandra Ortega García	420002495

1. Tarea 03.

1. ¿Qué es un lago de datos (data lake)? Indica en una tabla, las diferencias entre un lago de datos, un mercado de datos y un almacén de datos.

   Un *Data Lake* es un modelo de almacenamiento para grandes cantidades de datos en su forma bruta y sin procesar, permite almacenar datos de cualquier tipo, ya sean estructurados, semi-estructurados o no estructurados, sin imponer una estructura específica. Proporciona un repositorio centralizado para almacenar datos de diversas fuentes, como registros de aplicaciones, datos de sensores, registros de eventos, archivos de registro, datos de redes sociales y mucho más.

Data Lake	**Data Mart**	**Data Ware House**
Almacena datos en su forma bruta y sin procesar, lo que significa que puede contener datos estructurados, semiestructurados y no estructurados.	Contiene datos específicos y preseleccionados que están altamente estructurados y organizados para atender las necesidades de un grupo de usuarios particulares.	Almacena datos históricos y actuales en una estructura altamente organizada y optimizada para el análisis empresarial.
Ofrece una alta flexibilidad en términos de tipos de datos y cambios en la estructura de datos, lo que lo hace adecuado para exploración de datos.	Tiene una estructura más rígida y está diseñado para satisfacer las necesidades específicas de un grupo de usuarios, lo que lo hace menos flexible.	También tiene una estructura rígida y se optimiza para consultas predefinidas, lo que limita su flexibilidad.
Puede manejar grandes volúmenes de datos sin problemas y se puede escalar horizontalmente.	Por lo general, se configura para satisfacer las necesidades de un grupo de usuarios específico y, por lo tanto, no es tan escalable.	Es escalable, pero la escalabilidad puede ser costosa debido a su estructura rígida y optimización.
Es utilizado por científicos de datos, analistas y otros profesionales para explorar datos en bruto y desarrollar modelos de análisis.	Se crea para un grupo de usuarios específico, como un departamento o equipo, y se enfoca en sus necesidades de informes y análisis particulares.	Se utiliza para admitir análisis empresariales a nivel organizacional y proporciona datos consolidados y consistentes para toda la empresa.
Por lo general, se realiza un procesamiento posterior a los datos para estructurarlos y hacerlos listos para el análisis.	Los datos se extraen y transforman específicamente para satisfacer las necesidades de un grupo de usuarios.	Los datos se someten a un proceso ETL (extracción, transformación y carga) antes de almacenarse, lo que garantiza que estén listos para consultas y análisis empresariales.

2. Describe al menos 6 características de los almacenes de datos.

- **Centralización de datos**: Un almacén de datos centraliza datos de diversas fuentes de la organización, lo que facilita el acceso a información consolidada y coherente.
- **Estructura organizada**: Los datos se almacenan en una estructura organizada que facilita la consulta y el análisis. Por lo general, se utilizan esquemas en estrella o en copo de nieve para organizar los datos.
- **Integración de datos**: Los datos se extraen, transforman y cargan (ETL) en el almacén de datos, lo que implica que se integran y se estructuran para su uso en análisis y generación de informes.
- **Datos históricos**: Un almacén de datos almacena datos históricos, lo que permite el análisis de tendencias y patrones a lo largo del tiempo. A menudo mantienen un registro de cambios para rastrear quién ha modificado los datos y cuándo.
- **Escalabilidad**: Los almacenes de datos son escalables para manejar volúmenes crecientes de datos a medida que la organización crece.
- **Arquitectura dimensional**: Muchos almacenes de datos siguen una arquitectura dimensional, que organiza los datos en dimensiones (como tiempo, producto, ubicación) y hechos (medidas o métricas).

3. Considera que tienes el siguiente esquema de una base de datos relacional, para un aeropuerto internacional:

Vuelo	(*IdVuelo*, *Compania*, *AeropuetoSalida*, *AeropuertoLlegada*, *HoraSalida*, *HoraLlegada*, *FechaSalida*, *FechaLlegada*)
Aeropuerto	(*IdAeropuerto*, *NombreAeropuerto*, *Ciudad*, *País*)
Boleto	(*NumeroBoleto*, *IdVuelo*, *FechaVuelo*, *Asiento*, *Tarifa*, *Nombre*, *Apellidos*, *Sexo*)
Check – in	(*NumeroBoleto*, *HoraRegistro*, *NumeroMaletas*)

Diseña un almacén de datos, siguiendo la metodología de 4 pasos de Kimball: elije el proceso a modelar, define la granularidad, define las dimensiones (con sus respectivas jerarquías) y define los hechos y medidas.

De acuerdo a los pasos de Kimball tenemos lo siguiente:

1) **Selección del proceso a modelar**:

El proceso a modelar será el de viajar en avión, dónde requerimos un boleto para abordar, un número de vuelo que pertenece a una compañía encargada de los vuelos, y una fecha y hora para el registro y finalmente abordar. Todos estos datos están representados en el anterior esquema relacional y requerimos diseñar nuestro almacén con estos datos.

2) **Granularidad del proceso seleccionado**:

El nivel de granularidad que nos conviene tener es por fecha de día y hora, pues requerimos que nuestros pasajeros lleguen en cierto horario, nuestro horario de salida y de registro.

3) **Definición de las dimensiones**:

   Tendremos una dimensión con respecto a la fecha que indicará Año, Mes, Día, Hora. Otro nivel de detalle será el País de ubicación, ciudad. Por último queremos una dimensión para detallar el nombre del aeropuerto, el avión a abordar, y el asiento.

4) **Definición de los hechos y medidas**:

   Nuestra tabla de hechos serán `pasajeros` y tendremos los datos de `pasajeros-hora`, `Pasajeros-por-vuelo`, `pasajeros-checking`, así podemos encontrar la cantidad de pasajeros en un avión de un aeropuerto por hora, y otras distintas consultas.

4. Describe las características principales de los esquemas: estrella, copo de nieve y constelación. Proporciona un ejemplo de cada uno de ellos.

	Estrella	**Copo de nieve**	**Constelación**
Características	Un hecho central (tabla de hechos) rodeado de múltiples dimensiones (tablas de dimensiones). Las dimensiones contienen atributos que describen los detalles de los datos del hecho. Los datos son altamente desnormalizados para facilitar las consultas de análisis.	Similar al esquema en estrella, pero las tablas de dimensiones se descomponen aún más en subdimensiones. Se utiliza para ahorrar espacio y mejorar la eficiencia en sistemas donde las dimensiones tienen una alta cardinalidad.	Múltiples esquemas en estrella conectados entre sí. Útil cuando una organización necesita mantener múltiples almacenes de datos separados o cuando hay datos compartidos entre diferentes áreas funcionales.
Ejemplos	En una tienda podríamos tener una tabla de hechos que haga referencias a las ganancias por mes o trimestre, mientras que en las tablas de dimensiones podríamos tener la granularidad por tienda o estado.	Siguiendo con el ejemplo anterior podemos dimensionar los estados en municipios y tiendas en departamentos para tener una mayor granularidad, lo que nos permite tener otras conexiones a nuestra estrella y no necesariamente en el nodo central. También podemos involucrar más tablas de hechos.	Supongamos que nuestro modelo anterior funciona en México, pero hay modelos similares en otros países, un modelo que contenga estos modelos no conexos entre sí es un ejemplo de constelación.

5. ¿Cuál es el objetivo de la Arquitectura de bus? Indica las ventajas y desventajas.

   - **Objetivo.** El objetivo principal de la Arquitectura de Bus es proporcionar un enfoque centralizado y eficiente para la interoperabilidad de sistemas, lo que incluye la comunicación, el enrutamiento, la transformación de datos y la gestión de servicios.

- **Ventajas.**
  - Permite agregar, modificar o eliminar servicios sin afectar en gran medida otros componentes del sistema, lo que facilita la adaptación a cambios en los requisitos empresariales.
  - Proporciona un punto central de control para la administración y el monitoreo de servicios, lo que facilita la supervisión y el mantenimiento.
  - Los sistemas no necesitan conocer los detalles de implementación de otros sistemas, ya que la comunicación se realiza a través del bus, lo que mejora la separación de preocupaciones.
  - Permite que sistemas heterogéneos se comuniquen y compartan datos de manera efectiva, lo que facilita la integración de aplicaciones y sistemas dispares.
- **Desventajas.**
  - La configuración de una Arquitectura de BUS puede ser compleja y requerir una inversión de tiempo y recursos significativa al principio.
  - Si no se diseña y ajusta adecuadamente, un BUS puede introducir latencia en la comunicación entre sistemas.
  - La administración y el mantenimiento continuo de un BUS pueden requerir habilidades especializadas y recursos dedicados.

6. Explica las diferencias entre las implementaciones ROLAP y MOLAP, ¿cuál resulta mejor? Investiga a qué se refiere una implementación HOLAP.

ROLAP	MOLAP
En ROLAP, los datos se almacenan en bases de datos relacionales. Los datos analíticos se organizan en tablas relacionales utilizando lenguaje SQL.	En MOLAP, los datos se almacenan en una estructura multidimensional, generalmente en cubos OLAP. Los datos se organizan en dimensiones y medidas, lo que facilita un acceso rápido y eficiente a los datos analíticos.
Tiene un rendimiento más lento en comparación con MOLAP, ya que las consultas deben ser traducidas en SQL y ejecutadas en una base de datos relacional.	Ofrece un rendimiento más rápido, ya que los datos se precalculan y se almacenan en una estructura multidimensional, lo que permite respuestas más rápidas a las consultas analíticas.
Requiere más espacio de almacenamiento debido a la estructura de base de datos relacional y la redundancia de datos en tablas.	Utiliza menos espacio de almacenamiento, ya que los datos se almacenan de manera más eficiente en cubos multidimensionales sin redundancia significativa.
Ofrece una mayor flexibilidad en términos de manejo de datos complejos y relaciones entre ellos, lo que es adecuado para escenarios donde los datos son altamente normalizados.	Es más adecuado para datos que se pueden modelar de manera efectiva en una estructura multidimensional. Puede no ser tan flexible para relaciones de datos complejas.
Puede requerir un mayor esfuerzo de mantenimiento debido a la complejidad de las bases de datos relacionales subyacentes.	Tiende a requerir menos mantenimiento, ya que los datos se precalculan y se almacenan en una estructura que está diseñada para consultas analíticas.
Es más escalable en términos de capacidad para manejar grandes volúmenes de datos y cargas de trabajo complejas.	Puede ser menos escalable en comparación con ROLAP para casos de uso con requisitos extremadamente altos de rendimiento y escalabilidad.

El uso de estos modelos depende en gran medida de su uso, por ejemplo:

- Si los datos cambian con frecuencia, requerimos escalabilidad, entonces lo más indicado es usar el modelo ROLAP.
- Si en otro caso tenemos grandes cantidades de datos y requerimos consultas eficientes sin estar agrgando, el modelo indicado sería el MOLAP.

HOLAP (Hybrid Online Analytical Processing). Es una combinación de los modelos MOLAP y ROLAP, se basa en la idea de que no es necesario elegir uno de estos modelos puros, sino que se pueden aprovechar las ventajas de ambos enfoques. A continuación listamos algunos conceptos relacionados:

- Almacenamiento híbrido.
- Capacidad de elección.
- Transparencia para los usuarios.
- Optimización de consultas.

7. Considera un DWH con las dimensiones Cliente, Producto y Fecha y la medida ventas, la cual significa lo que se vendió de un cierto producto en una cierta fecha para un cierto cliente. Dibuja un esquema estrella asumiendo alguna jerarquía de conceptos para cada dimensión. Cuáles serían las consultas OLAP que se necesitan para obtener:

   - El total de ventas realizadas durante 2014.
   - El total de ventas para los Accesorios y Productos de Montaña en el primer y cuarto trimestre de 2011 en Francia y Alemania.
   - Las ventas que se tuvieron para los artículos de Turismo en las ciudades de Francia y Estados Unidos (asume que se tiene esa información y especifica en cada caso valores que decidas para al menos tres ciudades en cada caso).
   - Las ventas de todos los productos en Australia durante los meses del primer trimestre del 2011 (enero, febrero y marzo).
   - Las ventas que se tuvieron por cada producto que se tiene en Accesorios para las ciudades de Estados Unidos en cada mes de 2011.

   El esquema estrella para el DWH propuesto es el siguiente:

```
Cliente (IdCliente, Nombre, País)
Producto (IdProducto, Categoría, Subcategoría, Nombre)
Fecha (Fecha, Mes, Trimestre, Año)
Ventas (IdCliente, IdProducto, Fecha, Ventas)
```

   Las **jerarquías** de conceptos para cada dimensión son las siguientes:

   - **Dimension Cliente**:
     - País: España, Francia, Alemania, Estados Unidos, Australia
     - Ciudad: Madrid, Barcelona, París, Berlín, Nueva York, Los Ángeles, Sydney
   - **Dimension Producto**:
     - Categoría: Deportes, Moda, Electrónica
     - Subcategoría: Montaña, Turismo, Accesorios
     - Nombre: Casco, Ropa, Mochila, Cámara, Zapatos, Billete de avión, Hotel, Guía de viajes
   - **Dimension Fecha:**
     - Mes: Enero, Febrero, Marzo, Abril, Mayo, Junio, Julio, Agosto, Septiembre, Octubre, Noviembre, Diciembre
     - Trimestre: 1, 2, 3, 4
     - Año: 2011, 2012, 2013, 2014, ...

   Las **consultas OLAP** necesarias para obtener las respuestas a las preguntas planteadas son las siguientes:

     - El total de ventas realizadas durante 2014.

```sql
SELECT SUM(Ventas) AS VentasTotales
FROM Ventas
WHERE Fecha BETWEEN '2014-01-01' AND '2014-12-31';
```

- El total de ventas para los Accesorios y Productos de Montaña en el primer y cuarto trimestre de 2011 en Francia y Alemania.

```sql
SELECT SUM(Ventas) AS VentasTotales
FROM Ventas
WHERE Fecha BETWEEN '2011-01-01' AND '2011-12-31'
AND (Producto.Categoría = 'Accesorios' OR Producto.Categoría = 'Montaña')
AND (País = 'Francia' OR País = 'Alemania')
AND (Trimestre = 1 OR Trimestre = 4);
```

- Las ventas que se tuvieron para los artículos de Turismo en las ciudades de Francia y Estados Unidos (asume que se tiene esa información y especifica en cada caso valores que decidas para al menos tres ciudades en cada caso).

```sql
SELECT SUM(Ventas) AS VentasTotales
FROM Ventas
WHERE Producto.Categoría = 'Turismo'
AND (País = 'Francia' OR País = 'Estados Unidos')
```

- Las ventas de todos los productos en Australia durante los meses del primer trimestre del 2011 (enero, febrero y marzo).

```sql
SELECT SUM(Ventas) AS VentasTotales
FROM Ventas
WHERE País = 'Australia'
AND Fecha BETWEEN '2011-01-01' AND '2011-03-31';
```

- Las ventas que se tuvieron por cada producto que se tiene en Accesorios para las ciudades de Estados Unidos en cada mes de 2011.

```sql
SELECT Producto.Nombre, SUM(Ventas) AS VentasTotales
FROM Ventas
JOIN Producto ON Ventas.IdProducto = Producto.IdProducto
WHERE Producto.Categoría = 'Accesorios'
AND País = 'Estados Unidos'
GROUP BY Producto.Nombre, Fecha;
```

8. Explica los enfoques top-down y bottom-up para el desarrollo de un Almacén de Datos.

- **Top-Down**. En el enfoque top-down, el proceso de desarrollo del Almacén de Datos comienza con una visión estratégica y una comprensión completa de las necesidades comerciales y los objetivos de la organización.

  Se crea un modelo de datos multidimensional, que incluye dimensiones, jerarquías, métricas y relaciones, de acuerdo con los requisitos de análisis empresarial.

- **Buttom-Up**. En el enfoque bottom-up, el proceso comienza identificando las fuentes de datos existentes en la organización, como bases de datos operativas y sistemas de aplicaciones.

  Se desarrollan pequeños almacenes de datos departamentales o específicos para ciertas aplicaciones o necesidades. Estos almacenes se crean de manera más ágil y rápida.

Cómo vimos en clase, el enfoque Buttom-Up es más viable porque siempre puede unificar más datos y más departamentos de la empresa u organización.

9. Investiga cómo funcionan las operaciones OLAP Drill-across y Drill-through.

   Las operaciones OLAP de "Drill-across" y "Drill-through" son técnicas de análisis de datos utilizadas en entornos de Almacenes de Datos (Data Warehouses) y sistemas OLAP (Procesamiento Analítico en Línea) para explorar y analizar datos multidimensionales.

   - **Drill-across**. La operación de "Drill-across" permite a los usuarios acceder a datos detallados desde diferentes cubos OLAP que comparten una dimensión común. La idea es que, cuando se navega por un cubo OLAP, se puede "atravesar" (drill-across) a otro cubo que tiene una dimensión en común para obtener información adicional. Esto es particularmente útil cuando se necesita combinar datos de diferentes áreas funcionales o departamentos de una organización que se almacenan en diferentes cubos.
   - **Drill-through**. La operación de "Drill-through" permite a los usuarios acceder a datos detallados a nivel de registros subyacentes en la fuente de datos original. A diferencia del "Drill-across," que se refiere a diferentes cubos OLAP, el "Drill-through" se relaciona con los datos de origen, como bases de datos relacionales, y permite que los usuarios accedan a información más detallada y a menudo transaccional.

10. Un DWH es orientado a un tema, cuáles podrían ser los aspectos críticos en las siguientes organizaciones: una compañía manufacturera internacional, el banco de una comunidad local y una cadena hotelera nacional.

    - **Aspectos críticos en una compañía manufacturera internacional.**
      En una compañía manufacturera internacional, los aspectos críticos de un DWH pueden incluir:
      - **Datos de producción:** El DWH debe proporcionar datos de producción, como datos de inventario, datos de ventas y datos de calidad. Estos datos pueden utilizarse para realizar análisis de producción, como la optimización de la cadena de suministro y la mejora de la calidad de los productos.
      - **Datos de marketing:** El DWH debe proporcionar datos de marketing, como datos de clientes, datos de campañas y datos de ventas. Estos datos pueden utilizarse para realizar análisis de marketing, como la segmentación de clientes y la optimización de las campañas de marketing.

  - **Datos financieros:** El DWH debe proporcionar datos financieros, como datos de ingresos, datos de gastos y datos de rentabilidad. Estos datos pueden utilizarse para realizar análisis financieros, como la planificación y el presupuesto.

- **Aspectos críticos en un banco de una comunidad local.**

  En un banco de una comunidad local, los aspectos críticos de un DWH pueden incluir:

  - **Datos de clientes:** El DWH debe proporcionar datos de clientes, como datos de cuentas, datos de transacciones y datos de comportamiento. Estos datos pueden utilizarse para realizar análisis de clientes, como la detección de fraudes y la mejora de la satisfacción del cliente.
  - **Datos financieros:** El DWH debe proporcionar datos financieros, como datos de préstamos, datos de depósitos y datos de inversiones. Estos datos pueden utilizarse para realizar análisis financieros, como la gestión del riesgo y la planificación de la liquidez.
  - **Datos de operaciones:** El DWH debe proporcionar datos de operaciones, como datos de empleados, datos de procesos y datos de sistemas. Estos datos pueden utilizarse para realizar análisis de operaciones, como la mejora de la eficiencia y la reducción de costes.

- **Aspectos críticos en una cadena hotelera nacional.**

  En una cadena hotelera nacional, los aspectos críticos de un DWH pueden incluir:

  - **Datos de reservas:** El DWH debe proporcionar datos de reservas, como datos de habitaciones, datos de tarifas y datos de clientes. Estos datos pueden utilizarse para realizar análisis de reservas, como la optimización de las tarifas y la mejora de la ocupación.
  - **Datos de operaciones:** El DWH debe proporcionar datos de operaciones, como datos de empleados, datos de procesos y datos de sistemas. Estos datos pueden utilizarse para realizar análisis de operaciones, como la mejora de la eficiencia y la reducción de costes.
  - **Datos de marketing:** El DWH debe proporcionar datos de marketing, como datos de clientes, datos de campañas y datos de ventas. Estos datos pueden utilizarse para realizar análisis de marketing, como la segmentación de clientes y la optimización de las campañas de marketing.