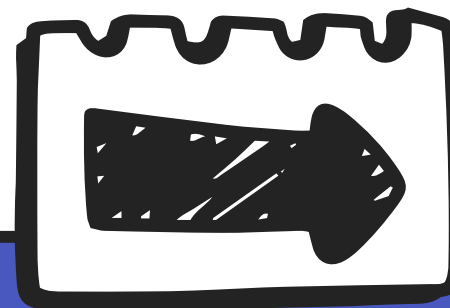




EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS

BIG DATA

ÁNGELES MUTHE LUIS ANTONIO DE JESÚS
GARCÍA AGUAZUL ÁNGEL
ORTIZ MEDINA OCTAVIO
RANGEL GARCÍA ALEXIS
RODRIGUEZ CORONA ALELI GERALDINE

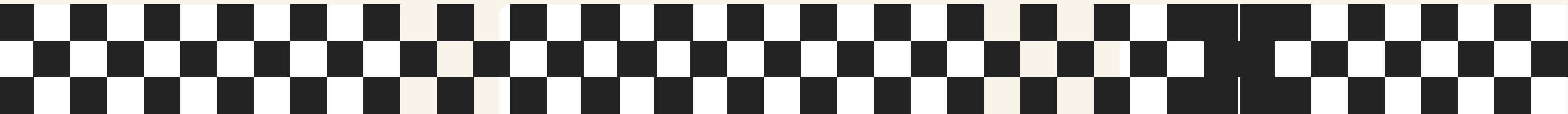


Origen

BIG DATA



- En el año 1989 **Erik Larson** utiliza por primera vez el término Big Data, en un artículo sobre el marketing y cómo se usarán los datos de los clientes, en los términos que actualmente conocemos.
- En 1997 Google lanza su motor de búsqueda y es el más usado, este acontecimiento favoreció a la creación de datos masivos.





BIG DATA

¿Qué es?

Es el manejo de grandes volúmenes de datos complejos que no pueden ser procesados con métodos tradicionales.

¿Cómo funciona?

Recolección de datos

Almacenamiento

Procesamiento

Análisis

Visualización e interpretación

¿Qué cantidades de datos maneja?

abarcan

Volumen

Cantidades masivas de datos

considerar

- Terabytes
- Petabytes
- Exabytes

Velocidad

Datos generados y procesados en tiempo real

Variedad

Datos estructurados, no estructurados y semiestructurados

como

- Texto
- Video
- Audio
- Sensores

Veracidad

Calidad y confiabilidad de los datos

Valor

Utilidad de los datos

para

Toma de decisiones

BIG DATA

¿Qué soluciona?

Datos masivos (estructurados y no estructurados)

Procesamiento en tiempo real (baja latencia)

Optimizar procesos

Escalabilidad horizontal

Identificación de patrones y anomalías

¿Para qué se usa?

Marketing y Ventas predictivos (machine learning/IA)

Turismo (satisfacción de clientes)

Médico (análisis prescriptivos)

Financiero (comportamientos y movimientos)

Toma de decisiones estratégicas



Herramientas de Software

Apache Hadoop

Almacenamiento y procesamiento masivo (batch).
Escalabilidad horizontal (de 1 servidor a miles de nodos).
Ecosistema robusto (HDFS, MapReduce, YARN).

Usado por: Facebook, LinkedIn, Netflix.

Apache Spark

100x más rápido que Hadoop (memoria RAM).
Soporta Python, Java, Scala, R.
Ideal para machine learning (MLlib) y stream processing.

Usado por: Uber, Airbnb, Databricks.

Elasticsearch

Indexación rápida de datos estructurados/no estructurados.
Integración con Kibana (visualización).
Escalable para logs, métricas y búsquedas.

Usado por: Netflix, eBay, Slack.

MongoDB

Flexibilidad con datos semiestructurados (JSON/BSON).
Escalabilidad horizontal automática.
Consultas rápidas y alta disponibilidad.

Usado por: Google, Adobe, Forbes.

EJEMPLO 1 MARKETING

- Analiza datos transaccionales, de navegación y sociales para entender preferencias del usuario.
- **Tecnologías:** Kafka para ingesta en tiempo real, Spark Streaming para procesamiento, y NoSQL (como Cassandra o MongoDB) para almacenamiento.



EJEMPLO 2 OPTIMIZACIÓN DE OPERACIONES



- Supervisa eventos en vivo para detectar anomalías o tomar decisiones automáticas
- **Tecnologías clave:** Apache Flink, Apache Storm, Elasticsearch con Kibana.



EJEMPLO 3

INVESTIGACIÓN CIENTÍFICA Y BIOINFORMÁTICA

- **Investigación Científica y Bioinformática:**
- Maneja y analiza datos genéticos, astronómicos o físicos en grandes volúmenes.
- **Tecnologías:** HPC (High-Performance Computing), Spark, Hadoop + herramientas especializadas como Apache Mahout.



**GRACIAS POR
SU ATENCIÓN**

