Data Warehousing

KEEPCODING
Tech School

Data Warehouse

Un DW es como una gigantesca biblioteca que almacena y organiza una gran cantidad de información sobre diferentes temas. En lugar de guardar libros, este "almacén de datos" guarda datos de una empresa, como ventas, clientes, inventarios y más.
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.
Cuando los gerentes, analistas o cualquier persona necesite 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.

Data Warehouse

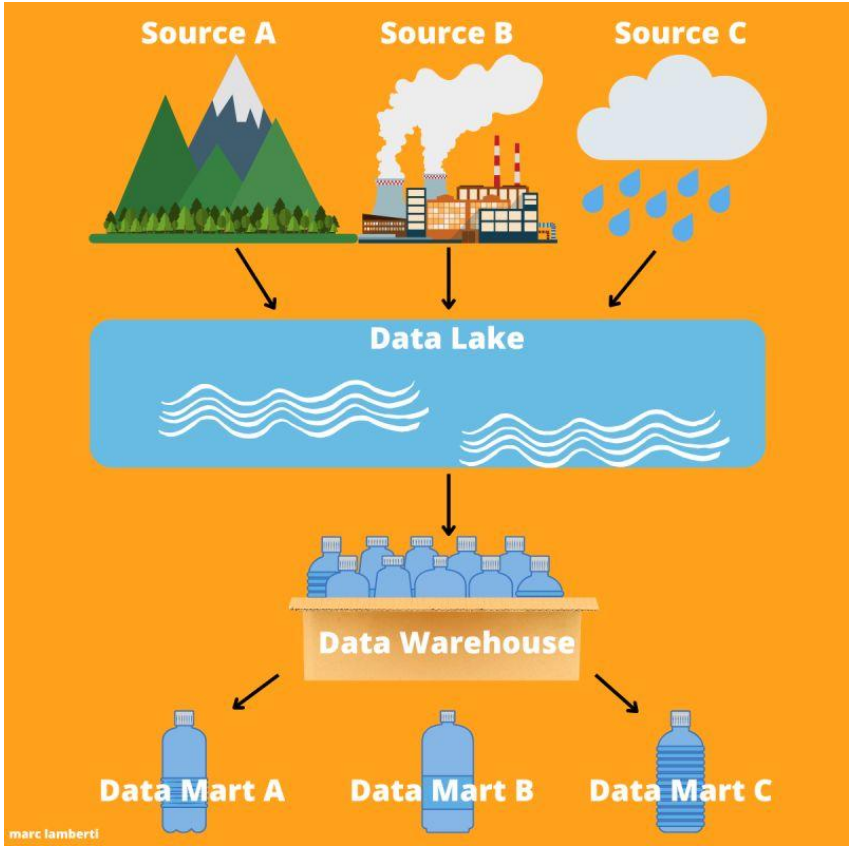
Un Data Warehouse (Almacén de Datos) es por tanto una base de datos centralizada, estructurada y orientada a temas, diseñada para almacenar y consolidar grandes volúmenes de datos provenientes de múltiples fuentes y sistemas de una empresa. Su objetivo principal es permitir el análisis y la generación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

Data 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.
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

Un Data Mart es como una tienda o una pequeña boutique que contiene una selección específica de datos importantes y útiles para un grupo de usuarios o un departamento en particular dentro de una empresa.
Imagina que una empresa tiene un gran almacén de datos (Data Warehouse) donde se guardan todos los datos de la compañía. Sin embargo, para facilitar el acceso y análisis de la información, se crean pequeñas tiendas de datos especializadas (Data Marts) que contienen datos relevantes para áreas específicas, como ventas, marketing, recursos humanos, etc.
Cada Data Mart está diseñado para satisfacer las necesidades de un grupo de usuarios concretos. 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.
La ventaja de los Data Marts es que permiten un acceso más rápido y sencillo a los datos específicos que cada departamento necesita, sin tener que buscar entre toda la información almacenada en el Data Warehouse. Esto facilita el análisis y la toma de decisiones para cada equipo de trabajo.
En resumen, un Data Mart es una tienda de datos especializada que contiene información relevante para un departamento o grupo de usuarios dentro de una empresa, lo que permite un acceso rápido y eficiente a la información específica que necesitan para sus actividades y decisiones.



De Marc Lamberti https://marclamberti.com/

Data Warehose – Data Lake – Data Mart

	Data Warehouse	Data Lake	Data Mart
Definición	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.
Estructura	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.
Propósito	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.
Usuarios	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.
Granularidad	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.
Escalabilidad	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.

ETL / ELT (Extract-Load-Transform)

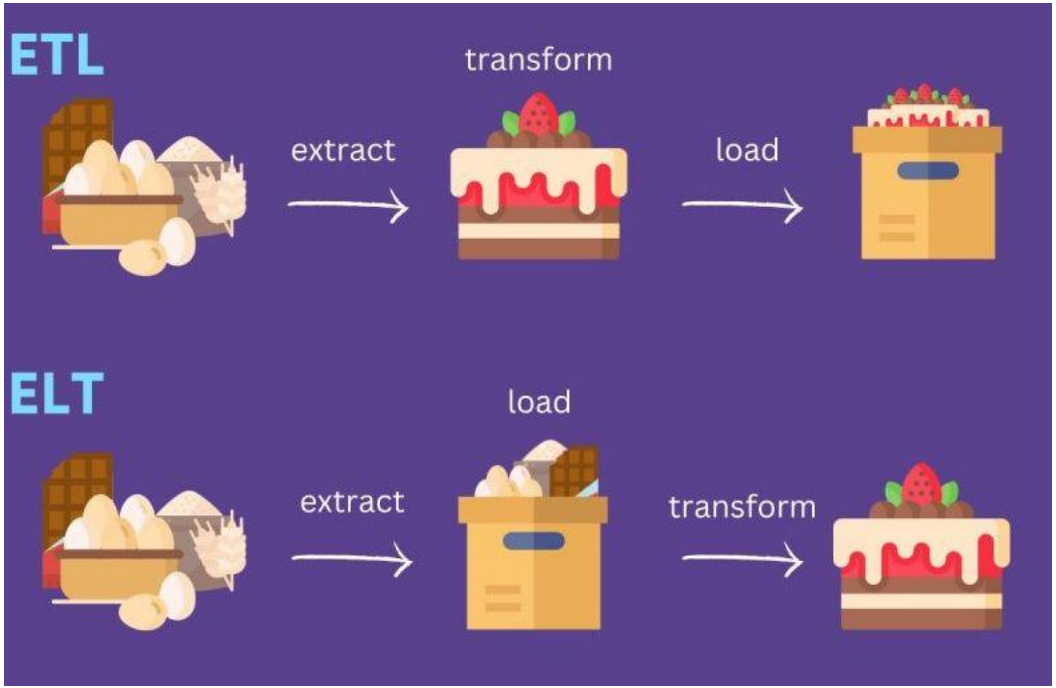
El proceso ETL es una metodología de integración de datos que se utiliza para mover y transformar datos desde múltiples fuentes heterogéneas hacia un destino común, como un Data Warehouse o un Data Mart.

- Extracción: En la primera etapa, los datos se extraen de diversas fuentes, como bases de datos, archivos, aplicaciones y APIs. Aquí se recopilan los datos necesarios para el análisis.
- Transformación: Una vez extraídos los datos, se lleva a cabo la etapa de transformación. Durante esta fase, los datos se limpian, se ajustan a un formato coherente y se realizan diversas operaciones para prepararlos y mejorar su calidad.
- Carga: Finalmente, los datos transformados se cargan en la base de datos destino, como el Data Warehouse, donde estarán listos para su análisis y consulta.

ETL / ELT (Extract-Load-Transform)

- La principal diferencia entre ETL y ELT radica en el momento en que ocurre la etapa de transformación.
- Con el ETL los datos sin procesar no están disponibles en el DW ya que se transforman antes de cargarlos.
- El ELT es más útil para el análisis de big data.
- ELT puede ser más eficiente al utilizar la potencia informática de los sistemas de almacenamiento de datos modernos.

ETL / ELT



De Marc Lamberti https://marclamberti.com/

Herramientas ETL

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

Base 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.

KEEPCODING

Tech School

Madrid | Barcelona | Bogotá

Datos de contacto