ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

NOMBRE DE ESTUDIANTE: Díaz Padilla Danny Sebastián

Laboratorio de:

ANALÍTICA DE DATOS – BIG DATA

Práctica Clase 20

Tema: Hadoop

Objetivos:

- Instalar la máquina virtual de Cloudera con Hadoop.
- Crear un esquema MapReduce en Python.
- Ejecutar MapReduce sobre un archivo de texto con el libro del quijote de la mancha.

Marco teórico:

Hadoop

Es una estructura de software de código abierto para almacenar datos y ejecutar aplicaciones en clústeres de hardware comercial. Proporciona almacenamiento masivo para cualquier tipo de datos, enorme poder de procesamiento y la capacidad de procesar tareas o trabajos concurrentes virtualmente ilimitados. [1]

Comandos

Listar todo el contenido en el directorio especificado:

hdfs dfs -ls /user/cloudera

Crear directorio en un sitio específicado.

hdfs dfs -mkdir /user/cloudera/input

Almacenar y poner un archivo local en el sistema de archivos Hadoop:

hdfs dfs -put quijote.txt /user/cloudera/input/

El siguiente commando utiliza una librería de map reduce llamada hadoop-streaming la cual permite realizar los trabajos de MapReduce y también con -input se agrega la entrada de los datos y con -output se establece el directorio de la salida del proceso.

hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar - input /user/cloudera/input -output /user/cloudera/output - mapper /quijote/mapper.py - reducer /quijote/reducer.py

Leer el contenido del archive resultante:

hdfs dfs cat /user/cloudera/output/part 00000 | head 1000

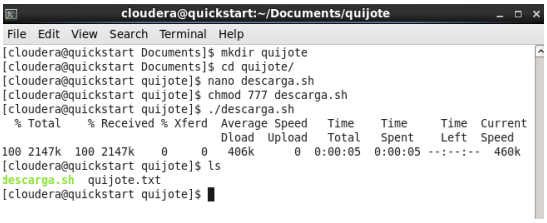
Desarrollo de la práctica:

Para el primer paso es crear un directorio llamado “quijote” y crear un archivo bash que permitirá descargar el libro del quijote.

El script contiene la siguiente línea de código:

curl http://www.gutenberg.org/cache/epub/2000/pg2000.txt -oquijote.txt

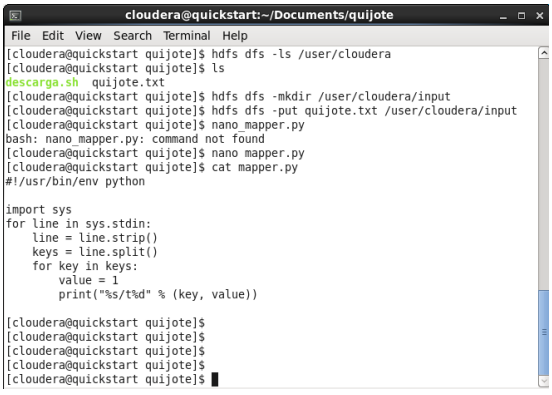
Posteriormente le agregamos permisos de ejecución y lo ejecutamos.



En la anterior figura se puede apreciar la descarga del fichero.

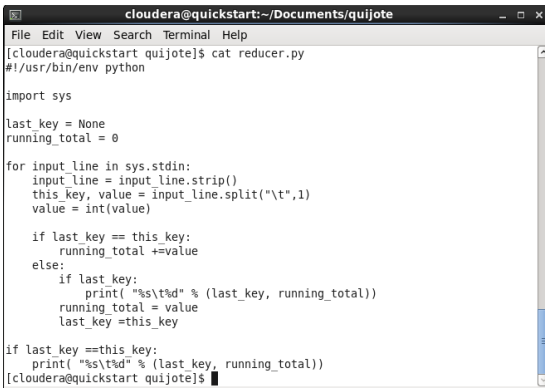
Luego preparamos el entorno de hadoop creando un directorio de entrada y poniendo el archivo de texto con el libro de el Quijote en ese directorio.

Se crea ahora el script que realizará el mapeo a todas las palabras, está hecho en python y tiene la siguiente estructura.



Se imprime el valor de 1 por cada palabra encontrada, es decir, se le da un puntaje a las palabras recibidas.

Para el archivo que realizará la reducción se crea otro fichero .py llamado reducer el cual obtiene un puntaje total de todas las palabras previamente procesadas por la función de mapeo.

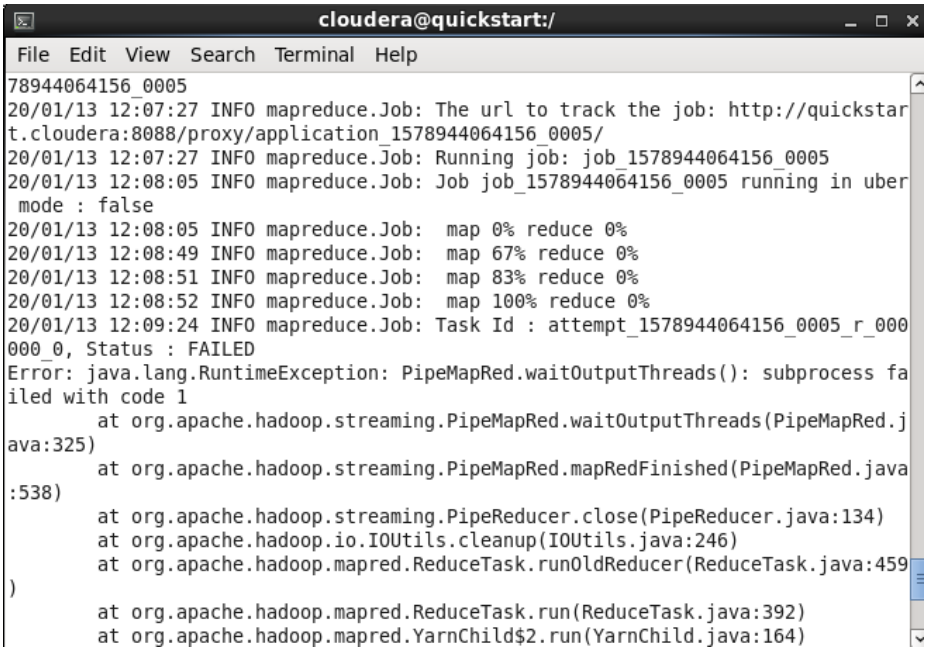


Mediante el comando:

hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar - input /user/cloudera/input -output /user/cloudera/output - mapper /quijote/mapper.py -reducer /quijote/reducer.py

se intenta realizar este procedimiento.

Sin embargo, aunque la función de mapeo y reducción alcanzar el 100% ambas tareas fallan.



El problema puede deberse a falta de compatibilidad o agotamiento de recursos.

En una ejecución normal, el resultado se podría ver como en la siguiente figura

```
[cloudera@quickstart ~]$ sudo -i
[root@quickstart ~]# cd ../..
[root@quickstart /]# cd quijote
[root@quickstart quijote]# nano mapper.py
[root@quickstart quijote]# nano reducer.py
[root@quickstart quijote]# hdfs dfs -ls /user/cloudera/output
[root@quickstart quijote]# hdfs dfs -ls /user/cloudera/output4
Found 2 items
-rw-r--r--   1 root cloudera          0 2020-01-08 16:51 /user/cloudera/output
_SUCCESS
-rw-r--r--   1 root cloudera    4005150 2020-01-08 16:51 /user/cloudera/output
part-00000
```

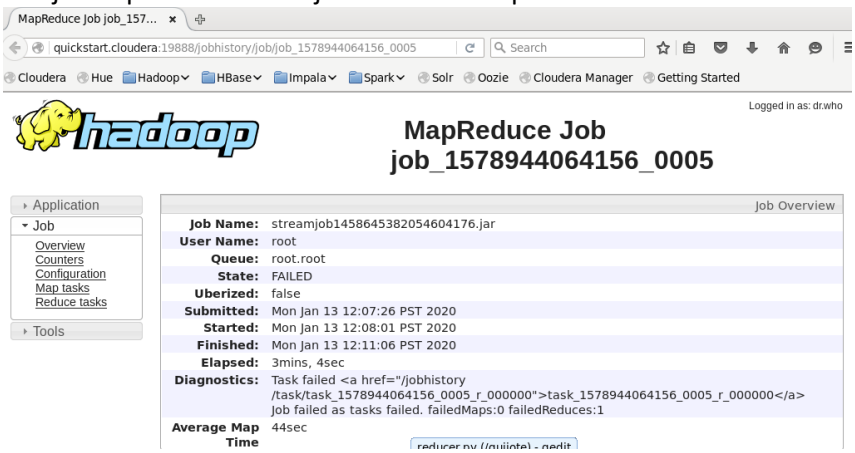
Y para mostrar todos los resultados se debe utilizar el comando: hdfs dfs -cat /user/cloudera/output/*

El cual muestra un conteo de cada palabra

```
única   2
única   3
única   4
única   5
única   5
única,  1
única,  1
única.  1
única.  1
único   1
único   2
único   3
único   4
único   5
único   6
único   7
único   8
único   9
único   10
único   11
único   11
```

Análisis de resultados:

El error generado parece ser ajeno a los archivos creados. Entrando al enlace proporcionado por el mismo "trabajo" se puede ver esa ejecución en tiempo real:



Sin embargo, no da una descripción específica del error y la línea de comandos solo señala líneas dentro de los scripts de hadoop.

Este problema puede ser causado por incompatibilidad o problemas de recursos ya que la máquina virtual ejecutada solicita 4GB de RAM y la máquina utilizada apenas cumple con el requisito

Conclusiones y recomendaciones:

- Se logró instalar la máquina virtual de Cloudera con Hadoop.
- Se creó un esquema MapReduce en Python.
- Se ejecutó MapReduce sobre un archivo de texto con el libro del quijote de la mancha, pero no se pudo cumplir la operación.
- El procesamiento distribuido otorgado por Hadoop permite aumentar la velocidad a la que se trata una cantidad gigante de datos.
- Hadoop proporciona un tratamiento del problema utilizando multihilos y todos los núcleos físicos disponibles en la máquina.

  - Se recomienda utilizar una computadora con un buen procesador y una cantidad de RAM mayor o igual a 8GB.
  - Para evitar problemas de permisos se recomienda ejecutar los comandos en un directorio fuera de root.

Bibliografía

[1]"¿Qué es Hadoop?", *Sas.com*. [Online]. Available: https://www.sas.com/es_pe/insights/big-data/hadoop.html. [Accessed: 13- Jan- 2020].