**Procesamiento de Grandes Bases de Datos**

**Actividad 1 - Framework**

Investigación sobre Apache Hadoop

**Introducción**

El crecimiento exponencial de los datos en la era digital obligó a desarrollar nuevas herramientas capaces de almacenar, procesar y analizar información a gran escala. En este contexto surge **Apache Hadoop**, un **framework** de código abierto que marcó un antes y un después en el procesamiento de grandes bases de datos.

**1. Origen y Contexto**

Apache Hadoop nació en el año 2006 dentro de la Fundación Apache como un proyecto de código abierto inspirado en las publicaciones técnicas de Google sobre el Google File System (**GFS**) y el modelo de programación MapReduce. Su creador inicial fue Doug Cutting, quien buscaba una herramienta capaz de procesar cantidades masivas de datos de forma distribuida y tolerante a fallos. El nombre *Hadoop* proviene del peluche de elefante amarillo del hijo de Cutting.

Hadoop se consolidó rápidamente como el pilar del ecosistema Big Data, utilizado por empresas tecnológicas de gran escala como Yahoo, Facebook, LinkedIn y posteriormente por industrias financieras, de telecomunicaciones y científicas.

**2. Arquitectura de Hadoop**

El ecosistema Hadoop se compone de cuatro módulos principales:

1. **HDFS (Hadoop Distributed File System):**
   * Divide los archivos en bloques de 64 o 128 MB.
   * Los distribuye en distintos nodos del clúster.
   * Cada bloque se replica 3 veces por defecto para asegurar redundancia.
2. **MapReduce:**

* Es el motor de procesamiento.
* Divide las tareas en dos fases:
  + **Map:** distribuye y procesa los datos localmente en cada nodo.
  + **Reduce:** combina los resultados parciales en un resultado global.

1. **YARN (Yet Another Resource Negotiator):**
   * Sistema de gestión de recursos y planificador de tareas.
   * Permite que múltiples aplicaciones se ejecuten de manera simultánea en el clúster.
2. **Common Utilities:**
   * Librerías y utilidades que soportan a los demás módulos.

**3. Ventajas**

* **Escalabilidad horizontal:** puede crecer añadiendo más nodos al clúster.
* **Tolerancia a fallos:** los datos se replican y se recuperan automáticamente en caso de fallos de hardware.
* **Flexibilidad:** permite almacenar datos estructurados (tablas), semiestructurados (XML, JSON) y no estructurados (imágenes, videos, texto plano).
* **Ecosistema amplio:** incluye herramientas como Hive (SQL sobre Hadoop), Pig (transformación de datos) y HBase (base de datos NoSQL).
* **Costos reducidos:** diseñado para funcionar sobre hardware convencional (*commodity hardware*).

**4. Limitaciones**

* **Baja eficiencia en tiempo real:** Hadoop está diseñado para procesamiento batch, por lo que no es apto para flujos de datos en vivo.
* **Latencia elevada:** el modelo MapReduce implica escribir y leer continuamente en disco, lo que ralentiza ciertos procesos.
* **Curva de aprendizaje:** su instalación y administración pueden ser complejas.

**5. Aplicaciones y Casos de Uso**

* **Empresas tecnológicas:** análisis de grandes volúmenes de clics y comportamiento de usuarios en plataformas web.
* **Sector financiero:** análisis de fraudes y detección de patrones en millones de transacciones históricas.
* **Salud:** almacenamiento y análisis de imágenes médicas y registros clínicos.
* **Ciencia e investigación:** procesamiento de datos provenientes de telescopios, sensores climáticos y experimentos genómicos.
* **Gobierno:** gestión de información censal y de seguridad pública a gran escala.

**6. Comparación con Frameworks Modernos**

Con el tiempo surgieron alternativas como **Apache Spark**, que procesa en memoria y ofrece mayor velocidad. Sin embargo, Hadoop sigue siendo relevante porque:

* Su **HDFS** continúa siendo uno de los sistemas de almacenamiento distribuido más confiables.
* Muchas organizaciones combinan **HDFS + Spark** para obtener lo mejor de ambos mundos.

**7. Conclusión**

Apache Hadoop es más que un **framework**: es el fundamento histórico y tecnológico del Big Data. Aunque en la actualidad existen soluciones más rápidas y ligeras, Hadoop sigue siendo utilizado globalmente por su robustez, capacidad de escalar y ecosistema maduro. Conocerlo es fundamental para comprender la evolución de las tecnologías de datos masivos y su impacto en los sectores productivos y científicos.