

Resumen Data Warehouse y BI

Unidad 1 - Data Warehouse

1. ¿Qué es un Data Warehouse? ¿Qué ventajas tiene? ¿Desventajas? ¿Para qué sirve? R: Es un sistema de almacenamiento orientado a temas, integrado, no volátil y variante en el tiempo para soporte de decisiones.

- Ventajas: Integración de datos, rendimiento optimizado para consultas, histórico completo, datos limpios y consistentes.
- Desventajas: Alto costo inicial, complejidad de implementación, latencia en actualización de datos.

2. ¿Por qué no podemos usar el transaccional como fuente de datos para el análisis? R: Porque está optimizado para transacciones CRUD, no para consultas complejas, impacta el rendimiento del sistema operativo, los datos están dispersos y sin integrar, y falta datos históricos.

3. ¿Cuál es la diferencia entre datos, información y conocimiento? R:

- Datos: Hechos en bruto sin contexto
- Información: Datos procesados con significado
- Conocimiento: Información aplicada con experiencia y juicio

4. ¿Qué puede llevar al fracaso a un sistema de información? R: Falta de apoyo gerencial, requerimientos mal definidos, datos de mala calidad, falta de capacitación de usuarios.

5. ¿Cuáles son los problemas frecuentes en las bases de datos? R: Inconsistencia de datos, redundancia, falta de integración, datos obsoletos o incorrectos.

6. ¿Cuáles son los principios al diseñar un DW? R: Orientado a temas, integrado, no volátil, variante en el tiempo.

7. ¿Cuáles son las etapas de diseño? R: 1. Análisis de requerimientos, 2. Diseño conceptual, 3. Diseño lógico, 4. Diseño físico, 5. Implementación.

8. ¿Qué tipos de DW existen? ¿Cuál es el que se desarrolla en el TP? R: Enterprise DW (corporativo completo), Data Mart (departamental - usado en el TP), Virtual DW (vista lógica).

9. ¿En qué se diferencia un DW de una BBDD Relacional? R:

- DW: Orientado a consultas, desnormalizado, datos históricos
- BBDD: Orientado a transacciones, normalizado, datos actuales

10. ¿Qué nos dice la tabla fact? R: La tabla fact muestra ID_Fecha, ID_Sucursal, ID_Cliente, ID_Articulo con métricas Importe_Venta y Cant_Venta.

¿Qué modelo se usa en el TP? ¿Qué otro modelo existe? R: Modelo estrella (usado en el TP). Otro modelo: Copo de nieve (snowflake).

11. ¿Qué es ETL? ¿Para qué sirve? R: Extract, Transform, Load. Sirve para extraer datos de fuentes, transformarlos y cargarlos al DW.

12. ¿Qué tipos de extracción existen? R: Full (extracción completa), Incremental (solo cambios), Delta (basada en timestamps).

13. Nombrar procesos de transformación R: Limpieza de datos, validación, normalización, agregación, derivación de campos.

14. ¿En qué momento del ETL se agregan las keys? R: Durante la fase de Transform del ETL.

15. ¿En qué momento del ETL se hacen los cálculos de métricas? R: Durante la fase de Transform del ETL.

16. ¿Qué pasa con los datos que ya están en el DW cuando se ejecuta el proceso ETL? R: Pueden ser: Refresh (se reemplazan), Append (se agregan nuevos), Update (se actualizan existentes).

17. ¿Cuáles son los beneficios del uso de ETL? R: Automatización, calidad de datos, integración, consistencia.

18. Nombrar fuentes de datos para el proceso ETL R: BBDD relacionales, archivos planos (CSV, TXT), APIs, sistemas ERP/CRM.

19. ¿Qué tipos de datos existen? R: Estructurados (tablas, BBDD), semiestructurados (XML, JSON), no estructurados (texto, imágenes, videos).

20. ¿Cuáles son las 3 principales V's del Big Data? R: Volumen (cantidad de datos), Velocidad (rapidez de generación), Variedad (tipos de datos diversos).

21. ¿Qué es Big Data? R: Conjuntos de datos tan grandes y complejos que requieren herramientas especializadas para su procesamiento y análisis.

22. ¿Qué es OLTP? R: Online Transaction Processing: Sistemas optimizados para transacciones frecuentes y rápidas.

23. ¿Cómo se carga la dimensión fecha? R: Se precarga con todas las fechas del rango de tiempo del DW, incluyendo atributos como día, mes, año, trimestre, etc.

24. ¿Qué datos se pueden descartar del transaccional? R: Datos temporales del sistema, logs técnicos, datos de auditoría interna, campos de control técnico.

Unidad 2 - OLAP

1. ¿Qué es OLAP? R: Online Analytical Processing: Tecnología para análisis multidimensional de datos.

- 2. ¿Cómo se relaciona con DW? R:** OLAP utiliza datos del DW para crear cubos multidimensionales que permiten análisis rápido.
- 3. ¿Cuáles son las características y ventajas de OLAP? R:** Análisis multidimensional, navegación intuitiva (drill-down, roll-up), cálculos complejos, respuestas rápidas.
- 4. ¿Qué tipos de medidas existen? R:** Aditivas (se pueden sumar como ventas), semi-aditivas (se suman en algunas dimensiones como inventario), no aditivas (no se pueden sumar como porcentajes).
- 5. ¿Cuál es la periodicidad de las medidas? R:** Pueden ser diarias, semanales, mensuales, trimestrales, anuales según el negocio.
- 6. ¿Qué es una dimensión? R:** Perspectiva o contexto para analizar los datos (tiempo, geografía, producto).
- 7. ¿Qué pueden contener? R:** Atributos descriptivos, jerarquías, claves primarias y foráneas.
- 8. ¿Qué es una jerarquía? R:** Estructura de niveles dentro de una dimensión (Año → Trimestre → Mes → Día).
- 9. ¿Qué es la multidimensionalidad? R:** Capacidad de analizar datos desde múltiples perspectivas simultáneamente.
- 10. ¿Qué nos dice la tabla fact? R:** Contiene las métricas del negocio y las claves foráneas hacia las dimensiones.
- 11. ¿Qué es la tabla de hecho? R:** Tabla central que almacena medidas cuantitativas del negocio.
- 12. ¿Cómo se arma? R:** Identificar proceso de negocio, definir granularidad, identificar dimensiones, identificar métricas.
- 13. ¿Qué relación hay entre tabla de hecho, dimensiones y modelo estrella? R:** La tabla hecho está en el centro conectada directamente con todas las dimensiones, formando una estrella.
- 14. ¿Qué son las dimensiones conformadas? R:** Dimensiones compartidas entre múltiples tablas de hecho que tienen la misma estructura y significado.

Unidad 3 - Data Mining y BI

- 1. ¿Qué es un data lake? R:** Repositorio de almacenamiento que contiene datos en su formato nativo hasta que se necesiten.
- 2. ¿Qué es data mining? R:** Proceso de descubrir patrones, relaciones y conocimiento en grandes conjuntos de datos.
- 3. ¿Cuáles son los pasos del data mining? R:** 1. Entendimiento del negocio, 2. Entendimiento de los datos, 3. Preparación de datos, 4. Modelado, 5. Evaluación, 6. Despliegue.

4. ¿Qué responde la primera etapa? R: Qué problema de negocio se quiere resolver y qué objetivos se persiguen.

5. ¿Qué implica la limpieza de datos? R: Eliminar duplicados, corregir errores, manejar valores faltantes, normalizar formatos.

6. ¿Qué técnicas hay en DM? R: Clasificación, clustering, reglas de asociación, regresión, árboles de decisión.

7. ¿Qué es BI? R: Business Intelligence: Conjunto de herramientas, aplicaciones y metodologías para convertir datos en información útil para la toma de decisiones.

8. ¿Qué implica la calidad de datos? R: Exactitud (datos correctos), completitud (datos completos), consistencia (datos coherentes), actualidad (datos actualizados).

9. ¿Qué son los datos críticos? R: Datos esenciales para el funcionamiento del negocio cuya pérdida o corrupción tendría impacto severo.

Unidad 4 - NoSQL

1. ¿Qué significan las siglas NoSQL? R: Not Only SQL: Bases de datos no relacionales que no usan SQL como lenguaje principal.

2. ¿Qué ofrecen? R: Escalabilidad horizontal, flexibilidad de esquema, alto rendimiento, disponibilidad.

3. ¿En qué se diferencian con las BBDD relacionales? R: Sin esquema fijo, no requieren ACID completo, escalabilidad horizontal, diferentes modelos de datos.

4. ¿Cuáles son los 3 principios de NoSQL? R: Escalabilidad (crecimiento horizontal), disponibilidad (sistema siempre accesible), tolerancia a particiones (funciona aunque fallen nodos).

5. ¿Qué tipos de NoSQL existen? ¿Cuál es la estructura de cada uno? R: Clave-Valor (Redis, DynamoDB), documentales (MongoDB, CouchDB), columnar (Cassandra, HBase), grafos (Neo4j, Amazon Neptune).

Unidad 5 - Datos Estructurados

1. ¿Qué es un dato estructurado? R: Datos organizados en formato fijo con esquema predefinido (tablas, campos específicos).

2. ¿Qué son los datos semiestructurados? R: Datos con cierta organización pero sin esquema rígido (XML, JSON).

3. ¿Qué tipos existen? R: XML, JSON, CSV con metadatos, archivos de configuración.

4. ¿Qué es una BBDD documental? ¿Qué ventajas y desventajas tiene? R: Almacena datos como documentos (JSON, XML).

- Ventajas: Flexibilidad de esquema, escalabilidad, consultas complejas.
- Desventajas: Menos madurez, posible inconsistencia, curva de aprendizaje.

5. ¿En qué se diferencia con las relacionales? R: Esquema flexible vs rígido, documentos vs filas, consultas de documentos vs SQL, escalabilidad horizontal vs vertical.