## 1. Tratamiento de datos

Se recopilan datos y se traducen en información usable

Pasos para este proceso;

- Descubrimiento de datos → es un proyecto en I+D.
- Adquisición o cobro
- Preparación
- Preprocesamiento
- Integración → proyecto de operación.
- Explotación (análisis, informe y visualización, acción)

  **Mugging de datos** → proceso inicial de refinar los datos para adecuarlos al usuario. Explotación, enriquecimiento, transformación e integración y validación de datos.

  - ☐ Front end → interfaz de usuario.
  - ☐ Back end → servidor, aplicación, base de datos…

  **Análisis - minería de datos:** proceso de predecir datos (resultados) a partir de grandes conjuntos de datos.

  **Informe y visualización:** diseñar la salida para que el usuario lo entienda y lo pueda usar.

## 2. Modelos de datos y tipos de datos.

Distinguimos:

- operaciones
- construcciones
- estructuras

Tipos de datos:

- Estructurado (20%) → longitud y formato definido (nº, fechas y grupos de palabras y números)

Modelo relacional → tecnología clave en la actualidad. Estructura y organiza grandes cantidades de datos de forma eficiente y efectiva. Usa SQL como lenguaje de consulta estructurado. Permite una gran flexibilidad en la gestión de datos y en la realización de consultas.

- Semiestructurado → datos con características consistentes y definidas. No se limita a una base rígida como los relacionales.
- Sin estructurar (80%) → no siguen un formato específico.

## 3. Gestión de datos

Recopila, mantiene y usa datos de forma segura, eficiente y rentable

Objetivo → ayudar a las personas/organizaciones a optimizar el uso de los datos y sacar el mayor provecho de ellos. También consiste en dar respuesta a los problemas que parecen hacer operativo un determinado proyecto de datos.

### Data management: Data storage

Persistencia → servicio más importantes que proporcionan las bases de datos operativos. Garantiza que los datos almacenados no serán modificados sin autorización y que estarán disponibles para el negocio.

Bases de datos relacionales → 1 o más relaciones representadas en tablas donde se almacenan los datos. Está formada por columnas y filas (la clave principal es la primera columna). → Esquema de base de datos.

SQL ha evolucionado en sintonía con la tecnología RDBMS.

CRUD: Crear, Recuperar, actualizar y eliminar son operaciones comunes y relacionadas que puede usar directamente una base de datos o a través de APIs.

NoSQL no se basa en el modelo tabla/clave de los RDBMS. Sus características son:

- Escalabilidad: capacidad de escribir datos en diferentes almacenes a la vez y sin limitaciones.
- Modelos de datos y consultas: usan marcos especializados para almacenar datos
- Diseño de persistencia: elemento más crítico de los NoSQL.
- Diversidad de interfaces
- Consistencia eventual: utilizan BASE en vez de ACID (RDBMS)

El almacenamiento de datos distribuidos consiste en una red donde los datos o información se almacenan en un nodo o computadora.

Las bases de datos distribuidas son las que recuperan rápidamente datos en muchos nodos.

Los almacenes de datos distribuidos tienen mayor disponibilidad y facilidad de acceso a escritura y lectura.

## Data management - Ingesta de datos.

Ingesta de datos: proceso de adquisición e importación de datos en un almacén de datos o una base de datos.

Si los datos se ingieren en tiempo real, cada registro se inserta en la base de datos a medida que se emite.

- Data in motion: Analizados a medida que se generan.
- Datos en reposo: recopilamos antes del análisis.
- Data Streaming: datos generados continuamente por miles de fuentes de datos que envían los registros simultáneamente en tamaños pequeños. En los data streaming systems, se computa en tiempo real un elemento de datos a la vez.

## Integración de datos

La integración de datos se ha centrado en el movimiento a través de middleware.

La integración de datos debe identificar:

- Origen de los datos
- Identificar las fuentes.
- Identificar formatos y lenguajes de scripting
- Recopilación de datos
- Datos de muestra

Las fuentes de datos no estructuradas necesitan moverse rápidamente a través de grandes distancias geográficas para su intercambio.

Para integrar datos en entornos de aplicaciones mixtas → obtener datos de un entorno de datos (origen) a otro entorno de datos (destino) (tecnologías de extracción, transformación y carga (ETL) se han usado para lograr esto en entornos de almacenamiento de datos tradicionales).

Las herramientas ETL se usan para transformar los datos en el formato requerido por el almacén de datos. La transformación se realiza en una ubicación intermedia antes de que los datos se carguen en un almacén de datos. ETL nos da la infraestructura subyacente para la integración mediante:

- Extraer: leer datos de bases de datos de origen
- Transformar: convertir en formato en uno que se ajuste
- Cargar: escribir los datos en la base de destino

La transmisión de datos (streaming data) y el procesamiento de eventos complejos son cada vez más importantes. La computación de transmisión está diseñada para manejar un flujo continuo de una gran cantidad de datos no estructurados.

## Recuperación de datos

Proceso de búsqueda, identificación y extracción de datos requeridos de una base de datos. Requieren escribir consultas o comandos de extracción de datos por parte de los usuarios en una base de datos.

Una base de datos está diseñada para hacer que los sistemas transaccionales se ejecuten de manera eficiente. Este tipo de bases de datos (OLTP) está diseñada para manejar transacciones pero no análisis.

Un almacén de datos (datawarehouse) es un tipo de base de datos que integra copias de datos de transacciones de sistemas de origen dispares y los aprovisiona para un uso analístico. Un datawarehouse es del tipo OLAP.

En un datawarehouse los datos se cargan en el almacén después de transformarlos en un formato bien definido y estructurado. Esto se llama esquema en la escritura.

Un data lake es un depósito de almacenamiento masivo con una enorme potencia de procesamiento y capacidad para manejar una gran cantidad de concurrencias, tareas… Un data lake garantiza que todos los datos se almacenen para un uso más adelante: esquema en escritura

Funciona de la siguiente manera: Los datos se cargan desde su fuente almacenada en su formato nativo hasta que se necesita, momento en el que las aplicaciones pueden leer libremente los datos y agregarle estructura.

Los datos se almacenan como un BLOB con identificador único.

**Infraestructura de almacenamiento y recuperación de datos.**
Jerarquía de memoria:

- El registro interno:
- La caché:
- RAM
- Disco duro
- Cinta magnética:

Las tecnologías emergentes de memoria principal no volátil (NVM) ofrecen una densidad de memoria mucho mayor, un costo por bit mucho menor y un consumo de energía en espera que la DRAM.8.

Sin embargo, la escalabilidad es una decisión entre hacer una máquina que hace que un servidor sea más potente versus agregar más máquinas.

El escalado vertical supone un mayor número de procesadores y RAM o el nuevo NVMM en el que cualquier operación funciona mejor con más memoria, pero su mantenimiento puede ser difícil y costoso y por supuesto posee limitaciones de crecimiento.
El escalado horizontal implica agregar más máquinas, posiblemente menos potentes, a una red relativamente más lenta. Las operaciones paralelas posiblemente serán más lentas, pero en la práctica es más probable añadir más máquinas.

## Calidad de los datos.

Calidad de los datos →
- Perfiles de datos.
- Análisis y estandarización de datos.
- Coincidencia de datos y limpieza de datos.

Generalización de perfiles de datos → proporcionan las métricas e informes que los propietarios de información necesitan.
Puede usar la generalización de perfiles de datos para:
- Analizar, clasificar e identificar los datos.

Análisis y estandarización de datos → capacidades de estandarización de datos.
Limpieza de datos → corregir los datos y que sean consistentes.
Coincidencia de datos → identificación de posibles duplicados para registro de cuentas, contactos…

## Seguridad de los datos

Seguridad → datos más sensibles → más seguridad.
Si un sistema de big data se implementa en la nube, necesitamos asegurar:
- Máquinas.
- Transferencia de datos mediante diferentes fases de la operación de datos.

**Técnicas de protección de datos**

- Cifrado
- Anonimización de datos
- Tokenización
- Controles de bases de datos.

## 4. Ejemplos de algunas arquitecturas posibles

La más simple es un sistema con RMDB. Se puede organizar de diferentes maneras;
- **1 nivel** → se colocan todos los componentes necesarios para una app o tecnología de software en un solo servidor o plataforma.
- **2 niveles** → arquitectura del servidor del cliente. Comunicación directa entre el cliente y el servidor. No existe intermediario.
- **3 niveles** → se separan sus niveles entre sí en función de la dificultad de los usuarios y cómo usan los datos presentes en la base de datos. Arquitectura más usada para diseñar un DBMS. Sus tres niveles son:
    - Base de datos
    - Aplicación: servidor de apps y programas que acceden a la base de datos. Media entre el usuario final y la base de datos.
    - Usuario: la aplicación proporciona varias vistas de la base de datos. Estas vistas son generadas por las apps que residen en el nivel de aplicación.

Para cosas más complejas tenemos que coordinar diferentes bases de datos y fuentes → DATAWAREHOUSE. Cuenta también con 3 niveles
- **Nivel inferior:** sistema de base de datos relacional. Se limpian los datos mediante herramientas de back - end.

- **Nivel intermedio:** Es un servidor OLAP, se presenta una vista abstracta de la base de datos. Media entre el usuario final y la base de datos.
- **Nivel superior**: capa de cliente front-end. Herramientas y API que conectan y se obtienen datos del almacén de datos (herramientas de consulta, de informes, de consulta administradas, de análisis y de minería de datos.

Modelo de Gartner:

- **Adquirir**: recoge todo tipo de datos útiles.
- **Organizar**: organiza de extremo a extremo, eso es LDW:

- Proporciona una arquitectura de gestión de datos moderna y escalable bien localizada para satisfacer las necesidades de datos y análisis de la empresa.
- Admite un enfoque de desarrollo que aprovecha la arquitectura y técnicas de almacenamiento de datos empresarial existentes en la organización.
- Establece una capa de acceso a datos compartidos que relaciona lógicamente los datos, independientemente del origen.

- **Análisis de la arquitectura de extremo a extremo** se puede ver dificultada por el aumento de la demanda.