Informática y comunicaciones

ies azarquiel

Curso especialización en inteligencia artificial y big data

Introducción Sistema Big Data

Profesor: Sebastián Rubio Valero

Septiembre 2025

1 UNIDADES DE DATOS

La ciencia de datos es un término general que incluye conceptos cómo big data, inteligencia artificial, minería de datos, aprendizaje máquina, aprendizaje profundo, etc. Todos estos términos los iremos desgranando durante el curso.
En la siguiente tabla se dan las diferentes unidades de medida de la información dadoque en ambientes de computacionales se emplea constantemente la numeración

byte (B)	8 bits
kilobyte (kB)	1000 bytes (10^3 bytes)
megabyte (MB)	1000 kilobytes (10^6 bytes)
gigabyte (GB)	1000 megabytes (10^9 bytes)
terabyte (TB)	1000 gigabytes (10^{12} bytes)
petabyte (PB)	1000 terabytes (10^{15} bytes)
exabyte (EB)	1000 petabytes (10^{18} bytes)
zettabyte (ZB)	1000 exabytes (10^{21} bytes)
yottabyte (YB)	1000 zettabytes (10^{24} bytes)

Table 1: Byte and Byte Multiples

en base 2 también existe el kibibyte (KiB), el cual corresponde a 1024 bytes (2 10). De igual modo,también existe el mebibyte (MiB = 220 bytes), el gibibyte (GiB = 230 bytes), y así toda la progresión hasta llegar al yobibyte (YiB = 280bytes).

Para hacernos una idea del volumen de datos que maneja la humanidad, según laspredicciones el volumen de datos en el mundo se calculaba en unos 4.4 zettabytes en 2013,y tiene un crecimiento exponencial según el cual se espera que pueda llegar a los 163 zettabytes para el año 2025. Toda esta información viene de diferentes fuentes:

- Datos de usuarios y/o clientes de instituciones y empresas.
- Datos generados por transacciones (compras, transferencias, ...).
- Datos adquiridos por sensores (de temperatura, de humedad, ...).
- Datos subidos a redes sociales (textos, imágenes, vídeos, ...).
- Datos relacionados con la salud (historiales y pruebas realizadas a pacientes).
- Datos de geolocalización (posicionamiento en cada momento según GPS).
- Datos guardados en logs (de todos los accesos que hacemos a páginas web).

- Datos producidos por el Internet de las cosas (de los diversos dispositivos IoT).
- Datos producidos por la genómica (cada vez que se secuencia un genoma).
- Datos de meteorología (información obtenida por satélites y las predicciones realizadasa partir de la misma).
- Datos producidos por cámaras (imágenes estáticas y vídeos producidos).
- Datos producidos por micrófonos (grabaciones de sonido producidas).
- Datos de RFID (aquellos con los que se tratan al realizar identificación porradiofrecuencia).

2 DATOS, INFORMACIÓN Y CONOCIMIENTO

En muchas ocasiones se habla indistintamente de datos e información, pero en el Big Data son coceptos, que aunque este relacionados, no tienen el mismo significado.

El **dato** es una representación sintáctica, generalmente numérica, que puede mane-

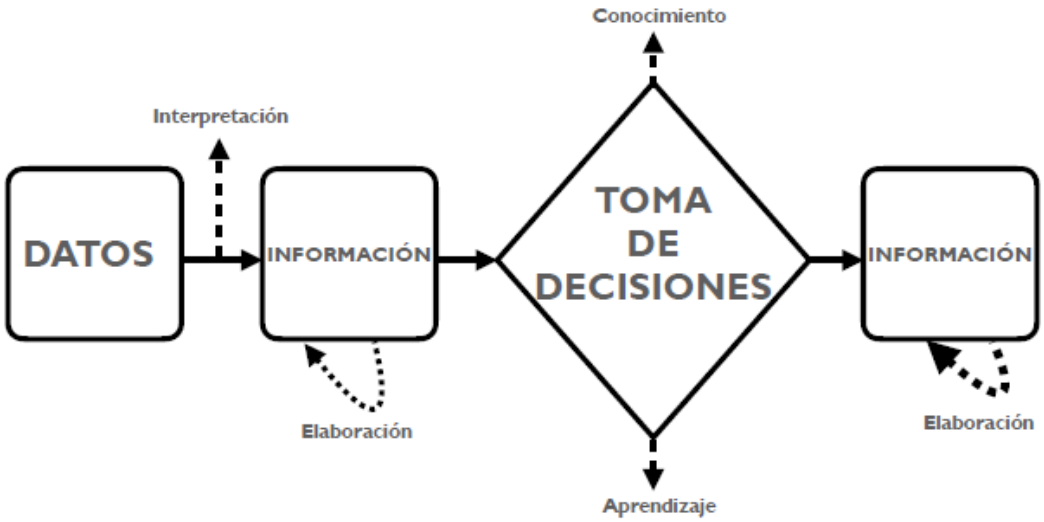

Figure 1: Enter Caption

jar un dispositivo electrónico - normalmente un ordenador - sin significado por sí solo. Sin embargo, el dato es a su vez el ingrediente fundamental y el elemento de entrada necesario en cualquier sistema y/o proceso que pretenda extraer información o conocimiento sobre un dominio determinado. En este sentido, 7 es un dato, como

también lo es π o como son los términos aprobado o suspenso.

Por su parte, la **información** es el dato interpretado, es decir, el dato con significado. Para obtener información, ha sido necesario un proceso en el que, a partir de un dato como elemento de entrada, se realice una interpretación de ese dato que permita obtener su significado, es decir, información a partir de él. La información es también el elemento de entrada y de salida en cualquier proceso de toma de decisiones. Partiendo de los datos del ejemplo anterior, información obtenida a partir de los mismos puede ser: El 7 es un número primo, π es una constante cuyo valor es 3; 141592653..., María ha aprobado el examen de conducir, Pablo está suspenso en matemáticas.

A partir de información, es posible construir **conocimiento**. El conocimiento es información aprendida, que se traduce a su vez en reglas, asociaciones, algoritmos, etc. que permiten resolver el proceso de toma de decisiones. Así pues, la información obtenida a partir de los datos permite generar conocimiento, es decir, aprender. El conocimiento no es estático, como tampoco lo es siempre el aprendizaje. Aprender, construir conocimiento, implica necesariamente contrastar y validar el conocimiento construido con nueva información que permita, a su vez, guiar el aprendizaje y construir conocimiento nuevo. Siguiendo con los ejemplos anteriores, el conocimiento que permite obtener que el 7 es un número primo puede ser el algoritmo de Eratóstenes.Por otra parte, el conocimiento que permite obtener el valor del número π puede extraerse de los resultados de los trabajos de Jones,Euler o Arquímedes, mientras que el aprobado de María en el examen de conducir y el suspenso de Pablo en matemáticas, se pueden obtener de la regla que en una escala de diez asigna el aprobado a notas mayores o iguales que 5 y el suspenso a notas menores.

Por tanto, datos, información y conocimiento están estrechamente relacionados entre sí y dirigen cualquier proceso de toma de decisiones (refe UCLM). En el siguiente ejemplo aunque trivial se puede ver la diferencia entre los tres conceptos.

1. El profesor corrige el examen de Pablo, que ha sacado un 3. Esta calificación, por sí sola, es simplemente un dato.

2. A continuación, el profesor calcula la calificación final de Pablo, en base a la nota del examen, sus trabajos y prácticas de laboratorio.*La nota final de Pablo es un 4* . Esto último es información.

3. ¿Ha aprobado Pablo? La información de entrada al proceso de decisión es su calificación final de 4 puntos, obtenida en el paso anterior. El conocimiento del profesor sobre el sistema de calificación le indica que una nota menor a 5 puntos se corresponde con un suspenso y, en caso contrario, con un aprobado.

4. La información de salida tras este proceso de decisión es que *Pablo está suspenso en matemáticas.*

3 CLASIFICACIÓN DE LOS TIPOS DE DATOS

3.1 En cuanto al tipo

Esta primera clasificación es con respecto al punto de via del tipo de operaciones que pueden hacer los ordenador sobre los dato:

- **Tipos simples**. También se le conoce como tipos primitivos, representan a un único valor
- **Tipos compuestos**. Son el resultado de la combinación de los tipos simples aparecen tipos compuestos, que representan un conjunto de valores a modo de estructura. Dentro de estos tipos nos podemos encontrar vectores, matrices, listas, conjuntos, registros, etc. Con tipos compuestos podemos representar el nombre de un producto, la imagen de la matrícula de un coche, en forma de matriz de bits, el audio de la transcripción de una conversación, o entes más complejos, como un coche o una persona.

3.2 Datos según el formato

Los datos también se pueden clasificar desde el punto de vista de como están organizados:

- **Datos estructurados**. Una colección de datos está estructurada cuando presenta un modelo o esquema organizativo. Es decir, todos los elementos de la colección responden a una misma organización, tanto en cuanto a tipos como a significado. Por jemplo los datos organizados según un esquema relacional, donde tenemos tablas y relaciones, a su vez cada tabla consta de registros compuestos de atributos.
- **Datos semiestructurados**. Son aquellos que no son estructurados, pero que presentan cierta organización. El primer rasgo con el que se puede identificar es que no tiene una estructura tabula, en forma de tabla de registros. Ofrecen mayor flexibilidad a la hora de definir la organización. Los formatos **XML** y **JSON** son ejemplos.
- **Datos no estructurados**. Son datos que carecen de estructura, documentos, mensajes, vídeo e imágenes entre otros. Estos datos al igual que los semiestructurados se suelen almacenar en **bases de datos NoSQL**

3.3 Datos en cuanto quién los genera

En esta otra clasificación se distinguen dos grandes grupos:

- **Datos generados por personas**
- **Datos generados por máquinas**

4 SOLUCIONES BIG DATA

En esta nueva era tecnológica en la que nos hayamos inmersos, a diario se generan enormes cantidades de datos, del orden de petabytes (más de un millón de gigabytes). Hoy en día, cualquier dispositivo como puede ser un reloj, un coche, un smartphone, etc está conectado a Internet generando, enviando y recibiendo una gran cantidad de datos. Tanto es así, que se estima que el 90en el mundo ha sido generado en los últimos años. Sin lugar a dudas, esta y las próximas generaciones serán las generaciones del **big data**

4.1 Modelo de las cinco uves

Esta realidad descrita anteriormente demanda la capacidad de enviar y recibir datos e información a gran velocidad, así como la capacidad de almacenar tal cantidad de datos y procesarlos en tiempo real. Así pues, la gran cantidad de datos disponibles junto con las herramientas, tanto hardware como software, que existen a disposición para analizarlos se conoce como **big data**. A día de hoy no hay una definición precisa de big data, pero si existe consenso en cuanto a las propiedades que deben cumplir, esto el modelo de **5 uves**

- **Volumen** En un sistema big data se generan grandes cantidades de datos, del orden de petabyte y exabyte. Esto hoy en día es bastante común en el comercio electrónico y resdes sociales
- **Velocidad**. Esta propiedad hace referencia a dos aspectos, el primero a la velocidad en la que se generan los datos, y la segunda a la velocidad en la que los datos son procesado.
- **Variedad**. Entre el 80-90% de los datos son no estructurados, correos electrónicos, mensajes, manuales, autios, videos, imágenes,etc.
- **Veracidad**. Para obtener los resultados esperados al aplicar tecnologías de big data, es muy importante que se pueda confiar en los datos, es decir que los

Figure 2: Enter Caption

datos representan exactamente la información que se espera. El problema de la calidad de los datos pasa por un preprocesamiento antes de que estos formen parte del entrenamiento modelo.

- **Valor**. El objetivo principal de los datos es convertirlos en información que nos permite desarrollar conocimiento

4.2 Arquitectura de un sistema Big Data

Supongamos que en en una primera fase se ha adquirido un conocimiento previo del caso de estudio y se han establecido unos objetivos para el proyecto (Big Data). Este análisis inicial marcará los puntos desde donde se parte (datos) a donde se quiere llegar (información). El proceso mostrado en la Figura se establece las etapas a seguir para extraer la información a partir de los datos. Este proceso consta de las siguientes etapas: (apuntes uclm)

- **Identificación de datos:** En esta fase se identifican todas las fuentes generadoras de datos que están disponibles y son de utilidad para el objetivo del
- **Preparación de datos:** En la primera etapa de preparación o pre-procesamiento los datos se almacenan en crudo. Esto permite crear repositorios con grandes volúmenes de información sin invertir mucho tiempo en la definición de su estructura e implementación. El objetivo de esta etapa es mejorar la calidad

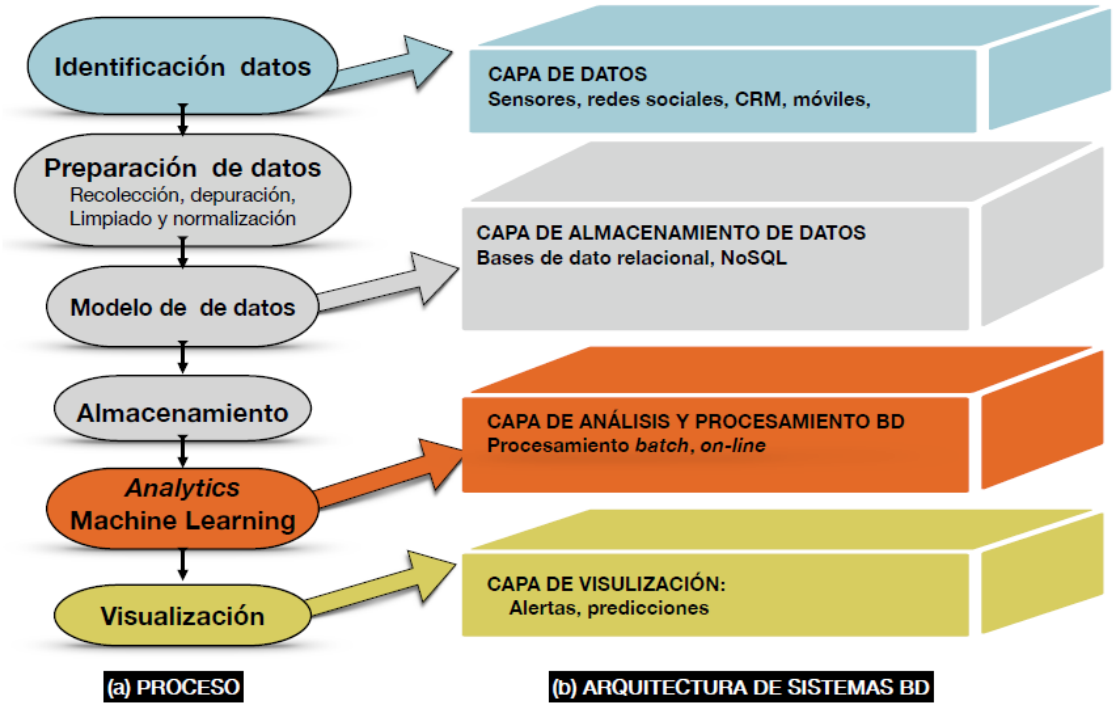

Figure 3: Enter Caption

de los datos. Se establecen modelos predictivos para los datos poco fiables, defectuosos, ausentes o desajustados en cuanto a atribuciones válidas.

- **Modelo de datos:** En esta etapa se define un modelo de datos con el objetivo de organizar, clasificar y especificar como se relacionan los datos existentes. También se establecen las condiciones que deben cumplir los datos para reflejar la realidad modelada y considerarlos válidos. También se definen las transformaciones necesarias que se deben realizar para poder ser utilizados. En la fase de almacenamiento se procede a la implementación del modelo en una base de datos.

- **Proceso analítico:** En esta etapa mediante técnicas de aprendizaje automático se buscan patrones de interés, determinación de tendencias, identificación de anomalías, etc.

- **Visualización:** En esta fase final se entregan los resultados obtenidos para su interpretación y documentación. En este punto, los datos finales han sido

limpiados, convertidos, seleccionados en base a atributos relevantes y enmarcados en representaciones visuales para ayudar a los humanos en su toma de decisiones.

En el proceso Análisis Machine Learning existen diferentes etapas:

- Análisis descriptivo, proporciona información sobre en rendimiento pasado del negocio y su contexto
- Análisis prescriptivo, tiene un enfoque más operativo con la idea de buscar soluciones
- Análisis predictivo, se basa en descubrir patrones, tendencia y relaciones que permiten explicar el comportamiento a partir de datos históricos con el fin de anticiparse al futuro.
- Análisis congnitivo, saber que esta sucediendo y por que

5 ARQUITECTURA DE ALMACENAMIENTO

5.1 Ejemplo de uso de un Data Hub

Puesto que no tenemos contacto director con estos sistemas de almacenamiento, los vamos estudiar utilizando ejemplos, de esta forma podremos tener una idea aproximada de cada uno de llos Sistema A: El inventario de la tienda física. Sistema B: La plataforma de ventas en línea (e-commerce). Sistema C: El sistema de gestión de relaciones con el cliente (CRM) que guarda los datos de contacto y el historial de compras. Sistema D: La aplicación móvil para los clientes. El problema; antes de tener un Data Hub, cada vez que un departamento necesitaba información, tenía que solicitarla directamente al sistema correspondiente. Por ejemplo:

El equipo de marketing quería lanzar una promoción basada en el historial de compras de los clientes, pero los datos de compra estaban en el Sistema B, mientras que la información de contacto estaba en el Sistema C. Tuvieron que pedirle al equipo de TI que exportara y uniera los datos de forma manual, un proceso que podía tardar días y a menudo resultaba en errores.

El equipo de servicio al cliente recibía una llamada de un cliente, pero no podía ver su historial de compras en línea (del Sistema B) desde el sistema de atención telefónica (Sistema A), lo que les impedía dar una respuesta rápida y personalizada.

La solución: Implementar un Data Hub "Tienda del Futuro" decide crear un Data Hub para centralizar y gobernar sus datos. Este hub se convierte en el "único punto de verdad" para la información de la empresa.

Integración: El Data Hub se conecta a los Sistemas A, B, C y D. Los datos se extraen de forma continua (casi en tiempo real) y se envían al hub.

Calidad y Gobernanza: Dentro del Data Hub, se aplican reglas para limpiar, estandarizar y validar los datos. Por ejemplo, se eliminan los registros duplicados de clientes y se asegura que el formato de los correos electrónicos sea el mismo en todas partes.

Distribución: El Data Hub pone los datos unificados a disposición de todos los sistemas y usuarios a través de APIs y otros servicios.

El resultado Con el Data Hub en funcionamiento, los problemas anteriores se resuelven fácilmente:

Para el equipo de marketing: Ahora pueden acceder directamente al Data Hub y obtener un conjunto de datos limpio que ya tiene la información de contacto y el historial de compras unificado. Pueden crear una campaña en minutos, dirigiendo la promoción a los clientes que compraron un producto específico en la tienda en línea y ofreciendo un descuento en su próxima compra en la tienda física.

Para el servicio al cliente: Cuando un cliente llama, el representante puede consultar el Data Hub para obtener una vista completa y unificada del cliente, incluyendo sus compras en línea y en la tienda física, lo que permite ofrecer una atención más rápida y personalizada.

En este ejemplo, el Data Hub actúa como un puente que conecta los diferentes sistemas, elimina la necesidad de procesos manuales y garantiza que todos los equipos trabajen con la misma información, precisa y actualizada.

5.2 Ejemplo de Data WareHouse

Un ejemplo clásico y comprensible de un Data Warehouse es la base de datos que usa una cadena de supermercados para el análisis de sus ventas a nivel nacional.

El problema antes del Data Warehouse Imagina que la cadena tiene cientos de tiendas, cada una con su propio sistema de cajas registradoras que almacena datos transaccionales (quién compró qué, a qué hora, con qué método de pago). Estos sistemas están optimizados para el día a día: procesar pagos rápidos y gestionar el inventario de cada tienda. Sin embargo, si el director ejecutivo quiere saber:

¿Cuál fue el producto más vendido en todas las tiendas de la región sur durante el último trimestre?

¿Qué tiendas tienen el peor rendimiento de ventas los fines de semana?

¿Cómo ha afectado el aumento del precio del arroz a la venta de pasta?

Responder a estas preguntas sería casi imposible con los sistemas transaccionales, ya que no están diseñados para consultas complejas que involucran grandes volúmenes de datos históricos de diferentes fuentes.

La solución: El Data Warehouse Para resolver este problema, la cadena de supermercados implementa un Data Warehouse.

Extracción, Transformación y Carga (ETL): Por la noche, cuando el tráfico de ventas es bajo, los datos de ventas de cada una de las tiendas son extraídos, transformados y cargados en el Data Warehouse.

Extracción: Se toman los datos de las bases de datos transaccionales de cada tienda.

Transformación: Se limpian y estandarizan los datos. Por ejemplo, se unifican los nombres de productos y las categorías (si la "Leche Entera" se llama de forma diferente en dos tiendas, se estandariza a un solo nombre). También se agregan datos de otras fuentes, como información demográfica del cliente o campañas de marketing.

Carga: Los datos limpios y transformados se cargan en la estructura de almacenamiento del Data Warehouse, que está optimizada para el análisis. Un ejemplo clásico y comprensible de un Data Warehouse es la base de datos que usa una cadena de supermercados para el análisis de sus ventas a nivel nacional.

El problema antes del Data Warehouse Imagina que la cadena tiene cientos de tiendas, cada una con su propio sistema de cajas registradoras que almacena datos transaccionales (quién compró qué, a qué hora, con qué método de pago). Estos sistemas están optimizados para el día a día: procesar pagos rápidos y gestionar el inventario de cada tienda. Sin embargo, si el director ejecutivo quiere saber:

¿Cuál fue el producto más vendido en todas las tiendas de la región sur durante el último trimestre?

¿Qué tiendas tienen el peor rendimiento de ventas los fines de semana?

¿Cómo ha afectado el aumento del precio del arroz a la venta de pasta?

Responder a estas preguntas sería casi imposible con los sistemas transaccionales, ya que no están diseñados para consultas complejas que involucran grandes volúmenes de datos históricos de diferentes fuentes.

La solución: El Data Warehouse Para resolver este problema, la cadena de supermercados implementa un Data Warehouse.

Extracción, Transformación y Carga (ETL): Por la noche, cuando el tráfico de ventas es bajo, los datos de ventas de cada una de las tiendas son extraídos, transformados y cargados en el Data Warehouse.

Extracción: Se toman los datos de las bases de datos transaccionales de cada tienda.

Transformación: Se limpian y estandarizan los datos. Por ejemplo, se unifican los nombres de productos y las categorías (si la "Leche Entera" se llama de forma diferente en dos tiendas, se estandariza a un solo nombre). También se agregan datos de otras fuentes, como información demográfica del cliente o campañas de marketing.

Carga: Los datos limpios y transformados se cargan en la estructura de almacenamiento del Data Warehouse, que está optimizada para el análisis.

Con licencia de Google Análisis: Una vez que los datos están en el Data Warehouse, los analistas de negocio y los directivos pueden usar herramientas de inteligencia de negocio (BI) para ejecutar consultas complejas y generar informes rápidamente. Las preguntas que antes tomaban días, ahora pueden ser respondidas en minutos.

"¿Cuál fue el producto más vendido?": El sistema puede analizar las ventas de todos los productos en todas las tiendas en un instante.

"¿Qué tiendas tienen peor rendimiento?": Se pueden comparar los datos de ventas por ubicación, día y hora.

Características del Data Warehouse en este ejemplo Orientado a un tema: El Data Warehouse está enfocado en un tema central (ventas) y agrupa los datos de diferentes fuentes relevantes (inventario, promociones, clientes).

Histórico: Los datos no se eliminan, sino que se acumulan con el tiempo, lo que permite a la empresa analizar tendencias a largo plazo.

No volátil: Una vez que los datos se cargan, no se cambian ni se eliminan. Esto garantiza que las consultas de hoy y de mañana se basen en la misma información histórica.

En resumen, el Data Warehouse es el cerebro analítico de la cadena de supermercados. Recopila datos de todas las operaciones diarias y los organiza de tal manera que permite a los líderes tomar decisiones estratégicas basadas en información consolidada y de alta calidad.

5.3 SISTEMA OLTP

Un sistema OLTP (Online Transaction Processing o Procesamiento de Transacciones en Línea) es un sistema de gestión de bases de datos diseñado para manejar un gran número de transacciones cortas y en tiempo real. El objetivo principal de estos sistemas es facilitar y automatizar las operaciones diarias de una empresa, como las ventas, los depósitos bancarios, las reservas de billetes y el control de inventario.

Características clave Transacciones cortas y rápidas: Las operaciones son pequeñas, como insertar, actualizar o eliminar registros de forma individual. Esto asegura que el tiempo de respuesta sea casi instantáneo para el usuario.

Alta concurrencia: Estos sistemas están optimizados para que miles de usuarios puedan realizar transacciones simultáneamente, sin interferir entre sí.

Integridad de los datos: Los sistemas OLTP siguen las propiedades ACID (Atomicidad, Consistencia, Aislamiento y Durabilidad) para garantizar que las transacciones sean confiables y que los datos permanezcan precisos.

Datos actuales y detallados: Los datos se almacenan de forma altamente detallada y reflejan el estado actual de la empresa.

Ejemplos comunes Cajeros automáticos y banca en línea: Cada retiro, depósito o transferencia es una transacción OLTP.

Sistemas de punto de venta (POS): Cuando escaneas un producto en un supermercado, se ejecuta una transacción OLTP para actualizar el inventario y registrar la venta.

Plataformas de comercio electrónico: Las compras en línea, las actualizaciones de inventario y el seguimiento de pedidos son procesados por sistemas OLTP.

Sistemas de reserva: Las reservas de vuelos, hoteles o entradas de cine son ejemplos de transacciones en tiempo real.

5.4 Base de datos de columnas

Las bases de datos de columnas (también conocidas como bases de datos orientadas a columnas o "columnar databases") son un tipo de sistema de gestión de bases de datos que almacenan los datos de una tabla por columnas en lugar de por filas.

A diferencia de las bases de datos relacionales tradicionales, que guardan la información de cada registro de forma secuencial (fila a fila), una base de datos de columnas agrupa todos los valores de una misma columna.

¿Cómo funcionan? Imagina una tabla simple de clientes con columnas para "ID", "Nombre" y "Ciudad".

En una base de datos tradicional (orientada a filas), los datos se almacenan de esta manera:

Fila 1: (1, Juan, Madrid)

Fila 2: (2, María, Barcelona)

Fila 3: (3, Pedro, Madrid)

Se lee y se escribe la fila completa.

En una base de datos de columnas, los datos se almacenan de esta otra manera:

Columna ID: (1, 2, 3)

Columna Nombre: (Juan, María, Pedro)

Columna Ciudad: (Madrid, Barcelona, Madrid)

Se lee y se escribe la columna completa.

Este enfoque es especialmente eficiente para cargas de trabajo analíticas (OLAP) y Big Data, donde las consultas suelen implicar la agregación de datos de unas pocas columnas sobre un gran número de filas. Al almacenar los datos por columnas, la base de datos solo necesita leer la información de las columnas relevantes para la consulta, lo que acelera significativamente el rendimiento.

Ejemplos actuales Actualmente, el mercado cuenta con varias bases de datos de columnas, tanto de código abierto como comerciales. Algunos ejemplos notables son:

Amazon Redshift: Un almacén de datos en la nube optimizado para análisis a gran escala.

Google BigQuery: Un servicio de análisis de datos a gran escala.

Apache Cassandra: Una base de datos NoSQL de columnas anchas, diseñada para gestionar grandes cantidades de datos distribuidos.

Apache HBase: Una base de datos NoSQL de código abierto modelada a partir de Google Bigtable.

ClickHouse: Una base de datos de gestión de columnas para el procesamiento analítico en línea (OLAP).

Snowflake: Una base de datos en la nube que separa el almacenamiento del cómputo.

Pregunta.

Siguiendo el ejemplo anterior ¿Cómo se produce el proceso de toma de decisiones para determinar si un número es primo?

5.5 Diferencia entre base de datos y almacen de datos

La diferencia principal entre una **base de datos** y un **almacén de datos** es su propósito y uso. Una base de datos está diseñada para el **procesamiento de transacciones en tiempo real** (OLTP), mientras que un almacén de datos está optimizado para el **análisis de grandes volúmenes de datos históricos** y la toma de decisiones empresariales (OLAP).

—

5.6 Base de Datos (Database)

Una base de datos es un sistema para almacenar y organizar datos de manera estructurada, a menudo para alimentar aplicaciones y sistemas operativos diarios. Su objetivo principal es garantizar la integridad y la velocidad en la entrada y actualización de datos. Piense en ella como la caja registradora de una tienda, donde cada venta se registra de manera rápida y precisa.

- Propósito: Procesamiento de transacciones en línea (OLTP).
- Datos: Almacena datos actuales y operativos.
- Volatilidad: Los datos son volátiles; se pueden leer, escribir, modificar y eliminar.
- Estructura:Típicamente sigue un modelo relacional normalizado para evitar la redundancia y garantizar la consistencia.
- Uso: Sistemas de gestión de clientes (CRM), sistemas de inventario, páginas web de comercio electrónico.

5.7 Data Lakes

Imagina un lago natural muy grande. A él fluyen ríos y arroyos de todo tipo: algunos con agua cristalina (datos estructurados), otros con lodo y sedimentos (datos semiestructurados), y otros que traen hojas, ramas y elementos varios (datos no estructurados). El lago lo almacena todo en su estado natural, sin procesar, para que luego puedas ir a pescar exactamente lo que necesites.

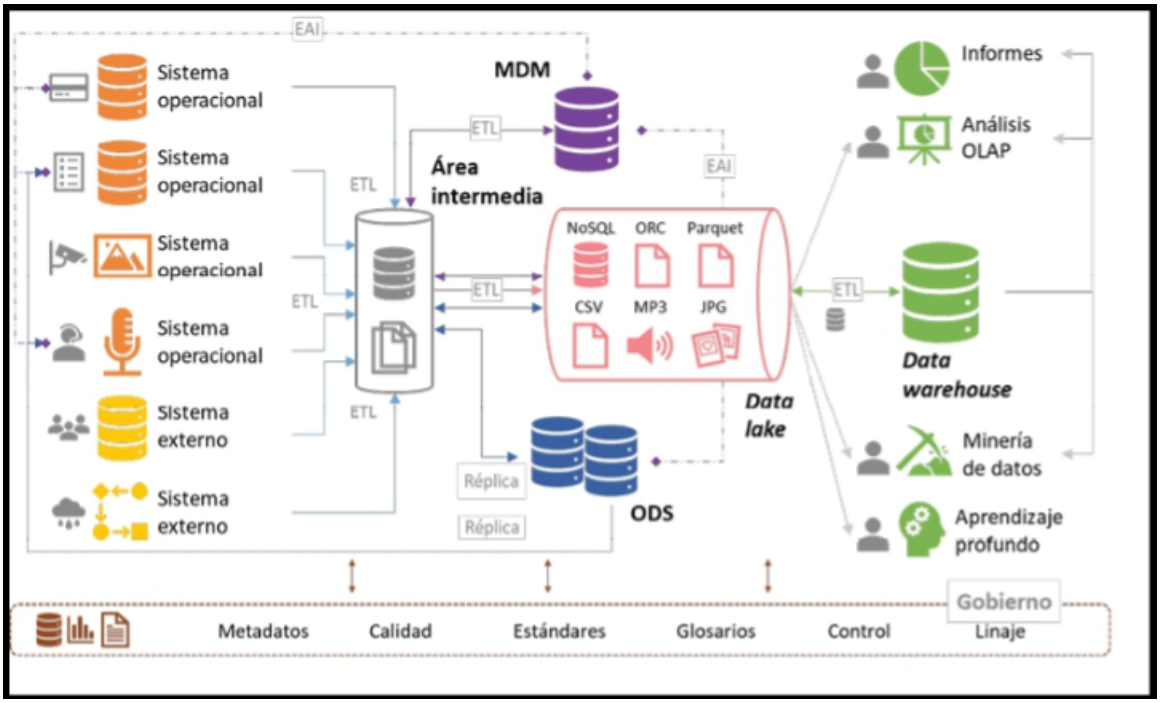

Figure 4: Enter Caption

Técnicamente, un Data Lake es un repositorio de almacenamiento centralizado que permite guardar todos tus datos, estructurados y no estructurados, a cualquier escala, en su formato original. A diferencia de los almacenes tradicionales (como los Data Warehouses), no necesitas definir la estructura, el esquema o el propósito de los datos antes de cargarlos. Esto te ofrece una flexibilidad enorme para el análisis futuro.

Características clave:

- Almacena todo tipo de datos: Desde bases de datos relacionales (datos estructurados), hasta archivos CSV, JSON, XML (semiestructurados), emails, documentos PDF, imágenes, videos, archivos de audio, logs de servidores, y datos de sensores IoT (no estructurados).

- Schema-on-Read (Esquema al leer): Esta es la diferencia fundamental. En un Data Warehouse tradicional se aplica un "Schema-on-Write" (esquema al escribir), lo que significa que debes modelar y transformar los datos antes de ingresarlos. En un Data Lake, los datos se cargan en crudo y el esquema se aplica después, solo cuando se leen para su análisis. Esto acelera enormemente la ingesta de datos.

- Alta escalabilidad y bajo coste: Se suele construir sobre tecnologías de almacenamiento en la nube (como AWS S3, Azure Data Lake Storage, Google Cloud Storage) o usando soluciones como Hadoop (HDFS). Estos sistemas están diseñados para escalar de forma masiva y a un coste por gigabyte muy bajo.

- Múltiples motores de análisis: Los datos en el lago pueden ser procesados por una variedad de herramientas: SQL, Python, R, machine learning, herramientas de BI, etc. Diferentes equipos (científicos de datos, analistas, ingenieros) pueden acceder a los mismos datos con sus herramientas preferidas.

5.8 Ejemplo de un Data Lake

Imaginemos una empresa de streaming llamada "StreamFlix".

El Desafío: StreamFlix necesita entender el comportamiento de sus usuarios para mejorar las recomendaciones, producir contenido exitoso y evitar la cancelación de suscripciones. Los datos provienen de muchas fuentes diferentes y en diversos formatos.

Fuentes de Datos (Los "Ríos" que fluyen al Lago):

- Datos Estructurados:
  - Base de datos de suscripciones: Información del usuario (ID, plan de pago, fecha de alta).
  - Catálogo de contenido: Metadatos de películas y series (ID, título, género, actores, director).
- Datos Semiestructurados
  - Archivos JSON de logs de clics: Cada interacción del usuario con la app (qué hizo clic, en qué minuto pausó, qué buscó).
  - Datos de la app móvil: Eventos de rendimiento de la aplicación (tiempos de carga, errores).
- Datos No Estructurados

- Archivos de video: El contenido multimedia en sí (las películas y series).
- Miniaturas de video: Las imágenes que se muestran para cada título.
- Comentarios de usuarios en texto libre: Reseñas escritas por los suscriptores.
- Registros de servidores (logs): Texto crudo con información técnica de cada petición.

El Data Lake en Acción:

1. Ingesta (Captura): Todos estos datos se vierten continuamente y en tiempo real en el Data Lake de StreamFlix, que está alojado en un servicio como Amazon S3. Se almacenan en su formato original, sin ser transformados. Una carpeta podría tener los logs del día, otra las nuevas miniaturas, otra la base de datos de usuarios (exportada diariamente).

2. Almacenamiento: El Data Lake guarda todo de manera segura y a bajo coste. Un ingeniero de datos podría organizarlo en "capas":

   - Capa Raw/Landing: Los datos exactamente como llegaron.
   - Capa Cleaned/Processed: Datos limpios y transformados (pero aún detallados).
   - Capa Curated/Analytics: Datos altamente procesados y listos para que los use un analista.

3. Análisis y Explotación (El "Pescar" en el Lago): Diferentes equipos acceden al lago para sus necesidades:

   - Equipo de Ciencia de Datos: Utiliza herramientas como Spark o Python para procesar los logs de clics y los comentarios de los usuarios. Aplican modelos de machine learning sobre los datos en crudo para crear un nuevo algoritmo de recomendación que prediga qué te gustará ver después.
   - Equipo de Analytics/Business Intelligence: Usa motores de consulta como AWS Athena o Presto (que permiten hacer SQL directamente sobre los archivos del lago) para generar un reporte semanal: "¿Cuál es la serie más vista en España en los últimos 7 días?".
   - Equipo de Marketing: Combina los datos de suscripción (estructurados) con las reseñas de usuarios (no estructurados) para detectar usuarios insatisfechos y lanzarles una campaña de email marketing con ofertas personalizadas.

- Equipo de Desarrollo: Analiza los logs de servidores no estructurados para identificar y resolver cuellos de botella en el rendimiento de la plataforma.

El Data Lake permite a StreamFlix romper los silos de información. En lugar de tener los datos en compartimentos separados y desconectados, todo converge en un único lugar. Esto permite realizar análisis transversales que serían imposibles o muy costosos en una arquitectura tradicional, impulsando la innovación y la toma de decisiones basada en datos.

6 BIBLIOGRAFIA

Sistemas Big Data Editorial Rama Apuntes de la Universidad de Castilla-La Mancha
Consultas realizadas DeepSeek y ChatGpt