

BigData

Data Warehouse en Big Data: Fundamentos, Arquitectura y Aplicaciones (Neo4J)

Lectura de CHAPTER 4: Enterprise Technologies and Big Data Business Intelligence

Luis Fernando Castellanos

Guarin

2025



UNIVERSIDAD
CENTRAL



1. Objetivos de Aprendizaje

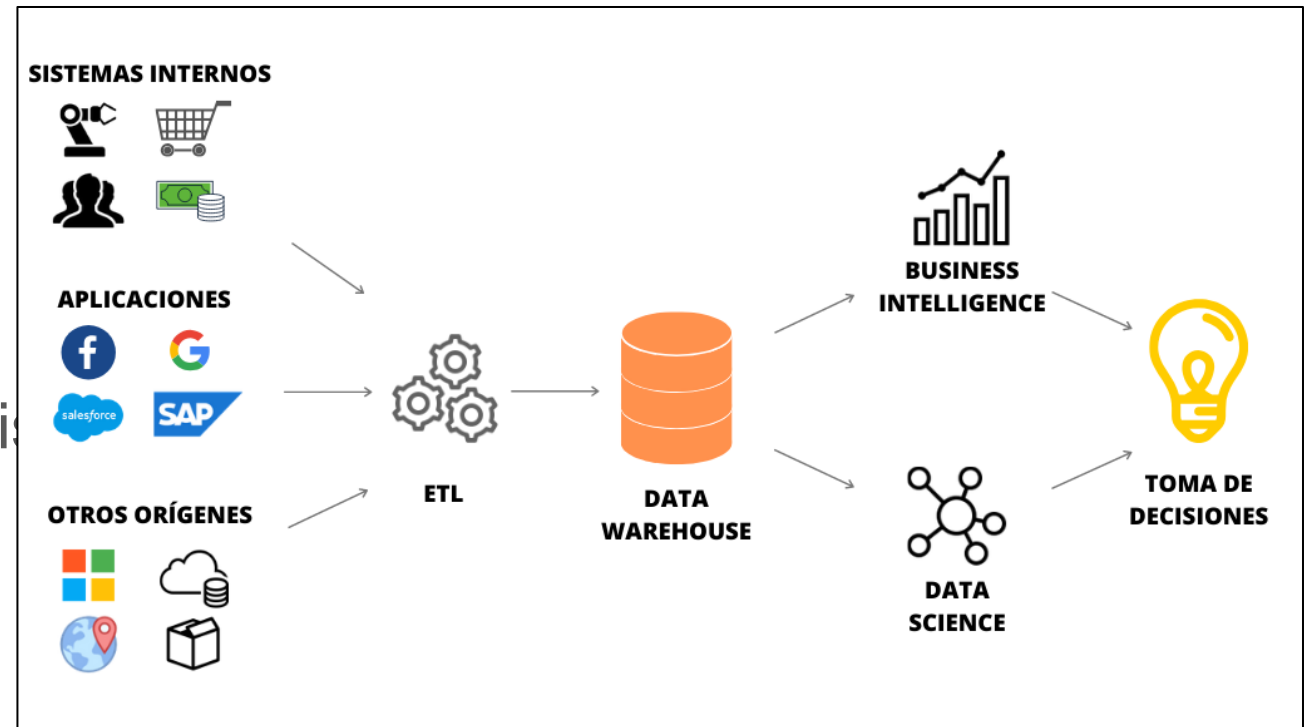
- Comprender los conceptos fundamentales de Data Warehouse y su evolución
- Identificar las diferencias entre OLTP y OLAP en contextos empresariales
- Analizar la arquitectura tradicional de Data Warehouse y su transformación ante Big Data
- Explorar las tecnologías de ETL y su adaptación a volúmenes masivos de datos
- Evaluar casos de uso reales de Data Warehouse en entornos de Big Data

2. Objetivos de Aprendizaje

- Definición: "Un Data Warehouse es un repositorio centralizado que almacena datos históricos e integrados de múltiples fuentes para soportar el análisis y la toma de decisiones"
- Características clave:
 - Orientado a temas específicos del negocio
 - Integrado (datos consistentes)
 - No volátil (los datos históricos no cambian)
 - Variable en el tiempo (mantiene la dimensión temporal)

3. Arquitectura Tradicional de Data Warehouse

- Componentes principales:
 - Fuentes de datos operacionales
 - Área de staging (preparación)
 - Capa ETL
 - Data Warehouse central
 - Data Marts departamentales
 - Herramientas de consulta y análisis
- El modelo de tres capas:
 - Capa de fuentes
 - Capa de integración
 - Capa de presentación
- Modelos de diseño: Esquema de estrella vs. Esquema de copo de nieve



<https://www.mistralbs.com/wp-content/uploads/2020/01/dwh-2.png>

4. OLTP vs OLAP

- **OLTP (Online Transaction Processing):**
 - Orientado a transacciones diarias del negocio
 - Alta concurrencia y tiempos de respuesta rápidos
 - Operaciones de lectura/escritura frecuentes
 - Optimizado para inserción y actualización
 - Almacena datos actuales del negocio

- **OLAP (Online Analytical Processing):**
 - Orientado al análisis multidimensional
 - Consultas complejas sobre grandes volúmenes de datos
 - Principalmente operaciones de lectura
 - Optimizado para consultas agregadas
 - Almacena datos históricos para análisis

5. ETL - Extract, Transform, Load

- Definición: "ETL es el proceso de extraer datos de sistemas fuente, transformarlos para cumplir con los requisitos del negocio, y cargarlos en un Data Warehouse"
- Fases del proceso ETL:
 - Extracción: Obtención de datos de fuentes heterogéneas
 - Transformación: Limpieza, normalización y aplicación de reglas de negocio
 - Carga: Inserción en el Data Warehouse según modelo dimensional
- Herramientas tradicionales: *Informatica PowerCenter, IBM DataStage, Microsoft SSIS*
- Desafíos con Big Data: **Volumen, velocidad y variedad**

6. Data Marts - Subconjuntos Especializados

- Definición: "Un Data Mart es un subconjunto de un Data Warehouse enfocado en un área específica del negocio"
- Tipos de Data Marts:
 - Dependientes (derivados del Data Warehouse)
 - Independientes (construidos directamente desde fuentes)
 - Híbridos (combinación de ambos enfoques)
- Ventajas:
 - Tiempo de implementación reducido
 - Satisface necesidades departamentales específicas
 - Mayor rendimiento para análisis especializados
- Integración con la estrategia global de Data Warehouse

7. El Desafío de Big Data para Data Warehouses

- **Las 5 V's de Big Data y su impacto en Data Warehousing:**
 - Volumen: Escala de petabytes y más
 - Velocidad: Datos en tiempo real y streaming
 - Variedad: Estructurados, semi-estructurados y no estructurados
 - Veracidad: Confiabilidad y calidad de los datos
 - Valor: Extracción de insights significativos
- **Limitaciones de las arquitecturas tradicionales:**
 - Costo prohibitivo para escalar verticalmente
 - Procesamiento por lotes insuficiente
 - Rigidez para incorporar nuevos tipos de datos

8. Arquitecturas Modernas de Data Warehouse para Big Data

- Evolución arquitectónica:
 - Logical Data Warehouse (LDW)
 - Data Warehouse en la nube
 - Arquitecturas híbridas (tradicional + Big Data)
- Tecnologías habilitadoras:
 - **Almacenamiento NoSQL**
 - Procesamiento distribuido (**Hadoop, Spark**)
 - **MPP (Massively Parallel Processing)**
 - Arquitecturas de streaming en tiempo real
- Integración con ecosistemas Big Data

Framework's que permite almacenar y procesar grandes volúmenes de datos en clusters de máquinas utilizando el paradigma **MapReduce y Machine Learning (Apache Spark)**

permite ejecutar múltiples tareas en paralelo a través de varios nodos o servidores, cada uno con su propia memoria y CPU, para mejorar el rendimiento y la escalabilidad, como son: **Amazon Redshift, Google BigQuery**

procesamiento de datos en tiempo real en lugar de esperar lotes:
Apache Kafka (mensajería distribuida para ingesta de datos)
Apache Flink (procesamiento en tiempo real con baja latencia)
Google Dataflow (integrado con GCP para procesamiento de streaming)




9. Data Lakes vs Data Warehouses

- Definición de Data Lake: "Repositorio que almacena datos en su formato nativo hasta que sean necesarios"
- Comparativa:
 - Data Warehouse: Datos estructurados, esquema predefinido, costoso pero optimizado
 - Data Lake: Todos los formatos, esquema flexible, económico pero requiere procesamiento posterior

Arquitectura de ref Data Warehouse

Caso de uso:
"Schema-on-read"

VS

Característica	Schema-on-read (Data Lake)	Schema-on-write (Data Warehouse)
 Momento de definición	En la lectura	En la escritura
 Flexibilidad	Alta	Baja
 Velocidad de ingesta	Rápida (sin transformación)	Lenta (requiere ETL)
 Precisión en consultas	Baja (datos pueden estar sin limpiar)	Alta (datos estructurados y validados)
 Casos de uso	Big Data, IA, Machine Learning	BI, Reportes, Análisis estructurado

10. Herramientas y Tecnologías para Data Warehousing en Big Data

Plataformas modernas de Data Warehouse:

- **Snowflake:** Data Warehouse nativo en la nube
- **Amazon Redshift:** Solución MPP en AWS
- **Google BigQuery:** Servicio serverless
- **Azure Synapse Analytics:** Plataforma unificada

Tecnologías complementarias:

- **Apache Hive:** Data Warehouse sobre Hadoop
- **Apache Drill:** SQL sobre múltiples fuentes
- **Presto:** Motor de consultas distribuidas

11. Sectores de Mayor Adopción del uso de DataWareHouse

1. **Financiero:** Para análisis de riesgo, detección de fraude y personalización de servicios
2. **Retail:** Para optimización de inventario, análisis de comportamiento del consumidor y marketing personalizado
3. **Telecomunicaciones:** Para análisis de redes, experiencia del cliente y **predicción de churn:** *identificar a los clientes que pueden dejar de usar un servicio o producto antes de que lo hagan*
4. **Público:** Para servicios ciudadanos, planificación urbana y estadísticas nacionales...en Colombia será que si?

12. Estándares y Normativas

Estándares internacionales:

- [ISO/IEC 27001: Seguridad de la información](#)
- [DAMA-DMBOK: Gestión de datos](#) **(libro)**

Marco normativo colombiano:

- [Ley 1581 de 2012: Protección de datos personales](#)
- [Decreto 1377 de 2013: Reglamenta la Ley 1581 de 2012](#): Define los procedimientos para la autorización del tratamiento de datos personales y los derechos de los titulares.
- [Ley 1341 de 2009: Ley TIC](#) (modificada por la [Ley 1978 de 2019](#)): Promueve el desarrollo tecnológico, incluyendo aspectos relacionados con la gestión y análisis masivo de datos.
- [Circular 002 de 2018 de la SIC: Estándares de seguridad](#)
- [CONPES 3920: Política de explotación de datos](#)

Consideraciones éticas y de privacidad en Big Data

Conclusiones y Tendencias Futuras

Evolución continua: Data Mesh, Data Fabric y arquitecturas descentralizadas

Automatización: DataOps y MLOps para Data Warehousing

Tiempo real: Convergencia entre streaming y procesamiento analítico

Democratización: Self-service BI sobre arquitecturas modernas

Recomendaciones para implementación en organizaciones colombianas:

Evaluar madurez tecnológica actual

Adopción gradual de capacidades Big Data

Formación continua del talento local

Aprovechamiento de soluciones cloud para reducir barreras de entrada

Ejercicios de la sesión



UNIVERSIDAD
CENTRAL



Gracias