Actividad práctica — Fundamentos de Big Data

1. Concepto de Big Data y sus 5V

Definición:  
Big Data es el conjunto de técnicas y tecnologías que permiten procesar, almacenar y analizar enormes volúmenes de datos, en distintos formatos y a gran velocidad, con el fin de generar valor y conocimiento útil para la toma de decisiones.

Esquema de las 5V:

Volumen: Cantidades masivas de datos generados (ej: redes sociales generan millones de posts por minuto).

Velocidad: La rapidez con la que los datos son producidos y procesados (ej: sensores de autos autónomos que procesan datos en tiempo real).

Variedad: Diferentes tipos y formatos de datos (ej: imágenes médicas, texto de redes sociales, transacciones bancarias).

Veracidad: Calidad y fiabilidad de los datos (ej: validación de datos en registros de pacientes hospitalarios).

Valor: Capacidad de convertir datos en beneficios reales (ej: personalización de productos en Amazon).

2. Diferencias entre datos estructurados, semiestructurados y no estructurados

| Tipo de datos | Definición | Ejemplos | Herramientas |
| --- | --- | --- | --- |
| Estructurados | Organizados en tablas y filas con formato rígido. | Bases de datos SQL, registros bancarios. | MySQL, PostgreSQL. |
| Semiestructurados | Datos con cierto orden pero sin esquema fijo. | JSON, XML, correos electrónicos. | MongoDB, Cassandra. |
| No estructurados | Datos sin un formato definido ni esquema. | Videos, audios, imágenes, redes sociales. | Hadoop, Spark, Elasticsearch. |

Los más usados en Big Data son los semiestructurados y no estructurados, por su flexibilidad y volumen creciente.

3. Bases de datos relacionales vs. bases de datos distribuidas

Diagrama comparativo (texto):

Modelo de datos:

Relacionales → Tablas estructuradas con claves.

Distribuidas → Documentos, grafos o columnas flexibles.

Escalabilidad:

Relacionales → Escalan verticalmente (mejor hardware).

Distribuidas → Escalan horizontalmente (más servidores).

Consistencia:

Relacionales → Alta consistencia (ACID).

Distribuidas → Eventual consistencia (CAP theorem).

Casos de uso:

Relacionales → Sistemas bancarios, ERP.

Distribuidas → Redes sociales, IoT, comercio electrónico.

Ejemplos:

Relacionales: MySQL, PostgreSQL.

Distribuidas: Cassandra, MongoDB.

4. Introducción a la computación distribuida

Definición:  
La computación distribuida es un modelo en el que varias computadoras trabajan de forma coordinada para procesar grandes volúmenes de datos. Es clave en Big Data porque permite dividir la carga de trabajo en múltiples nodos, logrando rapidez, escalabilidad y tolerancia a fallos.

Tecnologías principales:

Hadoop: Framework que almacena datos en clústeres y los procesa en paralelo mediante MapReduce. Ideal para datos masivos y no estructurados.

Spark: Motor de procesamiento en memoria que ofrece análisis en tiempo real, machine learning y procesamiento de streams mucho más rápido que MapReduce.

5. Retos y oportunidades del Big Data en la industria

Sector elegido: Salud

Retos:

Garantizar la privacidad y seguridad de los datos médicos.

Integrar información proveniente de múltiples sistemas hospitalarios.

Oportunidades:

Mejorar diagnósticos mediante análisis predictivo de historiales clínicos.

Desarrollar medicina personalizada basada en datos genómicos.

Caso de uso innovador:  
Un sistema nacional de telemedicina inteligente que use Big Data e IoT para monitorear en tiempo real pacientes crónicos (ej: diabéticos, hipertensos), enviando alertas automáticas al médico cuando se detecten valores fuera de rango.