



Big Data

Big Data, Datos Masivos, Macro Datos, llámalos como quieras, pero su importancia es inmensa.

- Enrique Ulises Báez Gómez Tagle
- Mauricio Ascencio Martínez
- Sara Rocío Miranda Mateos

Introducción

Conjuntos de datos grandes y complejos que no pueden gestionarse fácilmente con las herramientas tradicionales de gestión de bases de datos. Estos conjuntos de datos se caracterizan por su volumen, velocidad y variedad.



Volumen

Son masivos, con conjuntos de datos que pueden alcanzar petabytes de capacidad.

(Gestión de base de datos tradicional no es eficaz)



Velocidad

Siempre en movimiento, lo que dificulta su análisis y procesamiento en tiempo real.

(Necesitan algoritmos y herramientas complejas)



Variedad

Heterogéneos y se presentan en distintas formas, como texto, imágenes y audio.

(Desafío: estandarización)

Variedad del Big Data

Datos No Estructurados

Texto sin formato, imágenes, videos, archivos de audio, publicaciones en redes sociales, correos electrónicos, documentos PDF

Datos Semi-Estructurados

Archivos XML, JSON, archivos de registro y datos HTML

Datos Estructurados

Registros de bases de datos, hojas de cálculo y archivos CSV





Big Data y sus retos

Los retos surgen de acuerdo a la naturaleza del conjunto de datos. Estos son algunos de los retos:

1 Calidad de Datos

Suelen haber datos incompletos, incoherentes o mal organizados. Comúnmente, es necesario limpiar y normalizar los datos antes de proceder al análisis.

2 Seguridad y Privacidad

Pueden infringir la privacidad de las personas o contener información sensible de la empresa.

3 Costos

Se requieren potentes ordenadores y programas informáticos para su análisis.

(+ infraestructura, + personal)

∴ \$\$\$

Análisis de Big Data

Proceso de examinar grandes volúmenes de datos para descubrir patrones y correlaciones ocultos.

Análisis Predictivo

Aprendizaje automático o inteligencia artificial nos sirven para predecir acontecimientos futuros.

(Regresiones, correlaciones)

Análisis prescriptivo

Recomendar posibles acciones que pueden ejecutarse basándose en el análisis.

(Aprendizaje automático para la toma de decisiones)

Análisis descriptivo

Comprender eventos pasados y patrones dentro de un conjunto de datos.

Ayuda a detectar problemas o anomalías en un sistema.

Herramientas y tecnología (Apache)



Hadoop

Sistema distribuido para gestionar grandes tareas de procesamiento de datos.

Clusters + MapReduce



Spark

Procesamiento más rápido de conjuntos de datos + integración con DB, colas de mensajes y streaming



Cassandra

Maneja grandes cantidades de datos, ya sea que estén estructurados o no.



MapReduce

Modelo de programación y un marco asociado para procesar grandes conjuntos de datos. Fue desarrollado por Google y es fundamental para muchos sistemas de Big Data, Map Reduce consta de dos pasos principales.

| Map | Reduce |
|--|---|
| <p>En esta etapa, el conjunto de datos de entrada se divide en fragmentos más pequeños y se procesa de manera paralela. Cada operación de mapeo toma un par clave-valor y produce un conjunto de pares clave-valor intermedio.</p> | <p>En esta, todos los pares clave-valor intermedios se agrupan por clave y se procesan para generar el conjunto de datos de salida. La función de reducción toma una clave y un conjunto de valores para esa clave, y se combinan para reducir los datos.</p> |

MapReduce se puede utilizar para procesar petabytes de datos en miles de máquinas. Es especialmente útil para operaciones que requieren un procesamiento intensivo de datos, como la búsqueda y la indexación de texto, minería de datos, y el análisis de registros de máquinas.

Big Data en el mundo actual



Sanidad

Patrones de enfermedad y
diagnóstico predictivo



Transporte

Patrones de tráfico,
optimización de rutas



Finanzas

Tendencias de mercado y
reducción de fraudes

En el futuro

Para satisfacer necesidades cambiantes, algunas de las tendencias a futuro:

Edge Computing

Un enfoque descentralizado de la informática que permitirá recopilar y analizar los datos más cerca de su fuente.

Blockchain

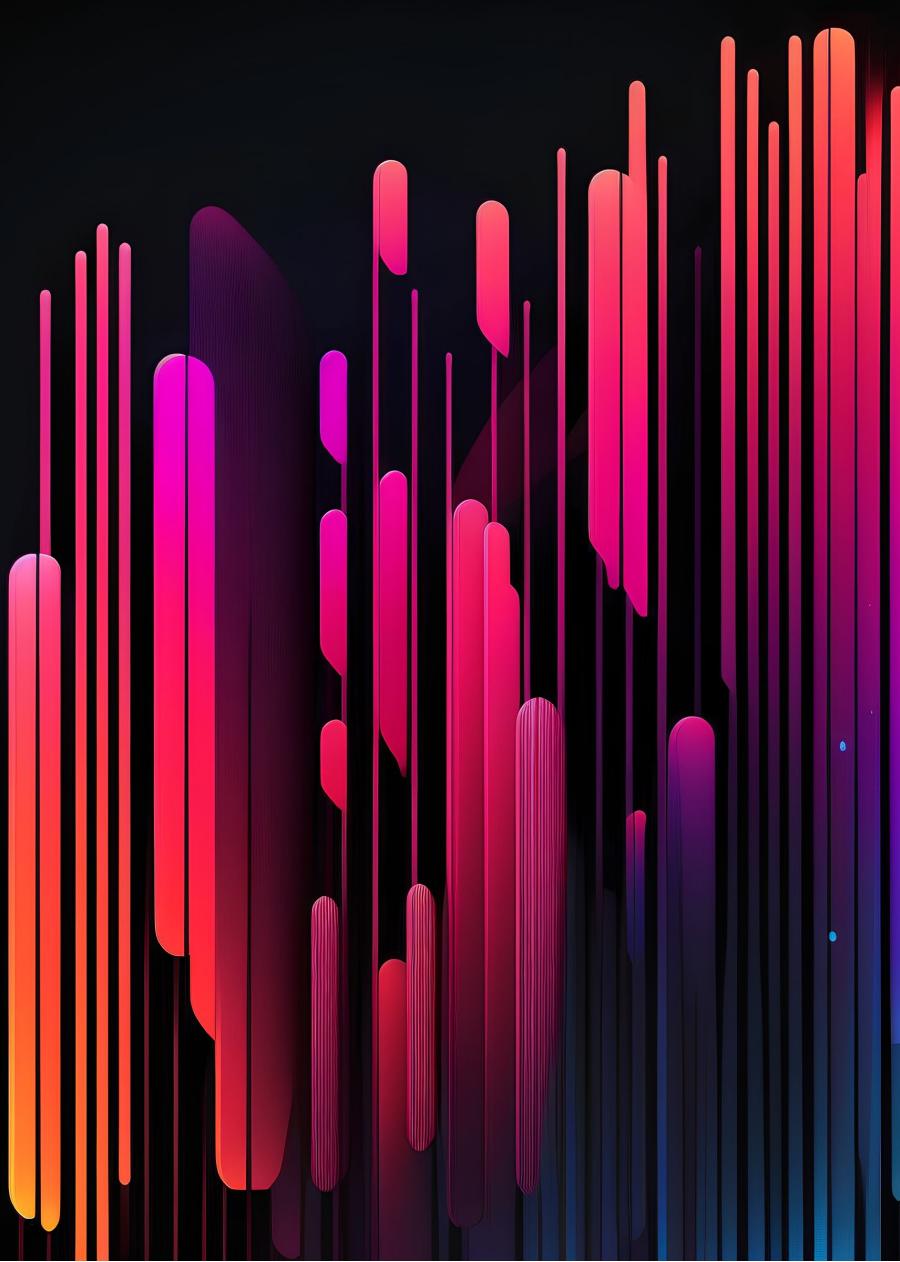
Una tecnología distribuida que ofrece una forma segura de almacenar y compartir datos. Tiene el potencial de transformar el almacenamiento y la comunicación de datos.

Inteligencia Artificial

Big data utilizado en el procesamiento de grandes cantidades de datos y análisis de los mismos utilizado por la IA.

Computación Cuántica

Ordenadores cuánticos capaces de procesar grandes cantidades de datos mucho más rápido que los ordenadores tradicionales.



Conclusión

Los Datos Masivos están transformando nuestra forma de vivir y trabajar. Presentan una serie de retos y oportunidades que requieren un pensamiento innovador y nuevas herramientas para resolverlos. A medida que los macrodatos sigan evolucionando, estimularán la innovación e impulsarán el cambio en muchos sectores.

Proyecto: Analizar opiniones de clientes

Objetivo: Analizar un conjunto de datos de reseñas de clientes y obtener métricas que nos ayuden a tomar decisiones.

Steps:

1. Data Collection:

- Asegurarnos de que nuestro conjunto de datos contiene información relevante, como opiniones de clientes, valoraciones, detalles de productos, información que necesitamos.

2. Data Preprocessing:

- a. Limpia el data set (entradas irrelevantes o duplicadas)
- b. Normalizar datos (eliminar caracteres especiales, convertir a minúsculas)

3. Data Storage:

- a. Elige una solución de Almacenamiento adecuada de acuerdo a las necesidades (Hadoop Distributed File System).
- b. Carga los datos preprocesados.

4. Data Analysis:

- a. Utiliza un framework de preprocesado para realizar el análisis (Apache Hadoop).
- b. Calcular estadísticas comunes (calificación promedio, palabras más usadas, análisis de sentimientos).
- c. Realizar un análisis profundo (modeling, clustering, recommendation systems).

5. Visualization:

- a. Analiza los resultados obtenidos para generar conclusiones.
- b. Identifica patrones, tendencias y correlaciones.
- c. Utiliza la analítica obtenida para mejorar la satisfacción del cliente y generar mejoras.



Preguntas

¿Cuales son las 3 características del Big Data?

¿Qué es un zettabyte?

¿Qué tecnologías se pueden utilizar para el Big Data?

¿Cuales son 3 beneficios del big Data en la actualidad?

¿En que nos ayudará el Big Data en el futuro?

¿Los videos, fotos, emails son ejemplos de que tipo de datos?