

ALMACENAMIENTO BIG DATA

Almacenamiento Big Data

En Big Data, los sistemas de almacenamiento son fundamentales para gestionar conjuntos de datos a gran escala de forma eficiente. Existen distintos tipos de sistemas, entre los que se incluyen

- **Bases de datos**
- **Sistemas de archivos distribuidos.**


Almacenamiento Big Data

En Big Data, los sistemas de almacenamiento son fundamentales para gestionar conjuntos de datos a gran escala de forma eficiente. Existen distintos tipos de sistemas, entre los que se incluyen

- **Bases de datos**
- **Sistemas de archivos distribuidos.**

BASES DE DATOS

Bases de Datos

1960-1980 

Empezaron con los “MainFrame” que eran sistemas cerrados.

En los 80’s empezaron con DBase

Bases de Datos

1980-1990  dBase  ORACLE  SAP

En los 80's empezaron con dBase como sistema gestor de base de datos doméstico.

También aparece Oracle como y SAP. Entornos profesionales para manejo, gestión y desarrollo de bases de datos corporativas.

Bases de Datos

1990-2000



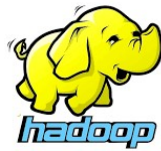
En los **90's** aparece Linux de la mano de con Linus Torvals y Richard Stallman con la revolución del software libre.

Aparecen las bases de datos relacionales como MySQL y PostgreSQL. Y para que no falten ingredientes a la salsa aparece Access de Microsoft.

Bases de Datos

2000-2010

En esta década es cuando aparece el Big Data. Y con este paradigma empazaron a aparecer aplicaciones tales como: Hadoop (2003), AWS, MongoDB, Redis, Hive, Neo4J (grafos), etc.



Bases de Datos

2010-Actualidad

Tenemos herramientas para proceso en Real Time tales como Apache Kafka, Apache Flink, no llegan a ser bases de datos como tal, hace referencia al procesamiento de señal.

Tenemos Amazon Athena, realmente no es una base de datos pero permite hacer consultas SQL sobre ficheros, eso muy interesante, etc.



Tipos Bases de Datos

Bases de Datos Relacionales: **RDBMS**

Las bases de datos relacionales organizan los datos en **tablas** (filas y columnas), donde cada tabla representa una entidad y las relaciones entre entidades se expresan mediante **claves externas**.

Se accede a los datos mediante **SQL** (Structured Query Language).

Tipos Bases de Datos

Nota: En Big Data se tiende a replicar la información, ya que importa más el acceso al dato más que su almacenaje, ya que esto último suele ser más barato.

En la práctica las ETL's, se hacen con JOIN de distintas tablas quedando una tabla mucho más grande.

Tipos Bases de Datos

¿Qué es un **JOIN** en SQL?

Un join en SQL es una operación que combina filas de dos o más tablas relacionadas en una base de datos. Permite recuperar datos de varias tablas y presentarlos en una sola tabla de resultados.

Tipos Bases de Datos

Ejemplo de **INNER JOIN**

Supongamos que tenemos dos tablas: empleados y departamentos. La tabla empleados tiene una columna `id_departamento` que se refiere a la columna `id` de la tabla departamentos.

```
SELECT *  
FROM empleados  
INNER JOIN departamentos  
ON empleados.id_departamento = departamentos.id;
```

Tipos Bases de Datos

Cientes

ID	Nombre	Apellidos	Direccion
1	Alicia	Gonzalez	C/ Con...
2	Bob	Garcia	Plaza ...
3	Carolina	Fernandez	Av. Medi...

Tarjetas

ID	ClientID	Num Tarjeta	Fecha expiracion
1	1	1234781	2023-01-01
2	1	8687913	2024-10-11
3	1	1237811	2025-06-17
4	2	1223898	2026-01-01

Tipos de BBDD: NoSQL

Las bases de datos NoSQL ofrecen modelos de esquema flexibles y no dependen de tablas estructuradas ni de consultas SQL. Están diseñadas para gestionar datos distribuidos a gran escala con distintos formatos.

```
[
  {
    "_id": ObjectId("..."),
    "id_usuario": 1,
    "nombre": "Juan Pérez",
    "email": "juan.perez@example.com",
    "pedidos": [
      {
        "id_pedido": 101,
        "producto": "Laptop",
        "cantidad": 1
      }
    ]
  },
  {
    "_id": ObjectId("..."),
    "id_usuario": 2,
    "nombre": "María Gómez",
    "email": "maria.gomez@example.com",
    "pedidos": [
      {
        "id_pedido": 102,
        "producto": "Smartphone",
        "cantidad": 2
      }
    ]
  }
]
```

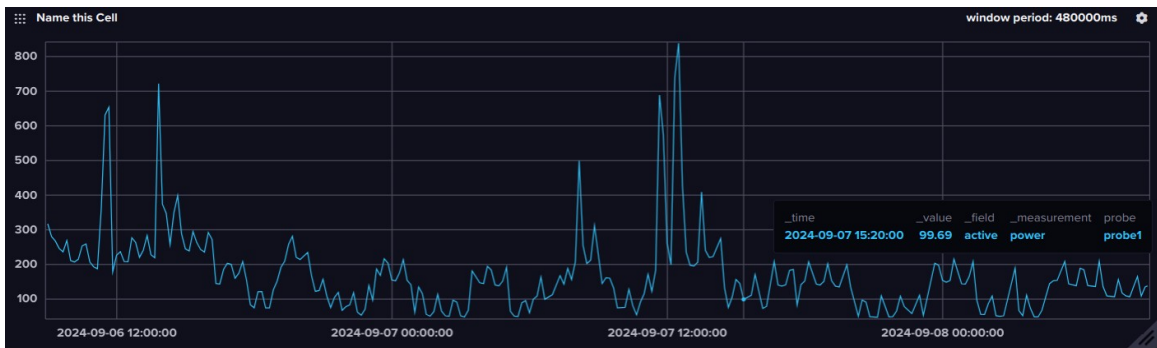
- **Document store:** Almacena documentos formato JSON
 - Redes sociales, Sistemas de gestión de contenidos (CMS)
- **Clave valor:** Los datos se guardan como pares clave-valor
 - Gestión de sesión, cacheo, carrito de la compra
- **Columnar:** Almacena el contenido en columnas en lugar de filas, diseñado para una eficiencia alta de lectura/escritura
- **Grafos:** Se centran en las relaciones entre puntos de datos utilizando estructuras gráficas de nodos y aristas.
 - Redes sociales, detección fraude, sistemas recomendación



Tipos de BBDD: Series Temporales

Diseñadas para manejar datos de series temporales en las que se recopilan puntos de datos a lo largo del tiempo. Estas bases de datos están optimizadas para leer y escribir secuencias de datos, generalmente con una marca de tiempo.

```
from(bucket: "home_climate")
|> range(start: v.timeRangeStart, stop: v.timeRangeStop)
|> filter(fn: (r) => r["_measurement"] == "power")
|> filter(fn: (r) => r["_field"] == "active")
|> filter(fn: (r) => r["probe"] == "probe1")
|> aggregateWindow(every: v.windowPeriod, fn: mean,
createEmpty: false)
|> yield(name: "mean")
```



Usado en:

- IoT
- Datos financieros
- Monitorización de sistemas

Tipos BBDD: De vectores

Una base de datos de vectores es una base de datos que puede almacenar vectores (listas de números de longitud fija) junto con otros elementos de datos.

Las bases de datos vectoriales generalmente implementan el algoritmo ANN o kNN para que uno pueda buscar en la base de datos con un vector de consulta para recuperar los registros de base de datos coincidentes más cercanos.



Chroma

 **Pinecone**



Weaviate

Muy populares con la llegada de RAGs y LLMs



Rey = Reina - Mujer +
Hombre