ACTIVIDAD #1:
Ejercicio: Taller de Big Data: Instalación y Prácticas

Instalación del Software:

Paso 1: Instalación de Hadoop

- Descarga e instala Hadoop desde el sitio web oficial.
- Sigue las instrucciones de instalación proporcionadas en la documentación oficial.
- Configura el entorno de Hadoop, incluyendo variables de entorno y archivos de configuración.

Paso 2: Instalación de Apache Spark

- Descarga e instala Apache Spark desde el sitio web oficial.
- Configura las variables de entorno para Spark.

Paso 3: Configuración de las Variables de Entorno en Linux/Mac

- Abre el archivo de configuración del perfil del usuario, como `.bashrc` o `.bash_profile`, ubicado en tu directorio de inicio.

- Añade las siguientes líneas al final del archivo, ajustando la ruta según donde hayas instalado Apache Spark:

```
export SPARK_HOME=/ruta/a/spark

export PATH=$PATH:$SPARK_HOME/bin
```

- Guarda el archivo y cierra el editor.

- Para aplicar los cambios, ejecuta el comando `source ~/.bashrc` o `source ~/.bash_profile` en la terminal.

Paso 4: Configuración de las Variables de Entorno en Windows:

- Ve a la configuración avanzada del sistema.
- Haz clic en "Variables de Entorno".
- En la sección "Variables de Usuario", haz clic en "Nuevo" y añade una nueva variable llamada `SPARK_HOME` con la ruta al directorio de instalación de Spark.
- En la variable `PATH`, añade `%SPARK_HOME%\bin`.
- Haz clic en "Aceptar" para guardar los cambios.

Paso 5: Prueba de la Instalación

Una vez que hayas configurado las variables de entorno, puedes probar la instalación ejecutando algunos comandos básicos de Spark en tu terminal:

- Abre una terminal o ventana de línea de comandos.
- Ejecuta el siguiente comando para iniciar el shell interactivo de Spark:

```
spark-shell
```

Paso 6: Instalación de MongoDB

- Descarga e instala MongoDB desde el sitio web oficial.
- Configura el entorno de MongoDB, incluyendo la configuración del servicio y la creación de una base de datos de prueba.

Prácticas Básicas:

Práctica 1: Manipulación de Datos con Hadoop

- Carga un conjunto de datos de ejemplo en el sistema de archivos distribuido de Hadoop (HDFS).
- Ejecuta algunas operaciones básicas de manipulación de datos, como contar líneas, filtrar datos, y calcular estadísticas simples utilizando comandos de Hadoop MapReduce.

Práctica 2: Procesamiento de Datos con Apache Spark:

- Carga los mismos datos de ejemplo en un DataFrame de Spark.
- Realiza algunas transformaciones y acciones básicas, como filtrado, agrupación y cálculo de estadísticas.
- Compara el rendimiento de Spark con el de Hadoop MapReduce para las mismas operaciones.

Práctica 3: Almacenamiento y Consulta de Datos con MongoDB:

- Importa los datos de ejemplo en una colección de MongoDB.
- Realiza consultas básicas para extraer información de la base de datos.
- Experimenta con índices y agregaciones para mejorar el rendimiento de las consultas.

Prácticas Avanzadas:

Práctica 4: Procesamiento de Datos en Tiempo Real con Apache Kafka y Spark Streaming:

- Configura un productor de datos para enviar datos de ejemplo a un tema de Kafka.
- Configura un consumidor de Kafka para procesar los datos en tiempo real utilizando Spark Streaming.
- Realiza análisis y procesamiento en tiempo real de los datos recibidos.

NOTA: Consulte https://www.paradigmadigital.com/dev/comunicacion-microservicios-apache-kafka/ para información similar.

Práctica 5: Análisis Predictivo con Machine Learning:

- Utiliza Spark MLlib para construir y entrenar un modelo de machine learning utilizando los datos de ejemplo.
- Evalúa el rendimiento del modelo utilizando métricas de evaluación adecuadas.
- Aplica el modelo para hacer predicciones sobre nuevos datos.

Consulte: https://learn.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-machine-learning-mllib-ipython