Resumen - Fundamentos de Data Warehouse y BBDD

Unidad 1 - Data Warehouse

Conceptos Fundamentales

Data Warehouse (DW): Repositorio centralizado que integra datos históricos de múltiples fuentes para análisis y toma de decisiones estratégicas.

ETL (Extract, Transform, Load): Proceso fundamental para:

- **Extract**: Obtener datos de sistemas fuente
- **Transform**: Limpiar, validar y convertir datos
- **Load**: Cargar datos en el DW

Tipos de Data Warehouse

Tipo	Descripción	Alcance
EDW	Enterprise Data Warehouse	Organizacional completo
Data Mart	Subset del EDW	Departamental/específico
ODS	Operational Data Store	Datos operativos recientes

Modelos de Datos

Modelo Estrella:

- Tabla central de hechos
- Dimensiones conectadas directamente
- Desnormalizado, consultas rápidas

Modelo Copo de Nieve:

- Dimensiones normalizadas
- Menor espacio de almacenamiento
- Consultas más complejas

Principios de Diseño DW

1. Foco en objetivos del negocio
2. Comenzar con el fin en mente
3. Pensar y diseñar un todo, construir de a poco
4. Colaborar con stakeholders

5. Construir metadata desde el inicio

Unidad 2 - OLAP (Online Analytical Processing)

Definición y Características

OLAP: Tecnología para análisis multidimensional rápido de grandes volúmenes de datos.

Características principales:

- Facilita consultas de business intelligence
- Utiliza estructuras multidimensionales (cubos)
- Navegación interactiva de datos

Operaciones OLAP

- **Drill Down**: Mayor nivel de detalle
- **Drill Up (Roll Up)**: Mayor nivel de agregación
- **Slicing and Dicing**: Diferentes perspectivas de datos
- **Consolidación**: Agrupaciones y acumulaciones

Dimensiones y Hechos

Dimensiones:

- Contextualizan las mediciones
- Ejemplos: Tiempo, Producto, Cliente, Geografía
- Contienen jerarquías (Año → Trimestre → Mes → Día)

Hechos:

- Datos medibles del negocio
- Métricas cuantitativas
- Determinados por las dimensiones

Tipos de Medidas

Tipo	Descripción	Ejemplo
Aditivas	Se suman en todas las dimensiones	Ventas, Cantidad
Semi-aditivas	Se suman solo en algunas dimensiones	Stock, Saldos
No aditivas	No se suman	Ratios, Porcentajes

Implementaciones OLAP

ROLAP (Relational OLAP):

- Usa bases relacionales
- Consultas SQL dinámicas
- Escalable, pero más lento

MOLAP (Multidimensional OLAP):

- Cubos pre-calculados
- Consultas muy rápidas
- Limitado en tamaño

HOLAP (Hybrid OLAP):

- Combina ROLAP y MOLAP
- Balance entre rendimiento y escalabilidad

Unidad 3 - Explotación de Datos

Data Lakes vs Data Warehouse

Data Lake:

- Almacena datos en formato nativo
- Soporta datos estructurados y no estructurados
- Mayor flexibilidad, menor governance

Data Warehouse:

- Datos procesados y estructurados
- Schema definido
- Mayor governance, menor flexibilidad

Data Mining

Proceso en 6 etapas:

1. **Definición del problema**
   - ¿Qué se busca predecir?
   - Definir métricas de éxito
2. **Preparación de datos**
   - Limpieza y consolidación
   - Manejo de datos faltantes

3. **Exploración de datos**
    - Análisis estadístico descriptivo
    - Identificación de patrones
4. **Creación de modelos**
    - Selección de algoritmos
    - Entrenamiento de modelos
5. **Validación de modelos**
    - Evaluación de rendimiento
    - Comparación de modelos
6. **Implementación y actualización**
    - Despliegue en producción
    - Monitoreo continuo

Business Intelligence (BI)

Definición: Transformación de datos en conocimiento para obtener ventaja competitiva.

Componentes:

- Herramientas de análisis
- Reportes y dashboards
- Análisis predictivo
- Visualización de datos

Calidad de Datos

Conceptos clave:

- **Valor**: Aumentar utilidad de los datos
- **Riesgo**: Reducir costos de mala calidad
- **Prevención**: Evitar errores desde el origen
- **Causa Raíz**: Solucionar problemas fundamentales

Unidad 4 - NoSQL

Características Principales

NoSQL ("Not Only SQL"):

- Alternativa a bases relacionales
- Mayor flexibilidad de esquemas

- Escalabilidad horizontal
- Optimizado para big data

Teoremas y Principios

Teorema CAP: Solo se pueden garantizar 2 de 3:

- **Consistency**: Consistencia de datos
- **Availability**: Disponibilidad del sistema
- **Partition tolerance**: Tolerancia a particiones

ACID vs BASE:

ACID (Bases relacionales):

- **A**tomicity: Todo o nada
- **C**onsistency: Reglas de integridad
- **I**solation: Transacciones independientes
- **D**urability: Persistencia garantizada

BASE (Bases NoSQL):

- **B**asically Available: Disponibilidad básica
- **S**oft state: Estado puede cambiar
- **E**ventual consistency: Consistencia eventual

Tipos de NoSQL

1. Key-Value:

- Estructura: Clave → Valor
- Ejemplos: Redis, Riak
- Uso: Caché, sesiones

2. Document:

- Estructura: Documentos JSON/XML
- Ejemplos: MongoDB, CouchDB
- Uso: Aplicaciones web, CMS

3. Graph:

- Estructura: Nodos y relaciones
- Ejemplos: Neo4j, OrientDB

- Uso: Redes sociales, recomendaciones

4. Wide Column:

- Estructura: Familias de columnas
- Ejemplos: Cassandra, ScyllaDB
- Uso: Big data, analytics

Unidad 5 - Datos de Otras Fuentes

Clasificación de Datos

Datos Estructurados:

- Formato fijo y predefinido
- Almacenados en tablas
- Fácil consulta con SQL

Datos Semi-estructurados:

- Formato autodefinido
- Ejemplos: XML, JSON
- Metadatos incluidos

Datos No estructurados:

- Sin formato específico
- Ejemplos: PDFs, imágenes, videos
- Requieren procesamiento especial

Formatos Comunes

XML (eXtensible Markup Language):

- Lenguaje de marcado
- Intercambio de datos entre sistemas
- Estructura jerárquica con etiquetas

JSON (JavaScript Object Notation):

- Formato ligero de intercambio
- Estructura clave-valor
- Ampliamente usado en APIs web

Bases de Datos Documentales

Características:

- Almacenan documentos completos
- Schema flexible
- Escalabilidad horizontal

Ventajas:

- Flexibilidad de estructura
- Desarrollo ágil
- Buen rendimiento en lectura

Desventajas:

- No siempre garantiza ACID
- Menos herramientas maduras
- Consumo de memoria en índices

Comparación de Tecnologías

OLTP vs Data Warehouse vs OLAP

Aspecto	OLTP	Data Warehouse	OLAP
Propósito	Transacciones operativas	Almacenamiento histórico	Análisis multidimensional
Usuarios	Muchos concurrentes	Pocos analistas	Analistas de negocio
Datos	Actuales, detallados	Históricos, integrados	Agregados, sumarizados
Modelo	Normalizado	Dimensional	Cubos
Consultas	Simples, rápidas	Complejas, batch	Analíticas, interactivas
Actualización	Continua	Periódica (ETL)	Pre-calculada

SQL vs NoSQL

Característica	SQL	NoSQL
Schema	Fijo	Flexible
Escalabilidad	Vertical	Horizontal
Consistencia	ACID	BASE
Consultas	SQL estándar	Específicas por tipo
Uso típico	Aplicaciones tradicionales	Big data, web scale

Conceptos Clave para Desarrollo

ETL Best Practices

- Validación de datos en cada etapa
- Manejo de errores y excepciones
- Logging detallado
- Procesos incrementales
- Monitoreo de calidad

Diseño Dimensional

- Identificar procesos de negocio
- Definir granularidad de hechos
- Identificar dimensiones
- Crear dimensiones conformadas
- Optimizar para consultas

Selección de Tecnología

Usar SQL cuando:

- Datos estructurados
- Necesidad de ACID
- Consultas complejas
- Equipo familiarizado

Usar NoSQL cuando:

- Datos no estructurados
- Escalabilidad horizontal
- Desarrollo ágil
- Big data

Glosario

Dimensiones Conformadas: Dimensiones compartidas entre múltiples tablas de hechos con definiciones consistentes.

Granularidad: Nivel de detalle de los datos almacenados en una tabla de hechos.

Slowly Changing Dimensions (SCD): Técnicas para manejar cambios en dimensiones a lo largo del tiempo.

Fact Table: Tabla central en modelo estrella que contiene medidas numéricas.

Star Schema: Modelo dimensional con tabla de hechos central rodeada de dimensiones desnormalizadas.

Snowflake Schema: Variación del modelo estrella con dimensiones normalizadas.

Data Lineage: Seguimiento del origen y transformaciones de los datos.

Master Data: Datos maestros que representan entidades clave del negocio.

Este resumen cubre los conceptos fundamentales para el desarrollo de soluciones de Data Warehouse y análisis de datos.