Lab 3

Roll No. : J070 – Saloni Jaitly, J072 – Saumya Nauni

Aim: Word Count Using Map Reduce

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

Codes:

//Map Reduce in HIVE

hive

CREATE TABLE FILES (line STRING);

LOAD DATA INPATH 'data1.txt' OVERWRITE INTO TABLE FILES;

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;

SELECT * FROM word_count ;

```
[cloudera@quickstart hive1]$ hadoop fs -put data.txt data1.txt
[cloudera@quickstart hive1]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> LOAD DATA INPATH 'data1.txt' OVERWRITE INTO TABLE FILES;
Loading data to table default.files
chgrp: changing ownership of 'hdfs://quickstart.cloudera:8020/user/hive/warehouse/files/data1.txt':
does not belong to supergroup
Table default.files stats: [numFiles=1, numRows=0, totalSize=50, rawDataSize=0]
OK
```

```
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_1614416156655_0002, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1614416156655_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1614416156655_0002
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-02-27 02:09:40,727 Stage-2 map = 0%,  reduce = 0%
2021-02-27 02:09:52,072 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.5 sec
2021-02-27 02:10:08,857 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.04 sec
MapReduce Total cumulative CPU time: 5 seconds 40 msec
Ended Job = job_1614416156655_0002
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/word_count
Table default.word_count stats: [numFiles=1, numRows=7, totalSize=54, rawDataSize=47]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.74 sec   HDFS Read: 7509 HDFS Write: 262 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.04 sec   HDFS Read: 4806 HDFS Write: 128 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 780 msec
OK
```

```
hive> 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;
Query ID = cloudera_20210227020808_9b7e516d-e93b-4849-8fc6-56dca2c70bbc
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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_1614416156655_0001, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1614416156655_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1614416156655_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-02-27 02:09:03,631 Stage-1 map = 0%,  reduce = 0%
2021-02-27 02:09:15,283 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.02 sec
2021-02-27 02:09:27,523 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.74 sec
MapReduce Total cumulative CPU time: 3 seconds 740 msec
Ended Job = job_1614416156655_0001
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
```

```
hive> SELECT * FROM word_count
    > ;
OK
This    2
a       2
hive    1
is      2
spark   1
tutorial        1
tutorial.       1
Time taken: 0.082 seconds, Fetched: 7 row(s)
```