Big Data en AWS - Laboratorio 7 - Procesamiento de datos de taxi de Nueva York usando Spark en EMR

En este laboratorio , utilizará Amazon EMR para ejecutar un trabajo de Spark que proceso el conjunto de datos 2015 New York Taxi.

El conjunto de datos completo incluye datos de los años 2009-2016 en todos los viajes de taxi en New York. El conjunto de datos utilizado en este laboratorio es un subconjunto de esos datos que utiliza solo en junio 2015. Hay aproximadamente 120 millones de registros en el conjunto de datos con formato csv que procesaremos en este laboratorio.

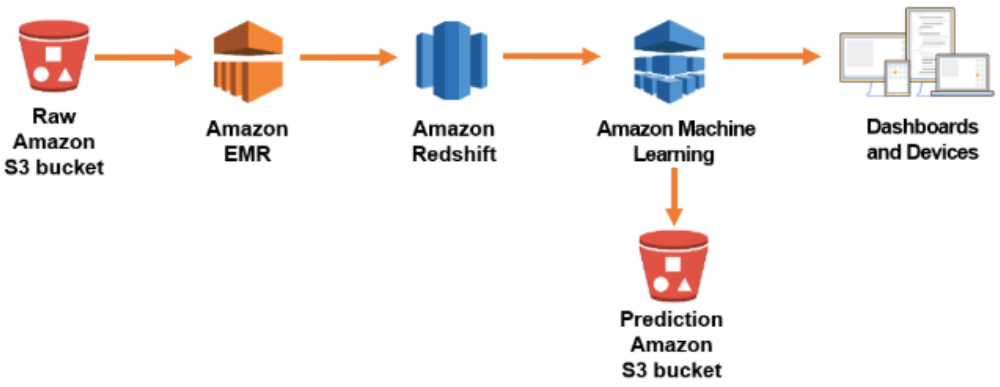
Su trabajo de Spark aprovecha la Api de Spark Dataframe para leer y procesar datos sucios. El trabajo de Spark también valida el formato y la integridad de los datos antes de guardar el conjunto de de datos en bucket de Amazon S3.

Los datos de salida se guardan en el mismo formato (*.csv) en un bucket de salida se guardan en un buckert de Amazon S3. Los datos procesados se cargan en Amazon Redshift para su posterior análisis.

Objetivos

Después de completar este laboratorio podrá:

- Use los pasos de Amazon EMR para procesar su aplicación Spark.
- Analice y procese conjuntos de datos con Spark en Amazon EMR.
- Importe los datos procesados a Amazon Redshift



En esta práctica de laboratorio, realizará la parte resaltada de la solución dada anteriormente. En este laboratorio, usted:

- Ejecuta Apache Spark en EMR de forma interactiva.
- Use Apache Spark para procesar datos de taxi de Nueva York.
- Use Apache Spark para transferir datos de Amazon EMR a Amazon Redshift directamente usando el conector Spark-Redshift

1. Acceso a la consola de administración de AWS

2. CARGUE UN SCRIPT DE PYTHON EN SU BUCKET DE AMAZON S3

En esta tarea, creará un bucket de Amazon S3 para almacenar los datos transformados y el script utilizado para procesar los datos.

A. Crear un depósito de salida de Amazon S3.

Antes de ejecutar la aplicación Spark, creará un bucket de Amazon S3 para contener el script de procesamiento de Python y los datos transformados.

Cada bucket de Amazon S3 requiere un nombre único, por lo que agregará un número aleatorio al nombre del bucket.

1. En la **consola de administración de AWS**, en el menú **Services,** haga clic en **S3.**
2. Hacer clic **Crear un cubo.**
3. Para el **nombre del depósito**, ingrese: **spark-lab-aquevedo**
4. Haz clic en **crear.**
5. Haz clic en nombre de tu cubo.
6. Hacer clic **Crear Carpeta.**
7. Crear una carpeta con el nombre: **Scripts**

B. Crear un depósito de salida de Amazon S3

Cargará un script de Python en su bucket de Amazon S3 para que el clúster de Amazon EMR acceda a él.

8. Haga clic derecho en este enlace y descargue el archivo a su computadora que está en el drive: **pypsark-lab7.py**

9. En la consola de administración S3, haga clic en la carpeta de secuencias de comandos.

10. Hacer clic Cargue, luego cargue el archivo que acaba de descargar.

3. **Ejecutar un trabajo Spark en Amazon EMR**
   En esta tarea, agregará un paso a su cluster de Amazon EMR para ejecutar su trabajo de Spark. Utilizará el script cargado en su bucket de Amazon S3 como una aplicación Spark. El script esta escrito en Python y usa la API Dataframe. El script eliminará puntos de datos erróneos del conjunto de datos, por ejemplo, algunos puntos de datos que se encuentran en el océano. El script procesa y conversa el conjunto de datos en su bucket de Amazon S3.

   11. En el menú **Servicios**, haga clic en **EMR**.
   12. En la página **Clúster**, haga clic en **labCluster**
   13. Haga clic en la pestaña **Pasos**, luego haga clic en **Agregar** paso.
   14. En el cuadro de diálogo **Agregar paso, configure:**
       a. **Tipo de paso:** aplicación Spark
       b. **Nombre:** New_York_Taxi_Preprocess
       c. **Ubicación de la aplicación:** s3://spark-lab-aquevedo/scripts/pyspark-lab6.py
       d. **Argumentos:** s3://spark-lab-123/output
   15. Haz clic en Agregar (Demora 5 minutos en completarse)
   16. Espere hasta que el trabajo tenga un estado de **completado.** Puedes hacer clic en actualizar cada 30 segundos para actualizar el estado.
   17. En el menú servicios, haga clic en S3.
   18. Haga clic en el bucket de Spark-lab y navega hasta la carpera de salida.
   19. Navegue a uno de los archivos de salida (comenzando con part-) y descárguelo a su computadora.
   20. Ver el archivo en un editor de texto como el Bloc de notas. Si el tiempo lo permite, puede descargar y abrir los otros archivos de salida.
       Estos datos transformados de su trabajo de Spark podrían usarse para crear visualizaciones como mapas de calor para las ubicaciones más ocupadas.

4. **Conectarse al nodo principal de Hadoop mediante SSH**
5. **Consultar datos de taxi de New York con Spark**
6. **Transferir datos en Amazon Redshift con Spark en Amazon EMR**
7. **Consultar los datos de Amazon Redshift**