Roll No: J016(Avneesh Dubey) & J054(Aayush Talekar)

Aim: Word count using Hive

Objectives:

1. To run hive command.

2. Copy Data file from Local to HDFS.

3. Generate a Word count query.

4. Display Word count of the file.

Code & Output:

first make a txt folder (name random.txt) and put it in Hadoop dfs:

remember you are in that directory or not

hdfs dfs -put random.txt random1.txt

Now start hive

CREATE TABLE FILES (line STRING);

LOAD DATA INPATH 'random1.txt' OVERWRITE INTO TABLE FILES;

```
hive> CREATE TABLE FILES1 (line STRING);
OK
Time taken: 0.099 seconds
hive> LOAD DATA INPATH 'random2.txt' OVERWRITE INTO TABLE FILES1;
Loading data to table default.files1
chgrp: changing ownership of 'hdfs://quickstart.cloudera:8020/user/hive/warehous
e/files1/random2.txt': User does not belong to supergroup
Table default.files1 stats: [numFiles=1, numRows=0, totalSize=152, rawDataSize=0
]
OK
Time taken: 0.507 seconds
```

CREATE TABLE word_count AS

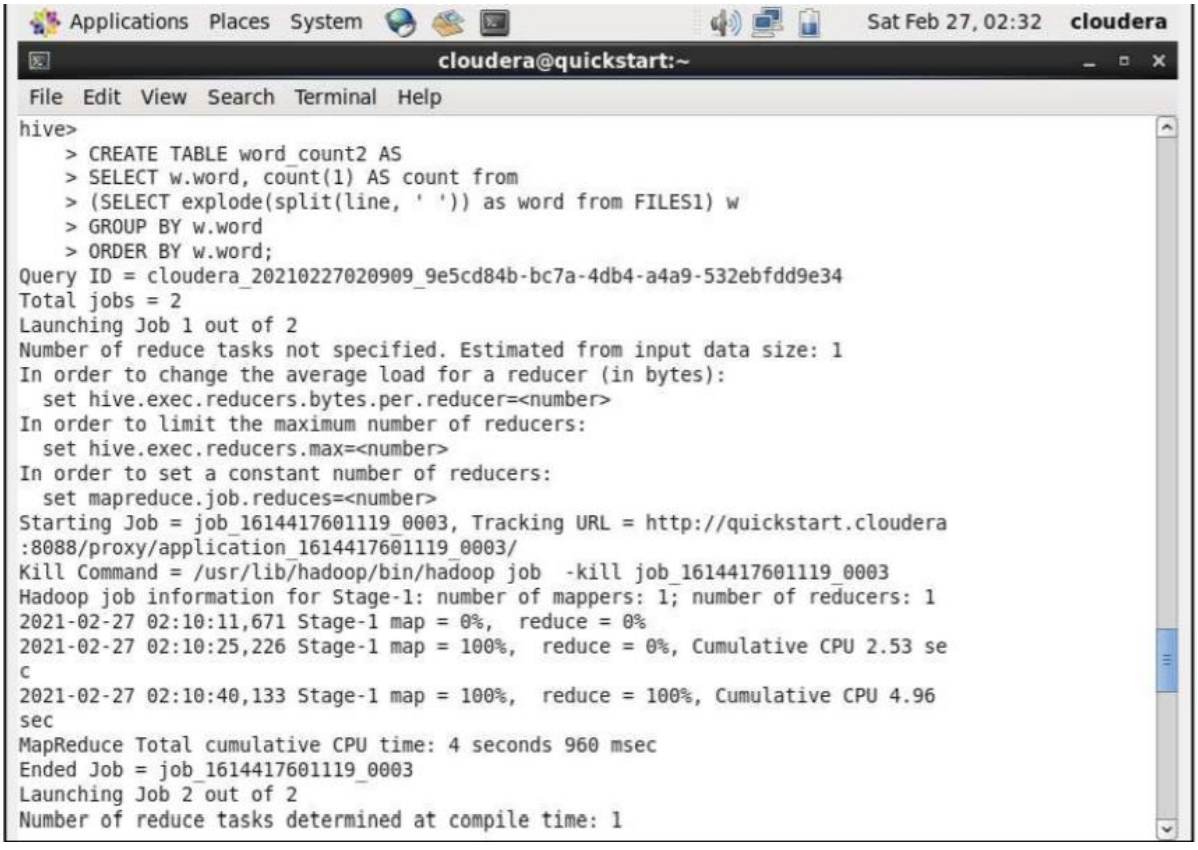
SELECT w.word, count(1) AS count from

(SELECT explode(split(line, ' ')) as word from FILES) w

GROUP BY w.word

ORDER BY w.word;

```
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1614417601119_0004, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1614417601119_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1614417601119_0004
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-02-27 02:10:57,260 Stage-2 map = 0%,  reduce = 0%
2021-02-27 02:11:11,249 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.69 se
c
2021-02-27 02:11:26,068 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.53
sec
MapReduce Total cumulative CPU time: 4 seconds 530 msec
Ended Job = job_1614417601119_0004
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/word_count2
Table default.word_count2 stats: [numFiles=1, numRows=24, totalSize=164, rawData
Size=140]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.96 sec   HDFS Read: 7622 HD
FS Write: 644 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.53 sec   HDFS Read: 5189 HD
FS Write: 240 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 490 msec
```




SELECT * FROM word_count;

```
Time taken: 93.042 seconds
hive> SELECT * FROM word_count2;
OK
.       8
2021    1
27th    1
BDT     1
Bye     1
February        1
It's    1
J078    1
My      1
Shah    2
Thank   1
This    2
Ujwal   2
a       1
command 1
has     1
hello   1
is      3
lab     1
name    1
no:     1
random  1
roll    1
you     1
Time taken: 0.08 seconds, Fetched: 24 row(s)
hive>
```