



# Big Data Aplicado

Curso de especialización en Inteligencia Artificial y Big Data.



# Programación





# Introducción: de los datos al conocimiento

**Dato:** representación sintáctica, generalmente numérica, que puede manejar un dispositivo electrónico - normalmente un ordenador - sin significado por sí solo.

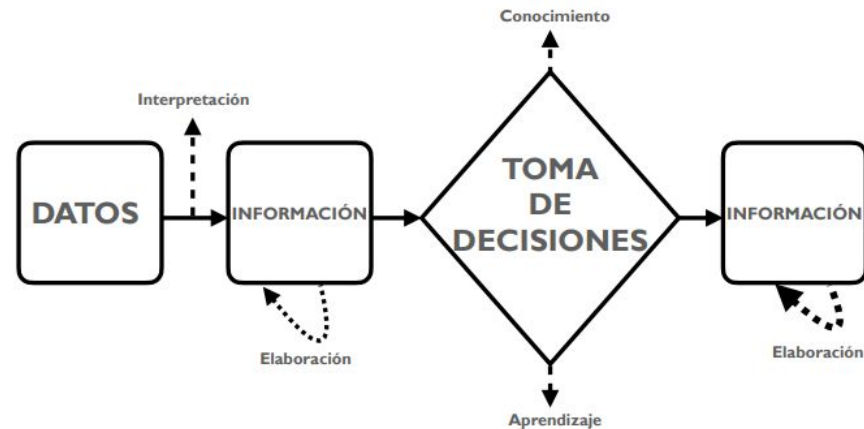
**Información** es el dato interpretado, es decir, el dato con significado.

→ 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

# Introducción: de los datos al conocimiento

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.



**Figura 1:** Relación entre datos, información y conocimiento en el proceso de toma de decisiones.

# Introducción: de los datos al conocimiento



Fuente: Hey, J.: The Data, Information, Knowledge, Wisdom Chaim: The Metaphorical Link



# Introducción: de los datos al conocimiento

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.*



# La carrera entre los datos y la tecnología

**Información:** El gran reto se basa en extraer información a través de los datos para generar conocimiento. Para ello será necesario que los datos y la tecnología deben estar alineados.

Obtener datos no ha sido siempre una tarea fácil. Esto es debido principalmente a que la gran cantidad de sensores disponibles en la actualidad, que permiten registrar magnitudes de cualquier proceso, no existía como a día de hoy.

Antes los procesos que se monitorizaban eran los procesos industriales realizados en grandes empresas. Por todos estos motivos, tradicionalmente se recurría a modelos de simulación que usaban modelos matemáticos, permitían generar datos realistas de un proceso. Los datos generados mediante simulación son conocidos como datos sintéticos mientras que los datos provenientes de las lecturas de un sensor se conocen como datos reales.

[Vídeo gemelos digitales \(digital twins\)](#)



# La carrera entre los datos y la tecnología

Para ello es necesario contar con la tecnología necesaria para su procesamiento. Así pues, el almacenamiento se presenta como el primer problema tecnológico a resolver.

Se plantean soluciones basadas en sistemas de información distribuida. Los sistemas de información distribuida permiten adquirir espacio de almacenamiento en servidores privados, dejando la gestión de estos servidores en manos del proveedor.

El segundo problema tecnológico es el procesamiento de los datos almacenados. Este aspecto cobra especial relevancia en función del caso de aplicación, pudiendo distinguirse entre:

- procesamiento on-line (en línea): los datos son procesados a medida que son generados, ya que se requiere una respuesta en tiempo real.
- procesamiento off-line (fuera de línea): no es necesario que los datos se procesen a medida que se generan.



# La carrera entre los datos y la tecnología

Ejemplo procesamiento online: en un sistema de control del tráfico que permite regular los semáforos en función del tráfico actual, el sistema debe regular el semáforo a medida que se van generando e interpretando los datos del tráfico en un instante de tiempo dado.

Ejemplo procesamiento off-line: en un sistema de detección del fraude bancario, comprobar si un cliente ha realizado algún movimiento fraudulento es una tarea que puede llevarse a cabo off-line, por ejemplo, haciendo un análisis de los movimientos del cliente en un momento dado, sin tener por qué diagnosticar cada movimiento que este va realizando.



# La carrera entre los datos y la tecnología

La computación distribuida, en donde múltiples máquinas realizan el procesamiento optimizando el rendimiento o la computación en la nube, que permite adquirir recursos de procesamiento al igual que se puede adquirir espacio de almacenamiento, son dos soluciones al problema del procesamiento.

Otras alternativas son la programación paralela y la programación multi-procesador, que permiten, respectivamente, aprovechar el paralelismo de múltiples hilos de ejecución dentro de un procesador y realizar el procesamiento dividiéndolo en múltiples hilos en diferentes procesadores.





## Los datos: de ayer y de hoy

La tecnología ha ido evolucionando para dar respuesta a la ingente cantidad de datos que ha comenzado a generarse. Esta evolución, o revolución, no está únicamente relacionada con la cantidad de datos (como se expuso en el anterior apartado) sino también con el tipo y el formato de los mismos.

En el pasado los formatos de archivos que se manejaban solían ser formatos de hojas de cálculo (.xlsx, .ods, .numbers etc) o ficheros separados por comas (.csv). Muy pocos eran los procesos en los que se trabajaba con otros tipos de datos como texto, imágenes, audio e incluso vídeos, ya que los formatos de estos tipos de datos eran limitados hace unos años, su procesamiento más complejo y la tecnología para ello aún en desarrollo.



## Los datos: de ayer y de hoy

- En cuanto al texto, las técnicas de inteligencia artificial y procesamiento del lenguaje natural hacen posible la extracción de conocimiento a partir de grandes volúmenes de textos, que pueden provenir de páginas web, archivos .pdf, redes sociales, etc.
- El desarrollo de hardware con mejores prestaciones y los nuevos modelos de programación permiten procesar en la actualidad grandes cantidades de imágenes, audios y vídeos con una gran variedad de técnicas de inteligencia artificial en tiempos razonables.
- Aparición de nuevos tipos y formatos de datos como los generados a partir de grafos. Estos datos se corresponden, por ejemplo, con datos geográficos obtenidos a partir de mapas como los generados en aplicaciones como Google Maps u Open Street Maps o datos de seguimiento y actividad en redes sociales de gran valor en campañas publicitarias entre otros muchos



## Los datos: de ayer y de hoy

- El desarrollo de bases de datos NoSQL más flexibles y escalables horizontalmente, que permiten almacenar y consultar eficientemente los diversos tipos de datos generados. Por ejemplo MongoDB, Cassandra, etc.
- La computación en la nube, que facilita el procesamiento masivo de datos sin necesidad de infraestructura propia. Servicios como AWS, Azure, GCP proporcionan recursos bajo demanda para ejecutar trabajos de big data.
- La aplicación de técnicas de aprendizaje automático y deep learning sobre conjuntos de datos masivos, permitiendo entrenar modelos más precisos para tareas como clasificación, predicción, reconocimiento de patrones, etc.
- El concepto de data lakes, repositorios centralizados que almacenan y organizan grandes cantidades de datos sin procesar, para su posterior análisis.

# Big Data

A diario se generan enormes cantidades de datos, del orden de petabytes. Se estima que el 90% de los datos disponibles en el mundo ha sido generado en los últimos años.

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.



# Big Data





# Big Data

Actividad:

- Lee el siguiente artículo.

HOW MUCH DATA IS GENERATED EVERY DAY IN 2023? (NEW STATS)

Read more at EarthWeb: How Much Data Is Generated Every Day in 2023? (NEW Stats) <https://earthweb.com/how-much-data-is-created-every-day/>

- Investiga y reflexiona sobre la importancia de los datos en la actualidad, y cómo su creciente generación afecta a la sociedad y a las empresas





# Big Data

Actividad, continuación.

- Según el informe "Data Never Sleeps 9.0", elaborado por Domo, en 2023 se generan más de 2,5 quintillones de bytes de datos cada día. Esto equivale a unos 73 zettabytes de datos al año.
- El crecimiento de los datos se debe a varios factores, entre los que se incluyen:
  - El aumento del uso de dispositivos conectados, como teléfonos inteligentes, tablets y ordenadores.
  - La proliferación de las redes sociales y las aplicaciones de mensajería instantánea.
  - La creciente popularidad de la nube y el almacenamiento en línea.



# Big Data

Actividad, continuación.

Año	Cantidad de datos generados (bytes)	Tráfico de Internet (%)	Redes sociales (terabytes)
2020	2,5 quintillones	90%	500
2025	463 exabytes	90%	2,5 petabytes

- Año: Año en el que se generaron los datos.
- Cantidad de datos generados (bytes): Cantidad total de datos generados en ese año.
- Tráfico de Internet (%): Porcentaje de los datos generados que se originaron en el tráfico de Internet.
- Redes sociales (terabytes): Cantidad de datos generados por las redes sociales en ese día.



# Big Data

Actividad, continuación.

- El aumento de los datos tiene un impacto significativo en la sociedad, ya que abre nuevas oportunidades para la innovación y el desarrollo. Sin embargo, también plantea desafíos, como la necesidad de desarrollar nuevas tecnologías para almacenar y procesar grandes cantidades de datos.

Especificidad	Datos
Cantidad de datos generados cada día	2,5 quintillones de bytes
Equivalencia en bytes	73 zettabytes
Factores que impulsan el crecimiento de los datos	Uso de dispositivos conectados, redes sociales, nube
Impacto de los datos	Oportunidades e innovaciones, desafíos

# Big Data

No existe una definición precisa del término pero los términos de datos masivos o grandes volúmenes de datos hacen referencia al big data. Por este motivo, a menudo el concepto de big data es definido en función de las características que poseen los datos y los procesos que forman parte de este nuevo paradigma de computación. Esto es lo que se conoce como las Vs del Big Data.

## LAS TRES V DEL BIG DATA



# Big Data

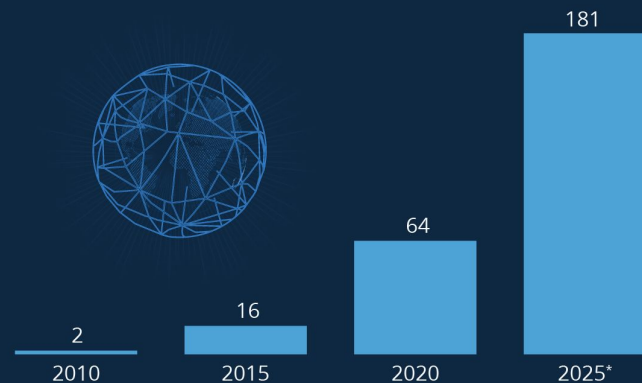
**Volumen** se refiere a la cantidad de datos que se generan y recopilan. El big data se caracteriza por un volumen de datos que es demasiado grande para ser procesado por las herramientas y técnicas tradicionales.

**Variedad** se refiere a la diversidad de los datos. El big data puede provenir de una variedad de fuentes, como sensores, dispositivos móviles, redes sociales y sistemas de transacciones. Los datos pueden ser de diferentes formatos, como texto, números, imágenes y audio.

**Velocidad** se refiere a la rapidez con la que se generan y recopilan los datos. El big data puede generarse y recopilarse a una velocidad que supera la capacidad de las organizaciones para procesarlos.

## El Big Bang del Big Data

Volumen estimado de datos digitales creados o replicados en todo el mundo, en zetabytes



Un zetabyte equivale a mil millones de gigabytes.

\* Previsión de marzo de 2021.

Fuentes: IDC, Seagate, Statista





# Big Data

El big data plantea desafíos para las organizaciones, ya que requieren herramientas y técnicas para almacenar, procesar y analizar grandes cantidades de datos de diferentes fuentes. Sin embargo, el big data también ofrece oportunidades para las organizaciones, ya que puede utilizarse para mejorar la toma de decisiones, la innovación y el servicio al cliente.

Aquí hay algunos ejemplos de cómo las organizaciones utilizan las tres V del big data:

**Volumen:** Las organizaciones utilizan el volumen de datos para identificar tendencias y patrones que no serían visibles con un conjunto de datos más pequeño.

**Variedad:** Uso de la variedad de datos para obtener una visión holística de sus clientes, productos y operaciones.

**Velocidad:** Las organizaciones utilizan la velocidad de los datos para tomar decisiones en tiempo real.

# Big Data

Dado que no existe una definición uniforme para el término big data, muchos autores definen el término en función de aquellas características que consideran más relevantes, por lo que es común encontrar “las cinco Vs del big data”, “las siete Vs del big data” o “las diez Vs del big data” según cada autor.



Figura 3: Definición de big data en base a “Las ocho Vs del big data”.

# Cluster de computadoras

Los clusters son conjuntos de máquinas interconectadas que trabajan como una única unidad para resolver cargas de trabajo de manera conjunta. Estos ofrecen varias ventajas, como alto rendimiento, alta disponibilidad, equilibrio de carga y escalabilidad.



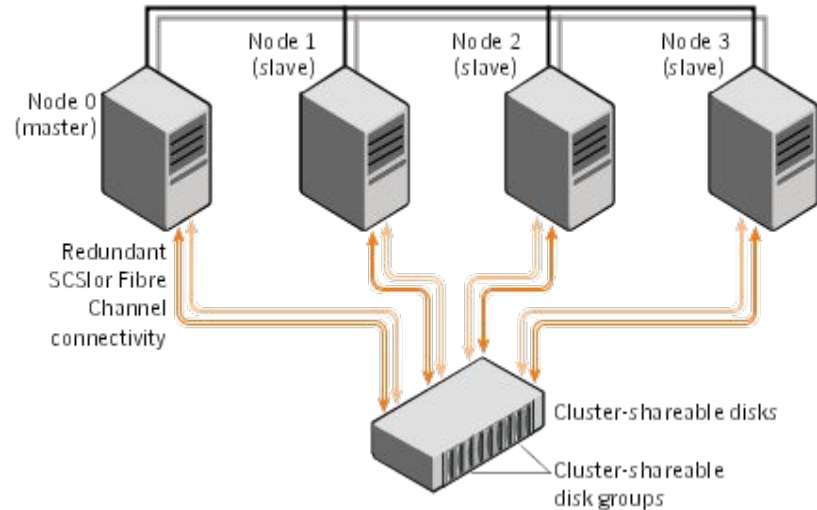


# Cluster de computadoras

## Ventajas:

**Alto rendimiento:** Los clusters permiten acelerar cargas de trabajo al dividir las en subtarefas y distribuir las entre los nodos, lo que posibilita resolver problemas complejos de manera eficiente.

**Alta disponibilidad:** La monitorización constante de los nodos en el cluster permite detectar fallos y tomar medidas para mantener los servicios y datos disponibles, ya sea reiniciando nodos caídos o respondiendo desde réplicas.

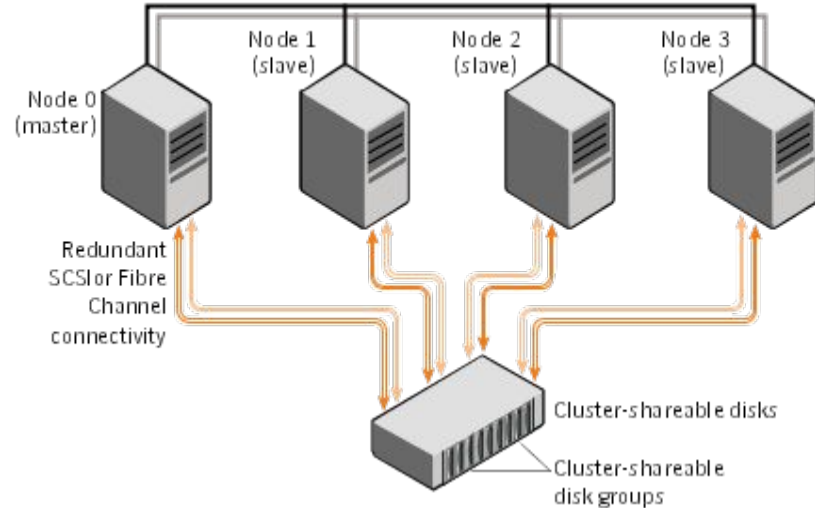


# Cluster de computadoras

## Ventajas:

**Equilibrio de carga:** Los algoritmos distribuyen las cargas de trabajo entre los nodos para evitar cuellos de botella, considerando factores como el tamaño del trabajo y la capacidad de procesamiento de cada nodo.

**Escalabilidad:** La capacidad de añadir nuevos nodos al cluster permite aumentar la potencia de cálculo de manera flexible, sin necesidad de estimaciones previas.





# Almacenamiento. Tipos

## Bases de Datos Relacionales:

Son sistemas de almacenamiento de datos estructurados que utilizan tablas para organizar la información. Cada tabla contiene filas y columnas, donde cada fila representa un registro único y cada columna representa un atributo específico. Estas bases de datos son conocidas por su estructura tabular y la capacidad de establecer relaciones entre diferentes tablas a través de claves primarias y claves foráneas. Ejemplos populares de sistemas de gestión de bases de datos relacionales (RDBMS) incluyen MySQL, PostgreSQL y Microsoft SQL Server.

Gestión de Empleados: Las empresas utiliza una base de datos relacionales para almacenar información sobre sus empleados (nombres, apellidos, edades y departamentos).



# Almacenamiento. Tipos

## Datasets:

Un dataset es un conjunto de datos que se agrupa en una colección organizada. Puede contener datos de cualquier tipo, como texto, números, imágenes o información en formato tabular. Los datasets se utilizan en diversas aplicaciones, desde análisis de datos hasta aprendizaje automático. Los datos dentro de un dataset suelen tener un propósito específico, como la investigación científica o la toma de decisiones empresariales.

**Análisis de Ventas Mensuales:** Un analista utiliza un dataset que contiene datos mensuales de ventas para realizar un análisis de tendencias y tomar decisiones estratégicas sobre estrategias de marketing y gestión de inventario.



# Almacenamiento. Tipos

## 3. Almacenes de Datos (Data Warehouses):

Infraestructura de almacenamiento diseñada para recopilar, almacenar y gestionar grandes volúmenes de datos de diferentes fuentes. Su objetivo principal es proporcionar una única fuente de verdad para el análisis de datos empresariales. Los almacenes de datos suelen estar optimizados para consultas y análisis, lo que facilita la obtención de información valiosa de los datos almacenados.

Análisis Empresarial: Una empresa utiliza un almacén de datos para consolidar información de ventas, inventario y finanzas de todas sus sucursales. Esto permite a los directivos acceder a informes detallados para la toma de decisiones empresariales.



# Almacenamiento. Tipos

## 4. ACID vs. CAP vs. BASE:

ACID (Atomicidad, Consistencia, Aislamiento, Durabilidad): Es un conjunto de propiedades que garantizan la integridad de las transacciones en una base de datos relacional.

- **Atómicas:** se ejecutan en su totalidad o no se ejecutan en absoluto.
- **Consistentes:** mantienen la integridad de la base de datos.
- **Aisladas:** las transacciones no interfieren entre sí.
- **Duraderas:** los cambios persisten incluso en caso de fallo del sistema.

CAP (Consistencia, Disponibilidad, Tolerancia a partición): El teorema CAP establece que en un sistema distribuido, es imposible garantizar simultáneamente la consistencia, la disponibilidad y la tolerancia a partición en caso de una falla de red.



# Almacenamiento. Tipos

## 4. ACID vs. CAP vs. BASE:

BASE (Básicamente Disponible, Suave, Eventualmente consistente): En contraposición a ACID, BASE se refiere a sistemas que priorizan la disponibilidad y la tolerancia a partición en sistemas distribuidos. Los sistemas BASE pueden ser "eventualmente consistentes", lo que significa que, con el tiempo, todos los nodos en el sistema llegarán a un estado consistente.



# Almacenamiento. Tipos

## Bases de Datos NoSQL y Orientadas a Grafos:

**Bases de Datos NoSQL:** Estas bases de datos se utilizan para gestionar datos no estructurados o semiestructurados, como documentos, gráficos o datos de series temporales. No siguen el modelo relacional tradicional y permiten una mayor escalabilidad y flexibilidad en la gestión de datos. Ejemplos incluyen MongoDB (base de datos de documentos) y Cassandra (base de datos de columnas ampliamente distribuida).

**Bases de Datos Orientadas a Grafos:** Estas bases de datos se utilizan para modelar y gestionar datos con relaciones complejas, como redes sociales o sistemas de recomendación. Utilizan estructuras de grafo para representar y almacenar datos, lo que facilita la búsqueda y navegación eficiente de relaciones. Ejemplos incluyen Neo4j y Amazon Neptune.





## Almacenes de datos. Bases de datos tradicionales.

Las bases de datos tradicionales están basadas generalmente en sistemas relacionales u objeto-relacionales. Para el acceso, procesamiento y recuperación de los datos, se sigue el modelo Online Transaction Processing (OLTP). El modelo OLTP (procesamiento de transacciones en línea), permite gestionar los cambios de la base de datos mediante la inserción, actualización y eliminación de información de la misma a través de transacciones básicas que son procesadas en tiempos muy pequeños.

Con respecto a la recuperación de información de la base de datos, se utilizan operadores clásicos (concatenación, proyección, selección, agrupamiento...) para realizar consultas básicas y sencillas (realizadas, mayoritariamente, en lenguaje SQL y extensiones del mismo).



## Almacenes de datos. Bases de datos tradicionales.

Las bases de datos relacionales son colecciones de datos integrados, almacenados en un soporte secundario no volátil y con redundancia controlada. La definición de los datos y la estructura de la base de datos debe estar basada en un modelo de datos que permita captar las interrelaciones y restricciones existentes en el dominio que se pretende modelizar.

A su vez, un Sistema Gestor de Bases de Datos se compone de una colección de datos estructurados e interrelacionados (una base de datos) así como de un conjunto de programas para acceder a dichos datos.



## Almacenes de datos. Bases de datos tradicionales.

La revolución en la generación, almacenamiento y procesamiento de los datos, así como la irrupción del big data, han puesto a prueba el modelo de funcionamiento, rendimiento y escalabilidad de las bases de datos relacionales tradicionales.

En este sentido, la inteligencia de negocio, más conocida por el término inglés business intelligence, investiga en el diseño y desarrollo de este tipo de soluciones. La inteligencia de negocio puede definirse como la capacidad de una empresa de estudiar sus acciones y comportamientos pasados para entender dónde ha estado la empresa, determinar la situación actual y predecir o cambiar lo que sucederá en el futuro. utilizando las soluciones tecnológicas más apropiadas para optimizar el proceso de toma de decisiones.



# Almacenes de datos. Bases de datos tradicionales.

Actividad: Business Intelligence.

- Lee el siguiente artículo.

[https://cloud.google.com/learn/what-is-business-intelligence?hl=es#:~:text=La%20inteligencia%20empresarial%20\(BI\)%20es,de%20decisiones%20estrat%C3%A9gicas%20y%20cotidianas.](https://cloud.google.com/learn/what-is-business-intelligence?hl=es#:~:text=La%20inteligencia%20empresarial%20(BI)%20es,de%20decisiones%20estrat%C3%A9gicas%20y%20cotidianas.)

- Reflexiona sobre el contenido para participar en el debate grupal.



# Almacenes de datos. Modelos de procesamiento.

## Procesamiento en paralelo y distribuido

El procesamiento en paralelo aprovecha las capacidades de los procesadores multinúcleo actuales para ejecutar diferentes hilos de forma concurrente. El sistema operativo reparte el tiempo de CPU entre los hilos.

El procesamiento distribuido utiliza un clúster de máquinas conectadas en red. Divide la carga de trabajo en subtarefas para ejecutar en paralelo en los distintos nodos. Requiere comunicación de datos entre nodos, siendo más rápida dentro del mismo equipo que entre equipos.

# Almacenes de datos. Modelos de procesamiento.



## Estrategias de procesamiento

- **Batch:** Procesamiento sin restricciones de tiempo, típicamente para analítica. Puede tardar horas o días. Se aplica sobre toda la cantidad de datos.
- **Transaccional:** Requiere tiempos de respuesta cortos, por debajo del segundo. Se usa en operaciones de OLTP con bases de datos relacionales.
- **Tiempo real:** Procesamiento de baja latencia para analítica interactiva con usuarios. Se suele usar en sistemas OLAP.
- **Streaming:** Debe procesar datos entrantes a la velocidad que llegan. Obliga a estructuras de datos actualizables en memoria.
- **OLTP:** Orientado a procesamiento de transacciones en línea. Usa bases de datos relacionales y tiempos de respuesta cortos.
- **OLAP:** Orientado a consultas analíticas en tiempo real. Usa bases de datos multidimensionales optimizadas para consultas complejas.



# Almacenes de datos. Modelos de procesamiento.

## Principio SCV

Establece que un sistema de procesamiento distribuido sólo puede tener 2 de estas 3 propiedades:

- Velocidad: tiempo de procesamiento desde que se reciben los datos.
- Consistencia: la precisión y coherencia de los resultados, lo cual puede depender de si se utilizan todos los datos disponibles o solo muestras.
- Volumen: cantidad de datos que se pueden procesar en un tiempo determinado.

**Un sistema distribuido puede ser rápido y preciso pero no manejar grandes volúmenes de datos, o puede ser preciso y manejar grandes volúmenes de datos pero no ser muy rápido.** La elección de estas propiedades depende de las necesidades y las limitaciones del sistema.

# Almacenes de datos. Modelos de procesamiento.

Los principales elementos involucrados en el procesamiento de datos en entornos de Big Data son:

- Fuentes de datos: bases de datos transaccionales, archivos, sensores, aplicaciones, redes sociales, etc.
- Ingestión y recolección de datos: procesos para conectarse a las fuentes, extraer los datos y unificarlos.
- Almacenamiento: sistemas como HDFS, datos en lago, bases NoSQL, etc. .
- Procesamiento distribuido: plataformas como Hadoop, Spark, Flink, etc.
- Estrategias de procesamiento: batch, transaccional, tiempo real, streaming.



# Almacenes de datos. Modelos de procesamiento.

Los principales elementos involucrados en el procesamiento de datos en entornos de Big Data son:

- Estrategias de procesamiento: batch, transaccional, tiempo real, streaming.
- Modelos de programación: MapReduce, SQL, Python, R. Formas de expresar el procesamiento a realizar.
- Infraestructura: clústers de nodos, redes de interconexión, sistemas de archivos distribuidos.
- Seguridad: control de acceso, encriptación, anonimización, etc.
- Gobernanza: definir políticas, calidad de datos, linajes, diccionarios.
- Analítica y visualización: para extraer insights y presentar resultados.

## DATA SOURCES

Internal data sources such as data from CRM system, ERP system, sales reports, etc.

External data sources such as government statistics and media channels



## DATA STORAGE

Big data storage software tools store, manage and retrieve massive amounts of data.



## DATA MINING

Data mining tools allow businesses to extract usable data from a huge set of raw data to find relationships, patterns, and anomalies.



SPSS Modeler



## DATA ANALYTICS

Although data mining tools incorporate data analysis, there are software designed specifically with advanced analytical capabilities.



## DATA VISUALIZATION

Data visualization software is also a type of data analytics tool. However, they are specifically designed to take the raw data and presenting it with beautiful and easy digestible visuals like graphs and charts.





## Almacenes de datos. Modelos de procesamiento.

Estas nuevas soluciones requerirán un modelo de procesamiento diferente a OLTP. Esto es así, ya que el objetivo perseguido por la inteligencia de negocio está menos orientado al ámbito transaccional y más enfocado al ámbito analítico. Las nuevas soluciones utilizan el modelo Online analytical processing (OLAP). La principal diferencia entre OLTP y OLAP estriba en que mientras que el primero es un sistema de procesamiento de transacciones en línea, el segundo es un sistema de recuperación y análisis de datos en línea. Por tanto, OLAP complementa a SQL aportando la capacidad de analizar datos desde distintas variables y dimensiones, mejorando el proceso de toma de decisiones.



## Almacenes de datos. Modelos de procesamiento. OLAP

Los sistemas OLAP están basados, generalmente, en sistemas o interfaces multidimensionales que proporcionan facilidades para la transformación de los datos, permitiendo obtener nuevos datos más combinados y agregados que los obtenidos mediante las consultas simples realizadas por OLTP. Al contrario que en OLTP, las unidades de trabajo de OLAP son más complejas que en OLTP y consumen más tiempo. Finalmente, en cuanto a la visualización de los mismos, los sistemas OLAP permiten la visualización y el análisis multidimensional a partir de diferentes vistas de los datos, presentando los resultados en forma matricial y con mayores posibilidades estéticas y visuales.



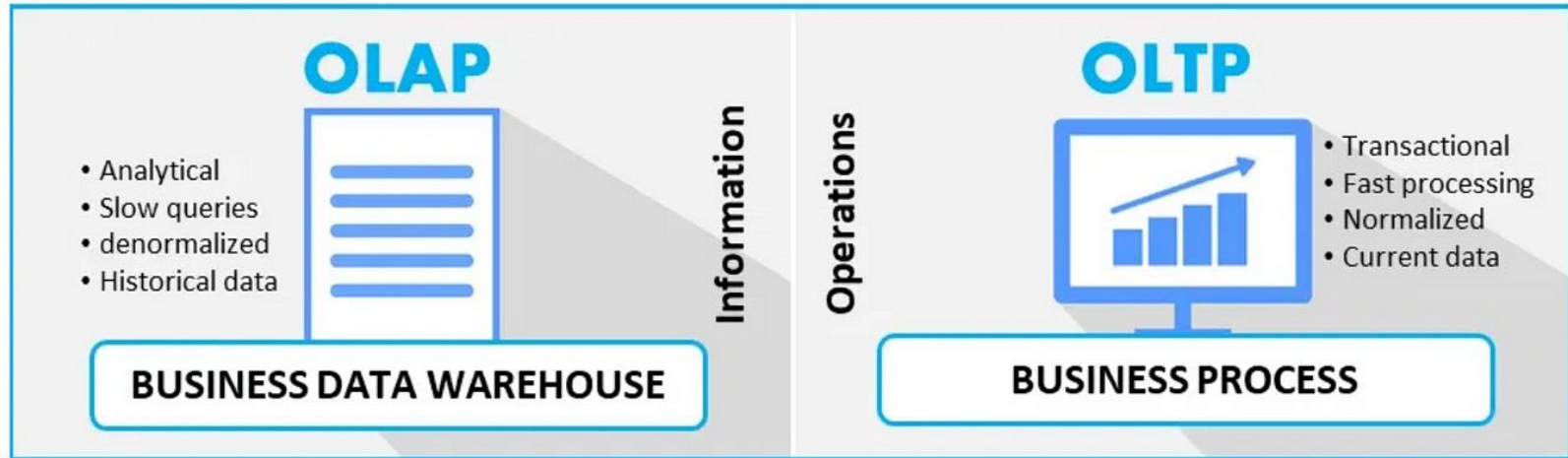
## Almacenes de datos. Modelos de procesamiento. OLAP vs OLTP

**Tabla 1:** Tabla resumen y comparativa entre OLTP y OLAP

	Bases de datos relacionales	Soluciones Business Intelligence
	OLTP	OLAP
<b>Concepto</b>	Sistema de procesamiento de transacciones en línea	Sistema de recuperación y análisis de datos en línea
<b>Funciones</b>	Gestión de transacciones: inserción, actualización, eliminación...	Análisis de datos para dar soporte a la toma de decisiones
<b>Procesamiento</b>	Transacciones cortas	Procesamientos de análisis complejos
<b>Tiempo</b>	Las transacciones requieren poco tiempo de ejecución	Los análisis requieren mayor tiempo de ejecución
<b>Consultas</b>	Simple, utilizando operadores básicos tradicionales	Complejas, permitiendo analizar los datos desde múltiples dimensiones
<b>Visualización</b>	Básica. Muestra los datos en forma tabular	Muestra los datos en forma matricial. Mayores posibilidades gráficas

Almacenes de datos. Modelos de procesamiento. OLAP vs OLTP

# OLAP Vs OLTP





# Almacenes de datos. Modelos de procesamiento. OLAP vs OLTP

## OLTP (Procesamiento de Transacciones en Línea):

**Propósito:** OLTP se utiliza para procesar transacciones comerciales en tiempo real. Su objetivo principal es mantener la integridad y consistencia de los datos transaccionales.

### Ejemplos:

- **Pedidos:** Cuando un cliente realiza una compra en línea, el sistema OLTP registra la transacción, actualiza el inventario y procesa el pago.
- **Reservas:** En una aerolínea, el sistema OLTP maneja las reservas de vuelos, asientos y emite boletos.
- **Actualizaciones de cuentas:** Los bancos utilizan OLTP para registrar depósitos, retiros y transferencias entre cuentas.

## OLAP (Procesamiento Analítico en Línea):

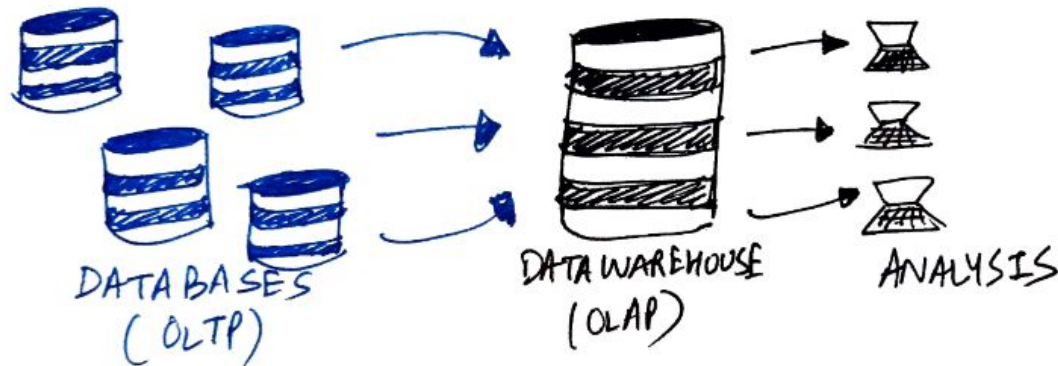
**Propósito:** OLAP se centra en el análisis de datos agregados desde diferentes perspectivas. Ayuda a tomar decisiones estratégicas basadas en información procesable.

### Ejemplos:

- **Informes financieros:** Un equipo financiero utiliza OLAP para analizar los ingresos, gastos y tendencias financieras de la empresa.
- **Minería de datos:** Los científicos de datos utilizan OLAP para descubrir patrones ocultos en grandes conjuntos de datos.
- **Planificación estratégica:** Una cadena minorista utiliza OLAP para evaluar el rendimiento de las tiendas, identificar productos populares y optimizar la colocación.

## Almacenes de datos. Modelos de procesamiento. OLAP vs OLTP

OLTP se enfoca en transacciones individuales y actualizaciones constantes de datos, mientras que **OLAP** se centra en análisis complejos y toma de decisiones estratégicas. Ambos sistemas son esenciales para el funcionamiento eficiente de una organización.





## Almacenes de datos. Modelos de procesamiento. OLAP vs OLTP

La arquitectura del software OLAP incluye dos componentes básicos:

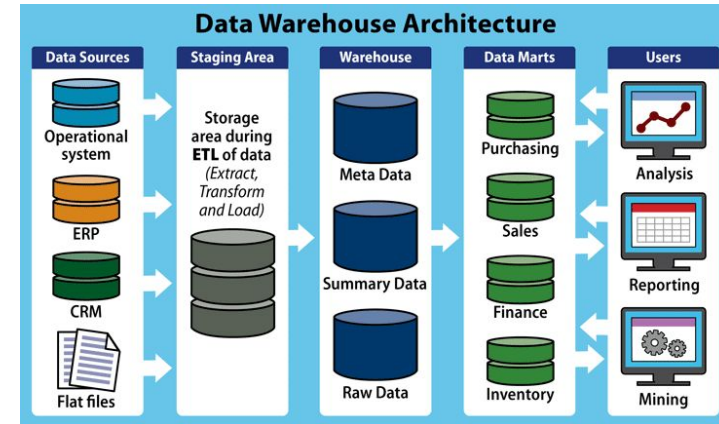
**Servidor OLAP:** proporciona almacenamiento de datos, realizando sobre ellos las operaciones necesarias y la formación de un modelo multidimensional a nivel conceptual.

**Cliente de procesamiento analítico en línea:** presenta al usuario una interfaz para el modelo de datos multidimensional, brindándole la capacidad de manipular datos convenientemente para realizar tareas de análisis.



# Almacenes de datos. Datawarehouse

Un almacén de datos, más conocido por el término data warehouse (en inglés), es una solución de business intelligence que combina tecnologías y componentes con el objetivo de ayudar al uso estratégico de los datos por parte de una organización. Esta solución debe proveer a la empresa, de forma integrada, de capacidad de almacenamiento de una gran cantidad de datos así como de herramientas de análisis de los mismos que, frente al procesamiento de transacciones, permita transformar los datos en información para ponerla a disposición de la organización y optimizar el proceso de toma de decisiones.



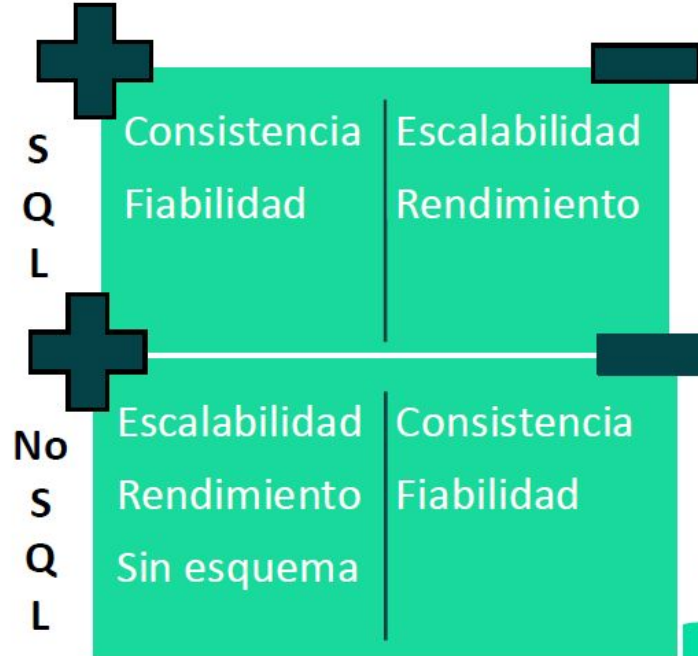



## Almacenes de datos. Bases de datos documentales u orientadas a documentos.

La gran variedad y heterogeneidad en los tipos de datos almacenados y procesados en los últimos años ha puesto puesto en cuestión de si las bases de datos relacionales son el modelo más óptimo para trabajar con según qué tipos de datos. Como alternativa a ellas, en los últimos años han proliferado las bases de datos NoSQL.

NoSQL es el término utilizado para referirse a un tipo de bases de datos que permiten almacenar y gestionar tipos de datos que tradicionalmente han sido difíciles de gestionar por parte de las bases de datos relacionales. Así pues, NoSQL hace referencia a bases de datos documentales, bases de datos orientadas a grafos, buscadores, etc.

Almacenes de datos. Bases de datos documentales u orientadas a documentos.





## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Comparativa.

Bases de datos SQL	Bases de datos NoSQL
Son bases de datos relacionales	Son bases de datos no relacionales o distribuidas
Utilizan lenguaje de consulta estructurado (SQL) y tienen un esquema predefinido	Tienen esquemas dinámicos para datos no estructurados
Son mejores para transacciones con múltiples filas	Son mejores para datos no estructurados como documentos o JSON
Son eficientes, flexibles y pueden ser accedidas fácilmente por cualquier aplicación	Ofrecen escalabilidad horizontal, lo que significa que simplemente se deben agregar más servidores para aumentar su carga de datos
Tienen esquemas rígidos, complejos y tabulares y típicamente requieren una escalabilidad vertical costosa	Son más flexibles y pueden manejar una variedad de tipos de datos
Son una buena opción cuando se trabaja con datos relacionados	Son útiles para infraestructuras modernas basadas en la nube
Ejemplos: MySQL, Oracle, PostgreSQL	Ejemplos: MongoDB, Cassandra, Couchbase, Amazon DynamoDB, Redis



## Almacenes de datos. Bases de datos documentales u orientadas a documentos. NoSQL.

Las NoSQL se caracterizan principalmente por:

**Independencia del esquema:** Al contrario que en las bases de datos relacionales, no es necesario diseñar un esquema para definir los tipos y estructura de los datos almacenados, permitiendo acortar el tiempo de desarrollo y facilitando las modificaciones de la estructura interna de la base de datos.

**No relacionales:** El concepto de relación de las bases de datos relacionales no existe en NoSQL. Por tanto, se trabaja con datos que no están normalizados, lo cual aporta flexibilidad en relación a los tipos y estructuras de datos que pueden ser almacenados.

**Distribuidas:** La cantidad de datos almacenados requiere de su almacenamiento en múltiples servidores, ya que un único servidor por potente que sea no podrá procesar en tiempos razonables tal cantidad de información. Este hecho permite utilizar hardware sencillo, ya que al utilizar múltiples servidores no es necesario que todos ellos tengan grandes prestaciones.



## Almacenes de datos. Bases de datos documentales u orientadas a documentos. NoSQL.

Las bases de datos documentales trabajan con documentos, entendidos como una estructura jerárquica de datos que, a su vez, puede contener subestructuras. Las bases de datos documentales pueden, efectivamente, trabajar con estos tipos de documentos. Sin embargo, el término documento en este contexto posee un mayor nivel de abstracción.

Los documentos pueden consistir en datos binarios o texto plano. Es posible que se traten de datos semiestructurados, cuando aparecen en formatos como JavaScript Object Notation (JSON) o Extensible Markup Language (XML). Por último, también pueden ser datos estructurados conforme a un modelo de datos particular como, por ejemplo, XML Schema Definition (XSD).



## Almacenes de datos. Bases de datos documentales u orientadas a documentos.

Actualmente, **XML y JSON** son los formatos de intercambio de datos más utilizados en el desarrollo de aplicaciones web.

### XML:

- XML (Extensible Markup Language) surge como una extensión de SGML, un lenguaje de marcado genérico creado para definir gramáticas de lenguajes.
- XML permite definir reglas para codificar documentos con una sintaxis específica y etiquetas personalizadas. Se compone de elementos (las etiquetas que encierran datos) y atributos (datos adicionales dentro de las etiquetas).
- Su orientación a documentos lo hace muy flexible para representar información estructurada de forma jerárquica y compleja. Sin embargo, al centrarse en la estructura, XML tiene cierta redundancia y verbosidad.





## Almacenes de datos. Bases de datos documentales u orientadas a documentos.

En XML, la estructura principal de un documento está formada por dos elementos: el prólogo (opcional) y el cuerpo. El prólogo contiene a su vez dos partes: la declaración XML que establece la versión del lenguaje, el tipo de codificación y si se trata de un documento autónomo y la declaración del tipo de documento. El cuerpo, por su parte, contiene la información del documento.

Supongamos que Carlos ha enviado un mensaje de whatsapp a Javier diciéndole que han quedado con los compañeros de trabajo a las diez de la noche en la puerta del Sol. Un documento XML que representa este mensaje como un documento, podría ser el mostrado en el listado.



## Almacenes de datos. Bases de datos documentales u orientadas a documentos.

XML:

Listado 1: Quedada en la puerta del sol

```
1 <?xml version="1.0" encoding="ISO-8859-1"?>
2 <whatsapp>
3     <para> Javier </para>
4     <de> Carlos </de>
5     <titulo> Quedada </titulo>
6     <contenido> A las 22:00 pm en la puerta del sol </contenido>
7 </whatsapp>
```

El listado 1 incluye en la primera línea el **prólogo del documento**, definiendo la versión y el tipo de codificación utilizada. A partir de la segunda línea, se define el cuerpo del documento que contiene, mediante etiquetas de apertura <> y cierre </>, los distintos atributos de los que se compone el documento.



## Almacenes de datos. Bases de datos documentales u orientadas a documentos.

JSON:

- JSON (JavaScript Object Notation) nace como un subconjunto de la notación de objetos JavaScript, orientado al intercambio de datos entre aplicaciones web.
- Soporta menos tipos de datos que XML (cadenas, números, booleanos, arrays, objetos), pero tiene una sintaxis más simple y compacta basada en pares clave-valor.
- Al enfocarse solo en representar datos, JSON resulta más ligero y rápido de parsear que XML. Es ideal para el intercambio de información simple entre cliente y servidor.



## Almacenes de datos. Bases de datos documentales u orientadas a documentos.

En JSON, la sintaxis del lenguaje tiene las mismas reglas que el lenguaje JavaScript, del cual proviene.

Los archivos JSON deben cumplir también otras reglas sintácticas adicionales. En primer lugar, un archivo JSON representará o bien un objeto, es decir, una tupla de pares clave-valor o bien una colección de elementos, es decir, un vector o array.

Los archivos JSON que representan objetos comienzan con una llave de inicio { y terminan con una llave de cierre }. Cuando se representa un vector, sus elementos se encierran entre corchetes []. Las cadenas y nombres de atributos del objeto deberán encerrarse entre comillas, así como todos los nombres de los atributos del objeto, separándose cada elemento del siguiente con una coma (,) no habiendo una coma después del último elemento.



## Almacenes de datos. Bases de datos documentales u orientadas a documentos.

Así pues, si se pretende representar en formato JSON el mensaje que ha enviado Carlos a Javier, el fichero JSON resultante sería el mostrado en el listado 2.

Este fichero define un objeto JSON que contiene una serie de atributos entrecomillados cuyo valor asociado son cadenas de caracteres que, por tanto, también van entrecomilladas.

Listado 2: Quedada en la puerta del sol

```
1 {  
2     "para": "Javier",  
3     "de": "Carlos",  
4     "titulo": "Quedada",  
5     "contenido": "A las 22:00 pm en la puerta del sol"  
6 }
```



## Almacenes de datos. Bases de datos documentales u orientadas a documentos.

Tabla 2: Tabla comparativa entre XML y JSON

	XML	JSON
Lenguaje fuente	SGML	JavaScript
Tipo Lenguaje	Orientado a datos	Orientado a documentos
Notación	Pesada	Ligera
Etiquetas inicio y fin	Sí	No
Comentarios	Sí	No
Espacios de nombres	Sí	No
Soporte tipos de datos	No	Sí

XML es más apropiado para documentos complejos con mucha estructura, mientras que JSON es mejor para transmitir objetos de datos simples de forma ágil y con menor sobrecarga. Ambos siguen siendo ampliamente utilizados como formatos universales de intercambio de información en el desarrollo web y de APIs.



## Almacenes de datos. Bases de datos documentales u orientadas a documentos.

Actividad: Utilice la siguiente aplicación para convertir el código XML de ejemplo a JSON y conteste a las siguientes preguntas: [XML to JSON Converter](#)

- ¿Cuáles son las diferencias entre XML y JSON?
- ¿Cómo se representan los datos en XML?
- ¿Cómo se representan los datos en JSON?

XML

```
<producto>
<nombre>Ordenador portátil</nombre>
<precio>1.200 €</precio>
<marca>Acer</marca>
<modelo>Aspire 5</modelo>
<características>
  <característica>Procesador Intel Core i5</característica>
  <característica>Memoria RAM de 8 GB</característica>
  <característica>Disco duro de 1 TB</característica>
  <característica>Pantalla de 15,6 pulgadas</característica>
</características>
</producto>
```



## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos.

Un grafo es un ente matemático compuesto por un conjunto de nodos o vértices y un conjunto de enlaces o aristas. Matemáticamente puede ser expresado por medio de la ecuación:

$$G = \{V, E\}$$

(Donde V representa el conjunto de nodos o vértices y E representa el conjunto de enlaces o aristas.



## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos.

El grafo de la siguiente figura representa un conjunto de ciudades conectadas por autovías, el conjunto  $V$  de nodos sería  $V = \{\text{La Coruña, Madrid, San Sebastián, Barcelona, Valencia, Sevilla, Cádiz}\}$ , mientras que el conjunto de enlaces  $E$  vendría dado por  $E = \{A-1, A-2, A-3, A-4, A-4-I, A-4-II, A-6, A-7\}$ .

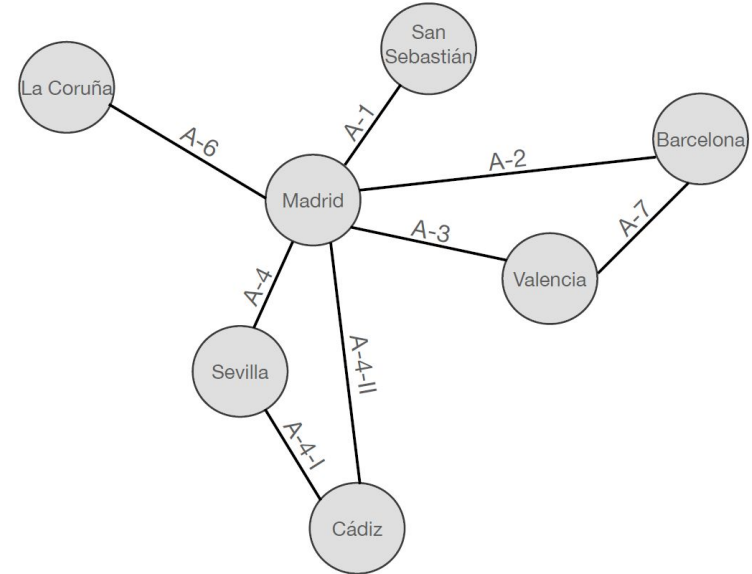


Figura 4: Grafo que representa las principales autovías de España



## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos.

Una base de datos orientada a grafos es, por tanto, un sistema de bases de datos que implementa métodos de creación, lectura, actualización y eliminación de datos en un modelo expresado en forma de grafo.

Existen dos aspectos fundamentales en este tipo de sistemas: el primero de ellos hace referencia al almacenamiento de los datos. En una base de datos orientada a grafos, los datos pueden almacenarse siguiendo el modelo relacional, lo que implica mapear la estructura del grafo a una estructura relacional, o bien, almacenarse de forma nativa utilizando modelos de datos propios para almacenar estructuras de tipo grafo.



## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos.

La ventaja de mapear los grafos a una estructura relacional radica en que la gestión y consulta de los datos se realizará de forma tradicional a través de un backend conocido como, por ejemplo, MySQL.

La ventaja del almacenamiento nativo de grafos radica en que existen modelos de datos e implementaciones que aseguran y garantizan el buen rendimiento y la escalabilidad del sistema.

El segundo aspecto importante es el procesamiento de los datos. El procesamiento nativo de los datos de grafos, el cual es beneficioso porque optimiza los recorridos del grafo cuando se realizan consultas aunque, en ocasiones, invierta demasiado tiempo y memoria en consultas que no requieren de recorridos complejos.



## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos.

Cualquier dominio puede ser modelado como un grafo. La motivación para requerir de sistemas de bases de datos específicos, orientados a trabajar con este tipo de datos, radica en tres aspectos principales:

- Rendimiento: Consultas más eficientes que se localizan en porciones del grafo, con complejidad constante al escalar. Superan a las BD relacionales.
- Flexibilidad: No requiere modelado exhaustivo previo. Se pueden agregar nodos y relaciones sobre la marcha sin modificar todo el modelo. Facilita la implementación.
- Agilidad: Gestión ágil de los datos gracias a las ventajas de rendimiento y flexibilidad. Permite metodologías de desarrollo ágil y diseño rápido de software que usa grafos.



## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos.

Lenguajes de marcado de grafos:

- GraphML
- eXtensible Graph Markup and Modeling Language (XGMML)
- Graph Exchange Language (GXL)
- Graph Modelling Language (GML)

La mayoría de ellos son variantes o extensiones del lenguaje XML para el modelado de grafos.

GraphML es uno de los lenguajes más extendidos para el modelado de datos en forma de grafo. Se trata de un lenguaje sencillo, general, extensible y robusto. La notación es muy similar a XML. A modo de ejemplo, el grafo mostrado en la figura 4 podría representarse en GraphML según se muestra en el listado



## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos.

GraphML

### GraphML

#### Definición:

- Derivado de XML
- Muy extendido para el modelado de grafos (GML, GXL, XGMML)
- Sencillo
- General
- Extensible
- Robusto





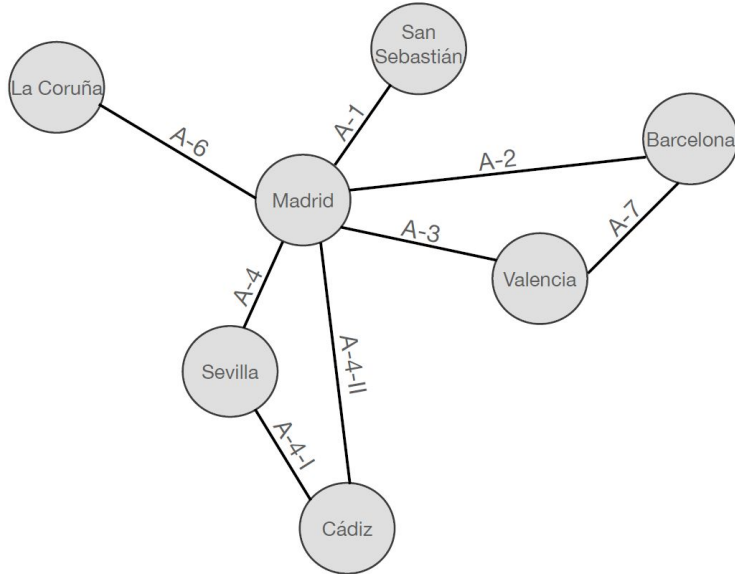
## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos.

- Se define un grafo llamado "Grafo\_Autovias" con aristas no dirigidas (edgedefault="undirected"), por lo tanto los enlaces son bidireccionales.
- Se especifican los nodos del grafo, que representan ciudades.
- Se definen las aristas entre nodos, indicando el identificador de cada arista y sus nodos origen y destino.
- Al ser no dirigido, el origen y destino son intercambiables.
- Si fuera un grafo dirigido, se debería especificar edgedefault="directed" y origen/destino no serían intercambiables.
- Permite modelar la red de autovías entre distintas ciudades como un grafo no dirigido en GraphML.

Listado 3: Grafo Autovías

```
1 <graph id="Grafo_Autovias" edgedefault="undirected">
2   <node id="La Coruña"/> <node id="San Sebastián"/>
3   <node id="Madrid"/> <node id="Barcelona"/>
4   <node id="Valencia"/> <node id="Sevilla"/>
5   <node id="Cadiz"/>
6   <edge id = "A-6" source="La Coruña" target="Madrid"/>
7   <edge id = "A-1" source="Madrid" target="San Sebastián"/>
8   <edge id = "A-2" source="Madrid" target="Barcelona"/>
9   <edge id = "A-7" source="Barcelona" target="Valencia"/>
10  <edge id = "A-3" source="Madrid" target="Valencia"/>
11  <edge id = "A-4" source="Madrid" target="Sevilla"/>
12  <edge id = "A-4-I" source="Sevilla" target="Cádiz"/>
13  <edge id = "A-4-II" source="Madrid" target="Cádiz"/>
14 </graph>
```

## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos.



Listado 3: Grafo Autovías

```
1 <graph id="Grafo_Autovías" edgedefault="undirected">
2   <node id="La Coruña"/> <node id="San Sebastián"/>
3   <node id="Madrid"/> <node id="Barcelona"/>
4   <node id="Valencia"/> <node id="Sevilla"/>
5   <node id="Cadiz"/>
6   <edge id = "A-6" source="La Coruña" target="Madrid"/>
7   <edge id = "A-1" source="Madrid" target="San Sebastián"/>
8   <edge id = "A-2" source="Madrid" target="Barcelona"/>
9   <edge id = "A-7" source="Barcelona" target="Valencia"/>
10  <edge id = "A-3" source="Madrid" target="Valencia"/>
11  <edge id = "A-4" source="Madrid" target="Sevilla"/>
12  <edge id = "A-4-I" source="Sevilla" target="Cádiz"/>
13  <edge id = "A-4-II" source="Madrid" target="Cádiz"/>
14 </graph>
```

Figura 4: Grafo que representa las principales autopistas de España



## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos.

- Ente matemático
- Conjunto de vértices o nodos ( $V$ )
- Conjunto de aristas o enlaces ( $E$ )



- Búsqueda de caminos mínimos
- Síntesis de circuitos
- Redes de comunicaciones
- Análisis de redes sociales
- ...

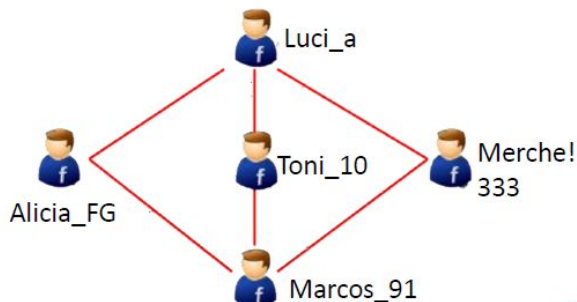


## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos. Ejemplos

### GraphML:

Ejemplo:  
Facebook

- $|V| = 5$
- $|E| = 6$
- Grafo no dirigido



### GraphML:

Ejemplo:  
Facebook

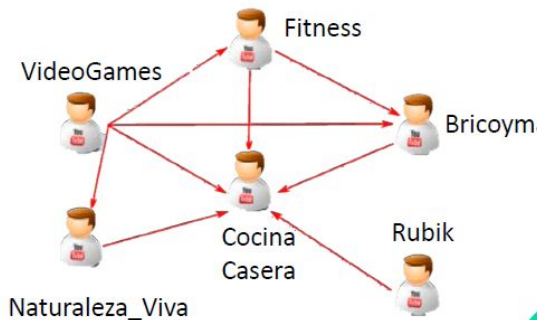
```
<graph id = "Grafo Facebook" edgedefault = "undirected">
  <node id = "Luci_a"/>
  <node id = "Alicia_FG"/>
  ...
  <edge id = "1" source = "Alicia_FG"
    target = "Luci_a"/>
  <edge id = "2" source = "Luci_a" target =
    "Toni_10"/>
  <edge id = "3" source = "Lucia_a" target
    = "Merche!333"/>
  <edge id = "4" source = "Alicia_FG"
    target = "Marcos_91"/>
  <edge id = "5" source = "Toni_10" target
    = "Marcos_91"/>
  <edge id = "6" source = "Merche!333"
    target = "Marcos_91"/>
</graph>
```

## Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos. Ejemplos

### GraphML:

- $|V| = 6$
- $|E| = 9$
- Grafo dirigido

Ejemplo:  
Youtube



### GraphML:

Ejemplo:  
Youtube

```
<graph id = "Grafo Youtube" edgedefault = "directed">
  <node id = "Fitness"/> <node id = "VideoGames"/>
  ...
  <edge id = "1" source = "VideoGames" target = "Fitness"/>
  <edge id = "2" source = "VideoGames" target = "Bricoyrn"/>
  <edge id = "3" source = "VideoGames" target = "Cocina_Casera"/>
  <edge id = "4" source = "VideoGames" target = "Naturaleza_Viva"/>
  <edge id = "5" source = "Fitness" target = "Bricoyrn"/>
  <edge id = "6" source = "Fitness" target = "Cocina_Casera"/>
  ...
</graph>
```



Almacenes de datos. Bases de datos documentales u orientadas a documentos. Grafos.

**GraphML:**

**Visualización:**

- yEd
- <https://www.yworks.com/products/yed>
- Gephi
- <https://gephi.org>



*Gp* Gephi



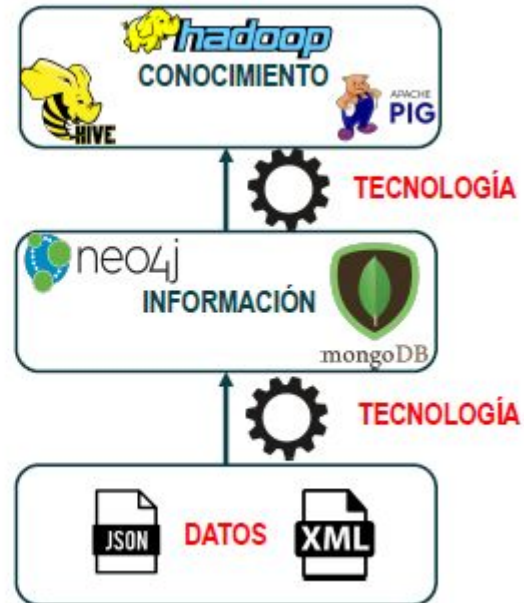
## Tecnologías Big Data.



## Tecnologías Big Data.

Toma de  
decisiones

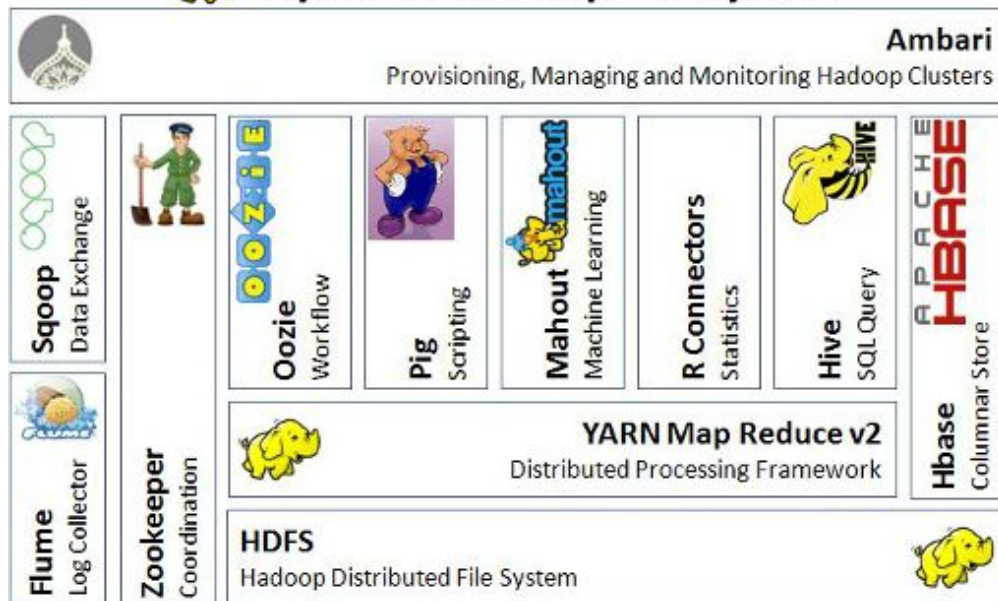
Datos vs  
Tecnología



## Tecnologías Big Data.



### Apache Hadoop Ecosystem





## Tecnologías Big Data.

# Big Data Landscape







## Ecosistema de herramientas, plataformas y soluciones de Big Data

- **Hadoop:** plataforma pionera para procesamiento distribuido de grandes volúmenes de datos. Incluye HDFS, MapReduce, YARN, Hive, HBase, entre muchos otros.
- **Spark:** motor de procesamiento en memoria altamente eficiente para trabajo batch y streaming. Compatible con ecosistema Hadoop.
- **Bases de datos NoSQL:** permiten escalar el almacenamiento y throughput, siendo aptas para datos no estructurados. Ejemplos: MongoDB, Cassandra, Redis.
- **Bases de datos NewSQL:** combinan escalabilidad de NoSQL con funcionalidades de SQL. Ejemplos: VoltDB, MemSQL.



## Ecosistema de herramientas, plataformas y soluciones de Big Data

- **Lambda architecture:** combina procesamiento batch y tiempo real en capas separadas que se unen en una vista de servicios.
- **Kafka:** sistema de mensajería distribuida de altas prestaciones, suele usarse para streaming.
- **Flink:** plataforma de procesamiento de flujos de eventos para analytics en tiempo real y streaming.
- **Cloud computing:** servicios en la nube como AWS, GCP y Azure proveen recursos escalables para Big Data.



## Ecosistema de herramientas, plataformas y soluciones de Big Data

- **Herramientas de ETL:** para extraer, transformar y cargar datos desde las fuentes.  
Ejemplo: Apache Nifi.
- **Machine learning:** algoritmos como regresión, árboles de decisión, redes neuronales.  
Se integran en pipelines.
- **Visualización:** Tableau, Power BI, Grafana para crear dashboards e informes a partir de los datos.