

BIG DATA & IA

BOOTCAMP

# Data Warehousing

# Data Warehouse



# Data Warehouse

Un Data Warehouse (Almacén de Datos) es una base de datos centralizada, estructurada y orientada a temas, diseñada para almacenar y consolidar grandes volúmenes de datos provenientes de múltiples fuentes y sistemas de una empresa. Su objetivo principal es permitir el análisis y la generación de informes para apoyar la toma de decisiones empresariales.



# Data Warehouse

Un DW es como una gigantesca biblioteca que almacena y organiza una gran cantidad de información sobre diferentes temas de una organización.

En esa biblioteca, los datos están clasificados en estanterías ordenadas de manera especial. Cada estantería contiene datos relacionados con un tema específico, como una estantería para ventas, otra para clientes, otra para productos, etc.

La magia del Data Warehouse radica en que estos datos se recopilan de diferentes lugares y sistemas de la empresa, como el sitio web, la tienda física, las redes sociales y más. Luego, todos esos datos se organizan cuidadosamente para que sea fácil acceder a ellos y analizarlos.




# Data Warehouse


Cuando los gerentes, analistas o cualquier persona necesitan obtener información sobre la empresa, pueden acudir a esta biblioteca de datos y buscar en las estanterías relevantes. Así, el Data Warehouse les proporciona una vista completa y clara de cómo está funcionando el negocio y les ayuda a tomar decisiones inteligentes.


En resumen, un Data Warehouse es como una poderosa biblioteca que almacena datos de una empresa y los ordena de manera inteligente para que cualquiera pueda buscar y entender la información de forma sencilla.



# Tipos de Data Warehouse

 **Tradicional (On premise):** Tiene servidores físicos donde se almacenan los datos.

 **Cloud:** Almacenamiento en la nube. SaaS, es un servicio que se ofrece como software sin necesidad de infraestructura o mantenimiento. Permite escalabilidad rápida.

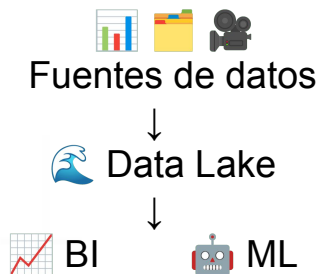
 **Virtual:** No se almacenan datos. Integraciones que realizan consultas en vivo a las fuentes.

# Data Lake

# Data Lake

Imagina un lago enorme donde puedes almacenar todo tipo de datos sin preocuparte por cómo están estructurados o formateados. Un Data Lake es como ese lago, pero para datos de una empresa u organización.

En lugar de utilizar bases de datos estructuradas y organizadas como en un Data Warehouse, el Data Lake guarda todos los datos tal como son: estructurados, semi-estructurados y no estructurados. Esto incluye archivos de texto, imágenes, videos, registros, documentos y más.







# Data Lake

El objetivo principal del Data Lake es tener un repositorio centralizado donde puedas almacenar cantidades masivas de datos sin importar su origen o formato. Esto es útil porque no siempre sabemos de antemano qué preguntas queremos hacer o qué información necesitamos. Con el Data Lake, podemos guardar todos los datos y luego analizarlos y extraer información valiosa más adelante.

Para acceder y analizar los datos en el Data Lake, se utilizan herramientas y tecnologías de procesamiento de datos, como SQL, Spark o Hadoop, que permiten realizar consultas y análisis complejos sobre el conjunto completo de datos.

En resumen, un Data Lake es como un lago gigante donde puedes almacenar todo tipo de datos, sin preocuparte por su estructura, para poder analizarlos y extraer información valiosa cuando la necesites. Es una herramienta poderosa para la ciencia de datos y el análisis de grandes cantidades de información.

# Data Mart



# Data Mart

Un Data Mart es un subconjunto de información especializado de un Data Warehouse, enfocado en un área específica de la organización, como ventas, marketing o finanzas.

Se diseña para que usuarios de negocio (analistas, responsables de área) puedan acceder rápidamente a la información que necesitan sin tener que consultar todo el Data Warehouse.

Por ejemplo, el Data Mart de ventas contendría información específica sobre clientes, productos vendidos y ventas, mientras que el Data Mart de marketing contendría datos sobre campañas, clientes potenciales y análisis de mercado.

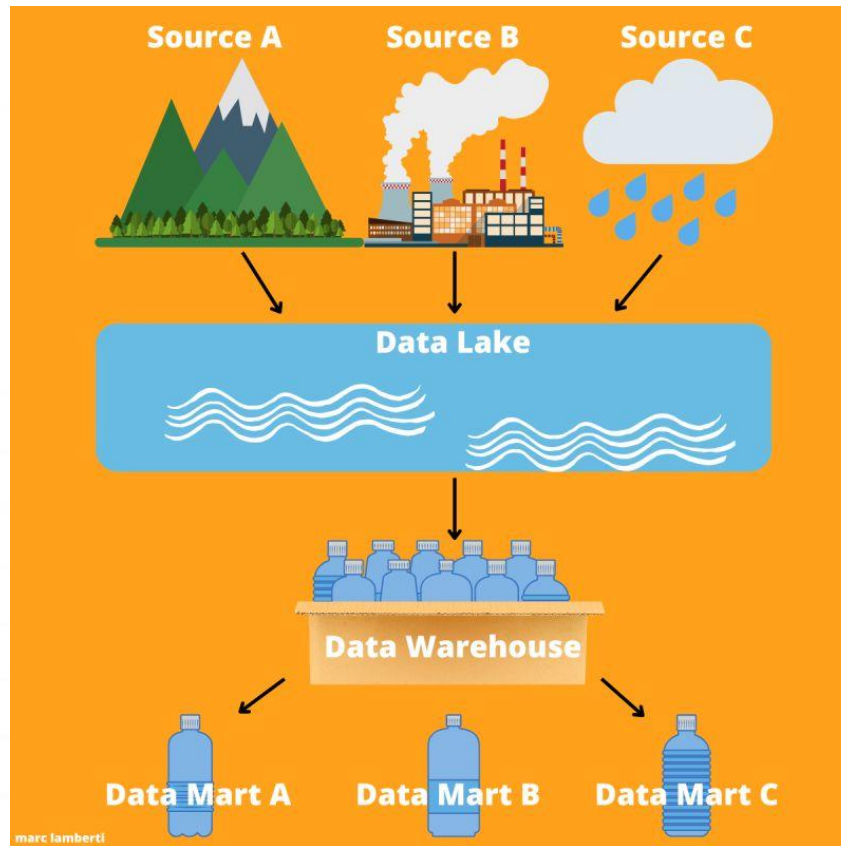
# Data Mart

**Enfoque temático** 🎯: cada Data Mart sirve a un área concreta.

**Subconjunto de datos** 📦: extrae solo lo relevante del Data Warehouse.

**Acceso más rápido** ⚡: consultas optimizadas para necesidades específicas.

**Simplicidad** 🧩: más fácil de usar que el Data Warehouse completo.



	Data Warehouse	Data Lake	Data Mart
<b>Definición</b>	Base de datos centralizada y estructurada que almacena datos históricos y actuales de una empresa.	Repositorio masivo y flexible que almacena datos en su forma original, independientemente de su estructura o formato.	Tienda de datos especializada que contiene información relevante para un grupo de usuarios o departamento específico dentro de una empresa.
<b>Estructura</b>	Datos altamente estructurados y organizados en tablas relacionales.	Datos no estructurados o semiestructurados almacenados tal y como son, como archivos, documentos, imágenes y más.	Datos organizados y adaptados para satisfacer las necesidades de un departamento o grupo de usuarios específico.
<b>Propósito</b>	Facilita el análisis empresarial y la toma de decisiones estratégicas.	Permite almacenar grandes volúmenes de datos diversos para futuros análisis y exploración.	Proporciona acceso rápido y sencillo a datos específicos para departamentos especializados.
<b>Usuarios</b>	Utilizado por diversos usuarios y equipos en toda la organización.	Accedido por científicos de datos y analistas para explorar datos y realizar análisis complejos.	Destinado a grupos de usuarios o departamentos con necesidades particulares de información.
<b>Granularidad</b>	Generalmente a nivel granular y detallado para análisis de tendencias a largo plazo.	Puede almacenar datos tanto a nivel granular como a nivel detallado, según sea necesario.	A menudo contiene datos más resumidos y agregados para un análisis más específico.
<b>Escalabilidad</b>	Escalable para manejar grandes volúmenes de datos de toda la organización.	Altamente escalable para almacenar y procesar grandes cantidades de datos sin estructura.	Escalable según las necesidades de cada departamento o grupo de usuarios.

# **Cómo llegan los datos al DW**



# ETL / ELT

ETL es un proceso fundamental en la integración de datos, para mover información desde diferentes fuentes hacia un Data Warehouse o un Data Lake.

- **Extract:** Capturar datos de diversas fuentes, como bases de datos, archivos, aplicaciones y APIs.
- **Transform:** Limpiar, depurar y ajustar los datos a un formato coherente. Se realizan diversas operaciones para prepararlos y mejorar su calidad.
- **Load:** Cargar los datos en la base de datos, donde estarán listos para su análisis y consulta.



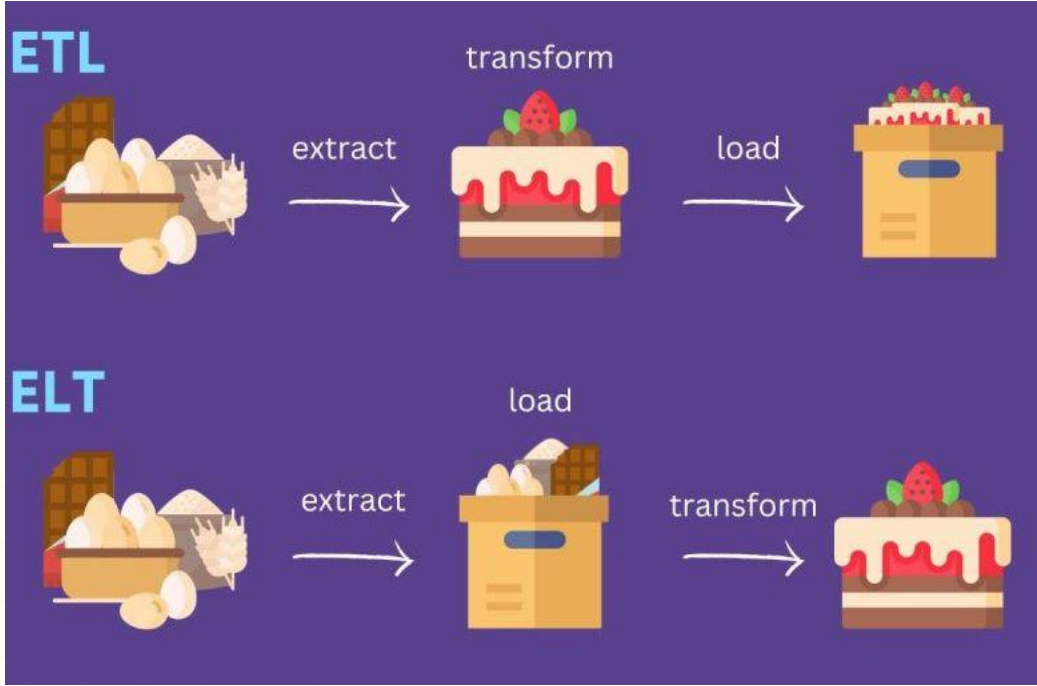


# ETL / ELT

- La diferencia entre ETL y ELT es el momento en el que se realiza la transformación de los datos.
- ETL, los datos en bruto no están disponibles en la base de datos. Se procesan antes de ello.
- ELT, es más útil para Big Data.
- ELT es más eficiente al utilizar la potencia informática de los sistemas de almacenamiento moderno.



# ETL / ELT





# Herramientas ETL

- Informatica PowerCenter
- SQL Server Integration Services (SSIS)
- Pentaho Data Integration (PDI)
- Talend
- Qlik
- Google Data Fusion

# Bases de datos OLTP y OLAP

- OLTP: es un tipo de base de datos diseñada para el procesamiento eficiente de transacciones en línea. Se utiliza para manejar operaciones diarias y rutinarias de una empresa, como registros de ventas, pedidos, reservas y transacciones financieras (insertar, actualizar base de datos).
- OLAP: es un tipo de base de datos diseñada para el análisis y la generación de informes empresariales. Se utiliza para realizar consultas complejas y análisis de grandes volúmenes de datos con el objetivo de obtener información valiosa para la toma de decisiones estratégicas.
- El DW es una base de datos OLAP.
- Las tablas en OLAP no están normalizadas.

# keep coding

