Spark

Unidad	Unidad 3
Entrega	Método de entrega de la actividad en el campus virtual

1. ¿Cuáles son los objetivos de la práctica?

1. **Integración de Apache Spark con Bases de Datos NoSQL:** Aprender a configurar Apache Spark para conectarse a bases de datos NoSQL y realizar operaciones de lectura y escritura.
2. **Procesamiento y Análisis de Datos del NFL Big Data Bowl:** Importar, almacenar y analizar el conjunto de datos utilizando Spark. Esto puede incluir limpieza de datos, transformaciones y cálculos estadísticos.
3. **Desarrollo de Competencias en Spark y NoSQL:** Mejorar las habilidades prácticas en el manejo de grandes volúmenes de datos y el uso de tecnologías de big data.
4. **Aplicar Buenas Prácticas de Desarrollo de Software:** Incluir documentación adecuada y pruebas unitarias para garantizar la calidad del código.

¿Qué es Spark?

Apache Spark es uno de los proyectos más populares en el ecosistema de Hadoop, y posiblemente, el proyecto de código abierto más desarrollado en Big Data. Aprovecha el procesamiento en memoria distribuido para realizar análisis de datos de forma eficiente y con un alto rendimiento. Es hasta 100 veces más rápido que MapReduce. Su simplicidad, rendimiento y flexibilidad lo han hecho popular no solo entre los científicos de datos, sino también entre los ingenieros, desarrolladores y otros perfiles interesados en Big Data.

Es un framework de programación distribuida, que ofrece un alto rendimiento, tanto para procesamiento por lotes, como para el interactivo. Tiene APIs para Java, Python, Scala y R, y tiene una cantidad significativa de proyectos relacionados. Se puede ejecutar aplicaciones Apache Spark localmente o distribuidas en un clúster.

2. Enunciado

1) **Preparación de los Datos:**
   a) Descargar el conjunto de datos del NFL Big Data Bowl 2024 de Kaggle.
   b) Explorar y entender la estructura y el contenido de los datos.
2) **Configuración del Entorno:**
   a) Instalar y configurar Apache Spark y las bases de datos NoSQL elegidas (MongoDB, InfluxDB, Cassandra, etc.).
3) **Carga y Almacenamiento de Datos:**
   a) Cargar los datos en Spark.
   b) Realizar cualquier limpieza o transformación de datos necesaria.
   c) Almacenar los datos procesados en la base de datos NoSQL seleccionada.

4) Preparar un reporte o presentación que describa las metodologías utilizadas.

3. Detalles de la entrega

1. Memoria en pdf que contenga:
    a. Capturas de pantalla y descripción de los pasos descritos anteriormente.
    b. Explicación de la configuración del sistema, el proceso de análisis de datos y las conclusiones.
2. Dockerfile o Docker-compose y scripts realizados