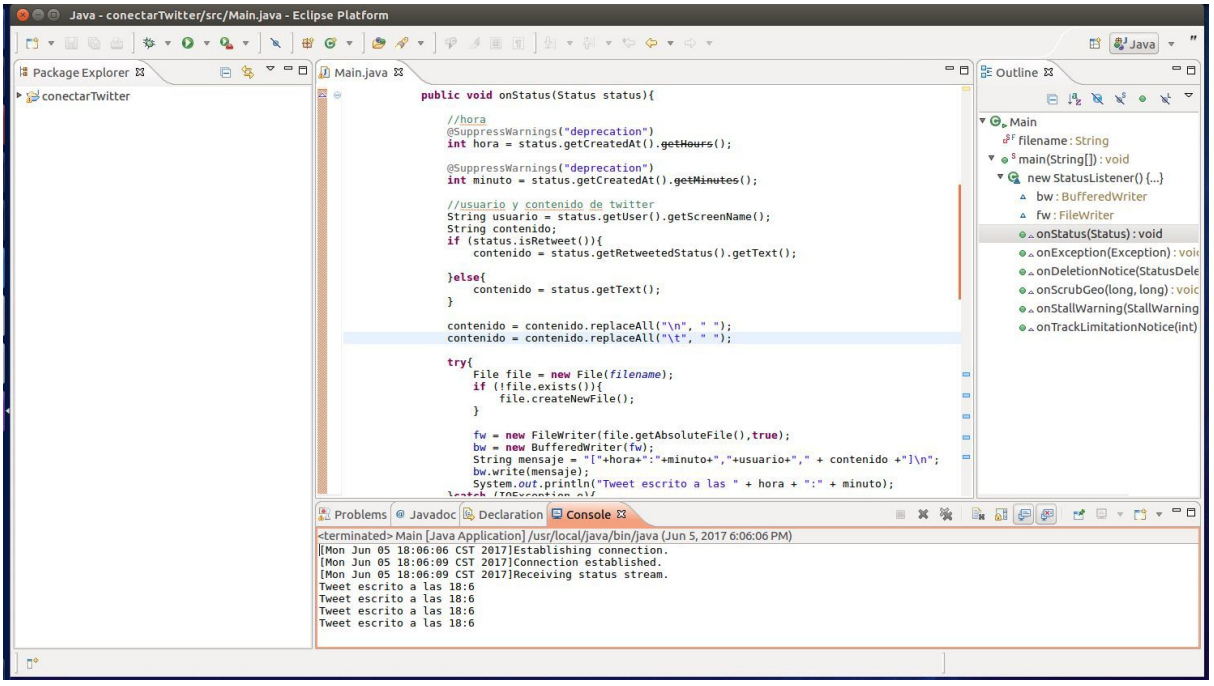
Instituto Tecnológico de Costa Rica
Escuela de Ingenieria en Computacion
Bases de datos ii
Grupo 40
Profesor Erick Hernandez
Estudiantes Jose Pablo Murillo, Carlos Villalobos, Esteban Santamaría
Manual de usuario proyecto 2: análisis de tweets

El siguiente manual de usuario se realizó en una máquina virtual corriendo Ubuntu 16.04 y Hadoop 2.7.

Obtención de datos de twitter

1. Abrir la aplicación en Eclipse.
2. Correr la aplicación presionando el botón verde en la barra superior de Eclipse.
3. La aplicación va a ir corriendo y registrando datos en “datos.txt”, que se encuentra en el mismo folder que el de la aplicación JAVA. La aplicación puede correr hasta que el usuario decide detenerla.

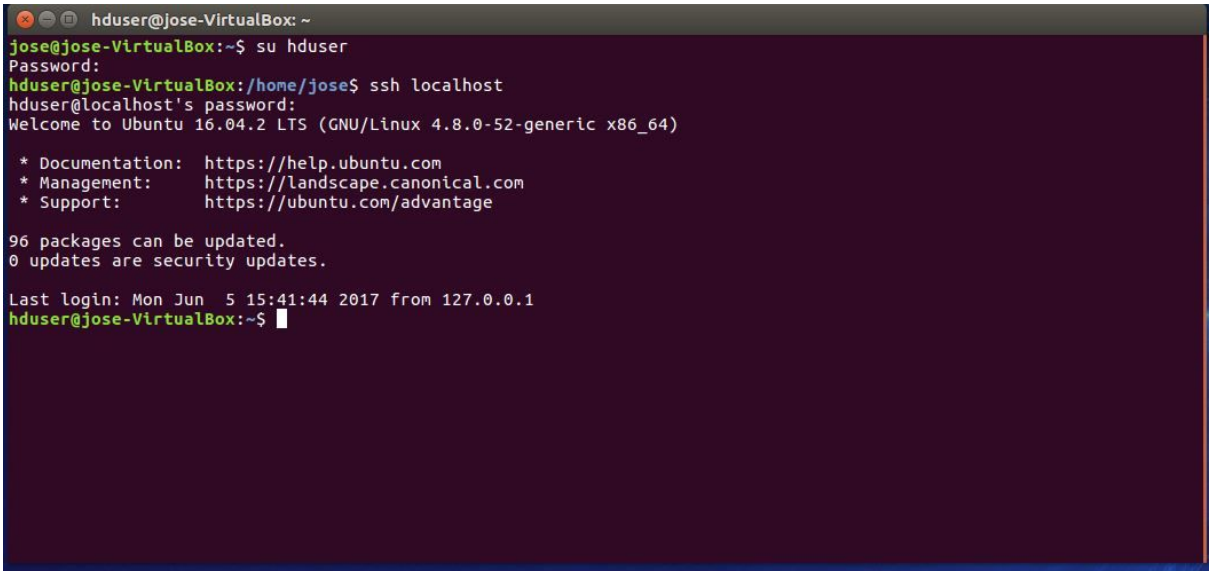


Ejecución del mapreduce usando Python

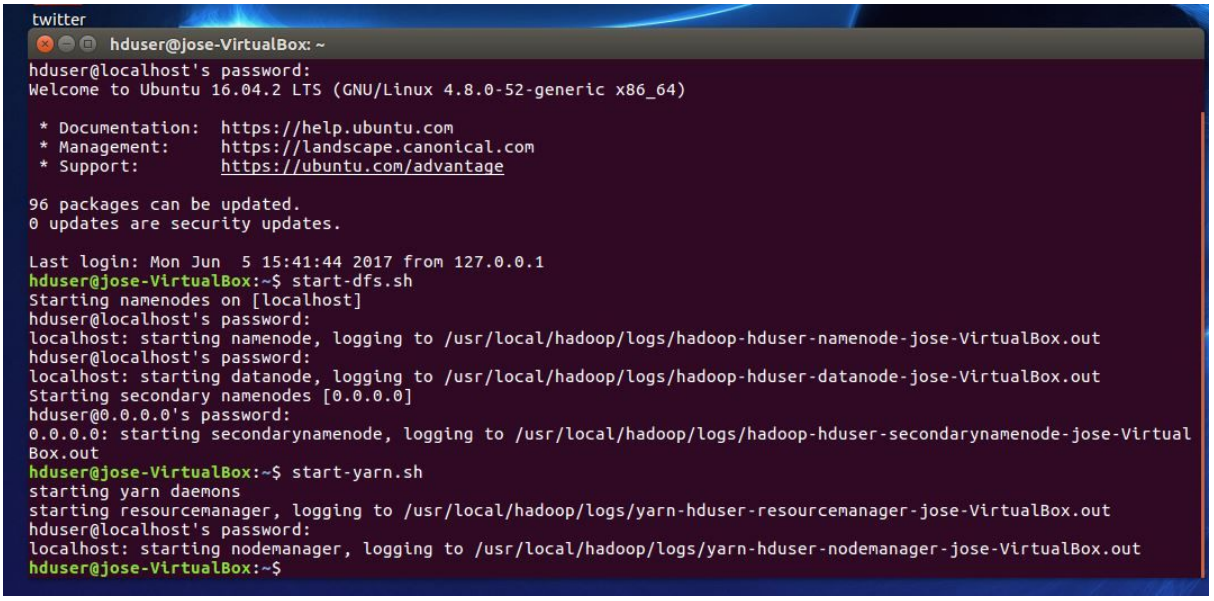
Para esta demostracion, se guardaron los archivos.py y se copio el archivo producto de correr la aplicacion JAVA “datos.txt” en el folder “twitter”. Por conveniencia se puso el archivo en el Desktop del usuario “jose”.

1. Ingrese como el usuario que fue definido para usar hadoop. En nuestro caso, se llama “hduser”.

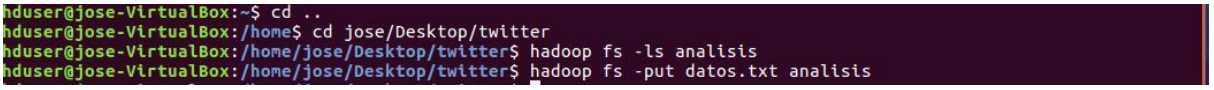
2. Inserte el comando ssh localhost para acceder al localhost de la máquina.



3. Ingrese los comandos start-dfs.sh y start-yarn.sh para inicializar los servicios para el servidor.



4. Dirijase a la carptera “twitter” donde se encuentran los archivos.py



5. Use los comandos “hadoop fs -mkdir analisis y hadoop fs -put datos.txt analisis para crear la carpeta donde se va a aguardar el archivo “datos.txt”.

NOTA: En caso de que le muestre un error con “ConnectionRefused” o “no such file or directory” puede ver la secuencia de pasos a seguir en el archivo “comandos”, incluido en los archivos de la aplicación.

6. Corra el siguiente comando: “$HADOOP_PREFIX/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -input analis/datos.txt -output output -mapper mapperTweets.py -reducer reducerTweets.py

En el comando anterior se especifa lo siguiente:

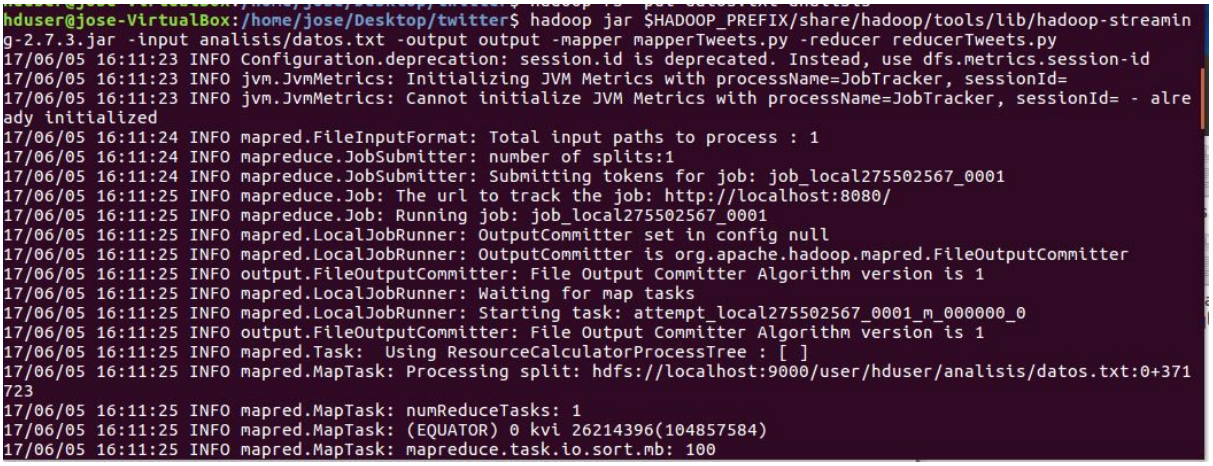
-input: archivo de datos que se va a usar para el map reduce.

-output: donde se guardan lo que retorne el reducer, en este caso nada.

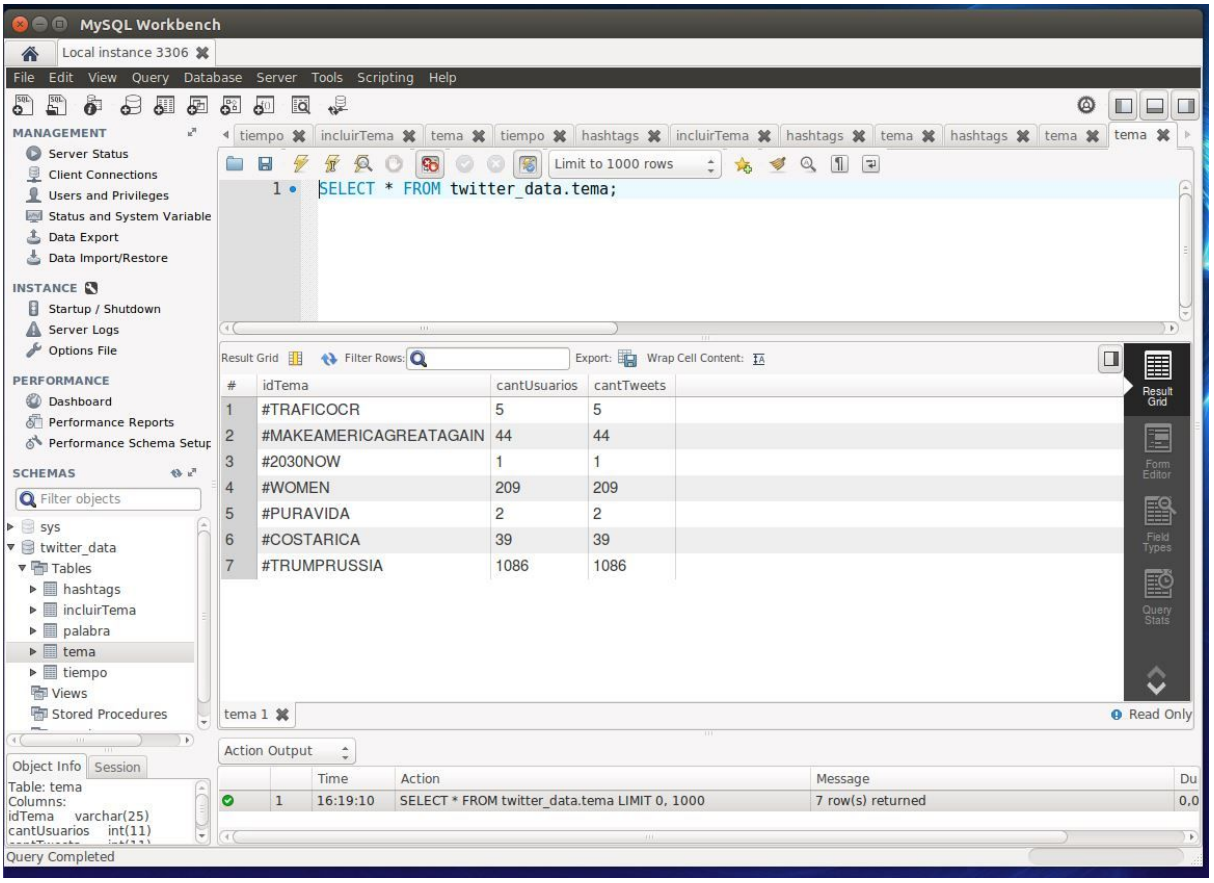
-mapper: el archivo mapper que va a mapear los datos

-reducer: el archivo reducer que va a analizar los datos pasados por el mapper.

Cuando presione enter, se va a comenzar el job como se muestra a continuación.



7. Una vez finalizado el job, se puede usar mySQL Workbench y examinar los resultados en las tablas.



8. Para visualizar los datos de manera gráfica, diríjase al navegador web e ingrese a la dirección “localhost”.

Se muestra una página donde se pueden realizar las siguientes consultas de datos para los temas que usaron para el desarrollo de la aplicación. Las acciones que se pueden hacer son las siguientes:

- Cantidad de usuarios que mencionaron el tema.
- Cantidad de tweets que fueron capturados por tema.
- El top 10 de palabras mencionadas por tema con su respectivo total de apariencias.
- La distribución de tiempo (en horas) dado un tema.
- Qué otros temas se mencionan para cierto tema y en cuantos tweets paso eso.

9. El usuario selecciona la accion y selecciona los temas sobre los que quire ver los datos.

10. Los datos son visualizados en un cuadro debajo de las opciones de los temas.