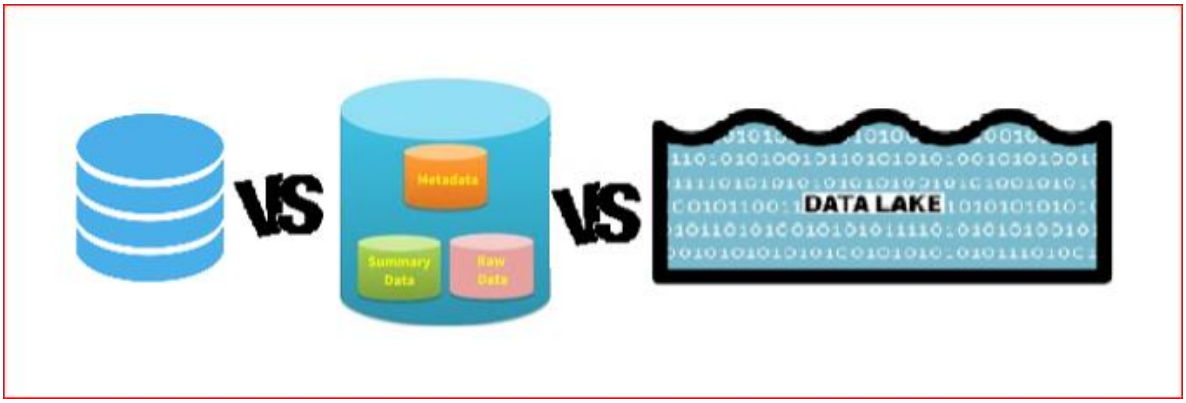
Base de Datos - Data Warehouse - Data Lake



Base de Datos, **Data Warehouse** y **Data Lake** son tres conceptos muy relacionados entre sí, pero no son lo mismo ni sirven para lo mismo, sirviendo cada uno a un propósito distinto.

Base de Datos

Para este ejercicio comparativo, nos referiremos a las bases de datos "clásicas", las "RDBMS", es decir, "Relational Database Management System" o, lo que es lo mismo, "Sistema de Gestión de Bases de Datos Relacionales". En otras palabras, las bases de datos SQL "de toda la vida" (SQL Server, Oracle, PostgreSQL, MySQL, etc.)

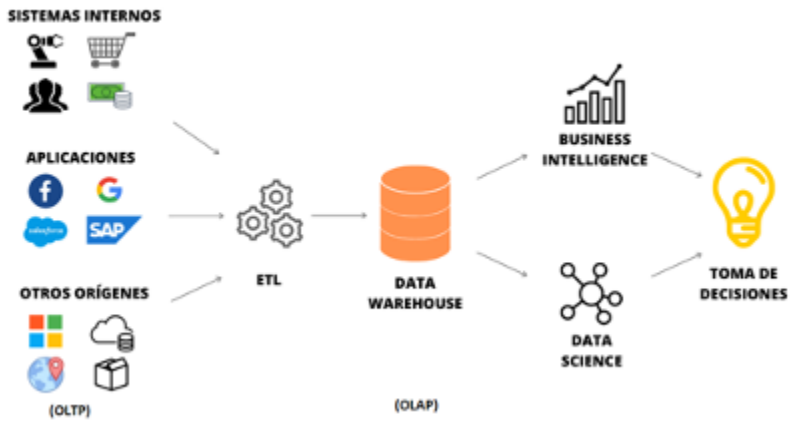
En una base de datos se almacena información estructurada, donde los datos se organizan en forma de tablas, filas y columnas (como una hoja Excel).
Ejemplo: Tabla Ventas, que contiene información acerca de todas las ventas realizadas en mi empresa. Cada fila de esa tabla corresponde a una venta. Y cada columna de una fila hace referencia a distintos atributos (propiedades) de una venta como, por ejemplo, día/hora de la venta, identificador de producto, identificador de vendedor, importe unitario, cantidad, importe total, código de cliente, etc.

Este tipo de almacenamiento se caracteriza por ser muy estructurado y estar **optimizado para realizar operaciones transaccionales** (inserción, borrado, actualización), es decir, para guardar los datos de softwares diseñados para guardar información (ERP, CRM, SGA, TPV, etc.), cuya arquitectura de datos sigue un diseño OLTP (Online Transaction Processing). Además, de forma habitual, **sólo tienen disponibles los datos más recientes**.

Por tanto, una base de datos con estructura OLTP, es el sistema de almacenamiento de datos que se encuentra debajo de los programas informáticos que usamos y que está diseñado para la introducción y modificación de datos en el sistema.

Data Warehouse (DWH)

Nos hemos dejado un tipo de operación fundamental, la **lectura (consulta)**. Cuando queremos analizar los datos que los softwares de gestión generan, la arquitectura de almacenamiento OLTP no es óptima, pues la escritura de sentencias SQL para consultar la información es compleja y, sobre todo, el tiempo de ejecución de esas consultas es muy elevado.
Para corregir esto, existe otra arquitectura, otra manera de organizar la información almacenada, llamada **OLAP (Online Analytical Processing)**. Un DWH es un sistema de base de datos que se construye encima de todas estas bases de datos OLTP y que se usa, típicamente, para **ser la fuente de datos de sistemas de Business Intelligence**.



El DWH obtiene la información de todos estos sistemas, limpia los datos, los unifica, les aplica reglas de negocio y, finalmente, los almacena con una estructura OLAP (típicamente, arquitectura en estrella -star schema- o copo de nieve -snow flake schema), creando así una capa de datos optimizada para ejecutar análisis de datos sobre ella. Al contrario que una base de datos, un DWH suele tener el **histórico de todos los datos** (ejemplo: no contiene solo la última dirección de un cliente, sino todas las direcciones que ha ido teniendo a lo largo de los años).
Por tanto, un DWH es una base de datos con estructura OLAP, que se encuentra por encima de las bases de datos transaccionales y que está diseñada para el análisis de los datos que hay en el sistema.

Data Lake (DL)

Hasta ahora hemos hablado de datos organizados de manera que su modificación es sencilla (OLTP) o su análisis es sencillo (OLAP). Pero todos tienen en común que son **datos estructurados**.
Con la aparición de tecnologías como Big Data, IoT, redes sociales, etc., se ha desarrollado la capacidad de procesar otro tipo de datos que **no son estructurados**, que a veces llegan en **tiempo real** y que, hasta ahora, simplemente ni se nos pasaba por la cabeza el poder analizarlos.
Nos referimos, por ejemplo, a:

- Fotos y vídeos, que pueden ser analizados por algoritmos de inteligencia artificial.
- Comentarios escritos en redes sociales, que pueden ser procesados por algoritmos de procesamiento natural del lenguaje para saber, por ejemplo, si un comentario dice algo positivo, negativo o neutro respecto a una marca o producto.
- Datos transmitidos por sensores IoT, que típicamente se transmiten almacenados en ficheros (a este tipo de almacenamiento también se le considera como datos semi-estructurados).
- Y todo, a menudo, siendo gestionado en tiempo real.

Aunque un DWH puede albergar datos no estructurados, cuando se combinan datos no estructurados con alta velocidad de procesamiento y un tamaño de almacenamiento fuera de lo habitual, lo normal es usar un data lake para gestionar datos en este escenario.

En resumen, **un Data Lake es un repositorio de datos centralizado, para el almacenamiento de datos estructurados y no estructurados, en tiempo real o no, donde los datos se almacenan tal cual están en el sistema origen**, es decir, sin aplicar ningún tipo de transformación sobre los mismos.

Similitudes y diferencias entre un DWH y DL

Un DWH y un Data Lake tienen ciertas características en común y algunas diferencias importantes.

Similitudes:

- Son repositorios centralizados de datos.
- Pueden almacenar datos estructurados.
- Pueden gestionar datos estructurados en tiempo real.

Diferencias:

- Un DL almacena datos no estructurados.
- Un DL no transforma los datos, sino que los almacena tal cual le llegan.
- Un DL no almacena los datos con una estructura óptima para su consulta (OLAP).
- Un DL tiene capacidad (de procesamiento y de almacenamiento) para gestionar datos no estructurados en tiempo real.

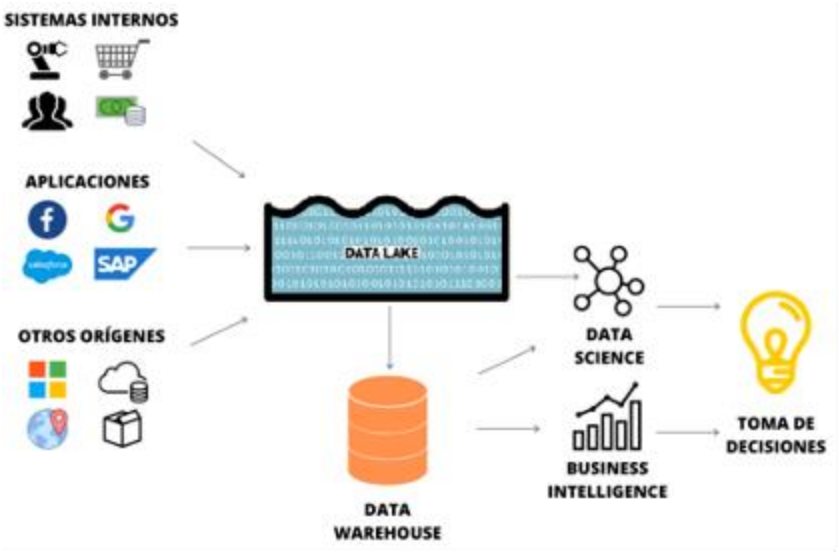
¿Data Lake o Data Warehouse?

Por un lado, un DWH sirve para el almacenamiento y análisis de datos estructurados.

Por otro lado, un DL sirve para el almacenamiento de datos estructurados y no estructurados, aunque no tanto para su análisis.

Lo normal es que queramos analizar los datos, lo que significa que siempre vamos a necesitar un DWH.

De esto puede deducirse que **podemos tener un DWH sin un DL, pero rara vez tendremos un DL sin un DWH**.



Por tanto, podríamos responder a la pregunta diciendo que:

Si disponemos solamente de datos estructurados, entonces usaremos un data warehouse.

Si disponemos tanto de datos estructurados como no estructurados, entonces:

- Usaremos un data lake para la captura de todo tipo de datos.
- Los datos estructurados los procesaremos y organizaremos adecuadamente para su análisis en un DWH, siendo el data lake la fuente de datos del DWH.
- Los datos no estructurados se quedarán en el data lake, donde serán accedidos por los procesos que deban analizarlos.