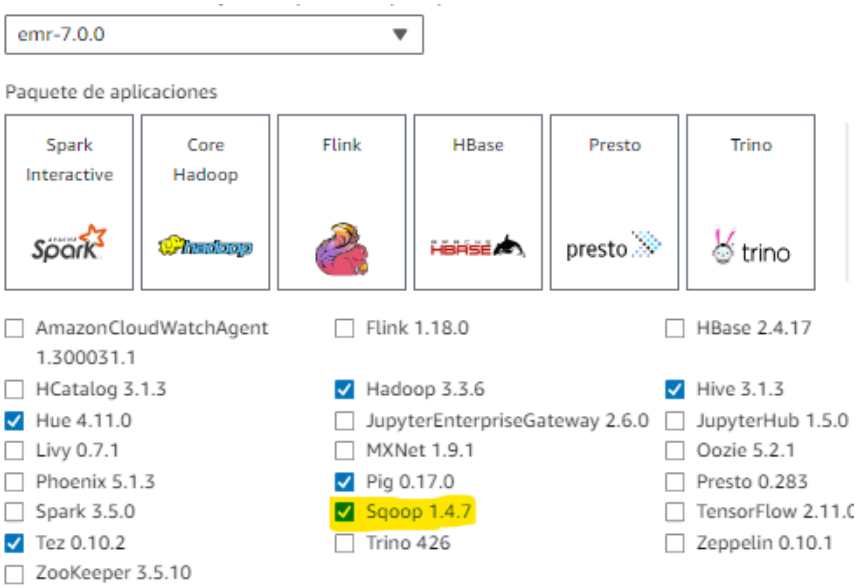
Enviar tareas como Pasos (*Steps*) a clústeres EMR.

CONTENIDO

APARTADO A

1.- Siguiendo los pasos que se explican en el tema, crea un clúster Hadoop EMR con 1 *master* y dos nodos. Selecciona la opción *Core Hadoop* (***versión 7.0.0***). No te olvides de seleccionar Sqoop, ya que lo utilizaremos en las prácticas siguientes.



INTRODUCCIÓN

- Utilizaremos el *dataset* https://www.kaggle.com/datasets/alimortezaie/online-retail.

CONTENIDO

APARTADO B

1. Viene en formato Excel. Desde el propio Excel puedes convertirlo a formato '*csv*'
2. Crea una carpeta con tu nombre en el directorio `user` del HDFS de EMR
3. Crea dentro de él una carpeta llamada `ventas` y sube a ella el '*csv*' que obtuviste anteriormente.

CONTENIDO

APARTADO C

USANDO Pig

1. Cargar los datos del *dataset* en PIG

Consultas:

1. ¿Cuántos registros tiene la tabla?

2. Mostrar registros con cantidades mayores o iguales a cero.
3. A partir de la consulta anterior, mostrar registros precio mayor a cero
4. A partir de la consulta anterior, mostrar solamente los registros con algún valor en el campo CustomerID.
5. ¿Cuántos registros tiene la última consulta?
6. Almacena la consulta final del punto 4 en un fichero llamado `ventas.csv` dentro de la carpeta de apartado **B**.

CONTENIDO

APARTADO D

Crear la tabla externa en Hive partiendo del fichero del punto anterior.

Conéctate al nodo maestro (SSH) o usa Hue.

1. Crear base de datos
2. Crear tabla externa sobre los datos RAW (CSV)
3. Hive no maneja muy bien el formato de fecha original, conviértelo a `d/M/yyyy H:mm`
4. Crea la misma estructura de tabla pero particionada por año y mes.
5. Inserta los registros del punto 1.2 en la tabla particionada.

CONTENIDO

APARTADO E

Consultas con **HIVE**:

Análisis de clientes

1. 10 clientes con mayor gasto total
2. Clientes con más compras (cantidad de facturas)

Análisis de productos

3. 10 productos más vendidos
4. 10 Productos más rentables (suma de precio unitario por cantidad)

Análisis geográfico

5. Países con mayor volumen de ventas

Análisis temporal

6. Ventas totales por mes (suma de precio unitario por cantidad)
7. Hora del día con más actividad

CONTENIDO

APARTADO F

SQOOP

1. En tu servidor MySql en la máquina EC2 crea una base de datos y una tabla para almacenar los datos del fichero `ventas.csv.`
2. Exporta con SQOOP los datos de `ventas.csv` a la tabla que creaste en el punto anterior.