LAB 2

MAP REDUCE IN HIVE

Vidhi Kapoor – J021

Kartikay Laddha – J025

To escape from safe mode - hdfs dfsadmin -safemode leave

[cloudera@quickstart Lab_1]$ hadoop fs -put WCFile.txt data1.txt

[cloudera@quickstart Lab_1]$ 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> CREATE TABLE FILES (line STRING);

FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. AlreadyExistsException(message:Table FILES already exists)

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': User does not belong to supergroup

Table default.files stats: [numFiles=1, numRows=0, totalSize=1364, rawDataSize=0]

OK

Time taken: 1.673 seconds

CREATING TABLE:

```
hive> CREATE TABLE word_count1 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_20210303230808_d66c8b3a-8370-4cff-afa5-e6624df016ed

```
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_1614834423086_0003, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1614834423086_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1614834423086_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-03 23:09:03,156 Stage-1 map = 0%,  reduce = 0%
2021-03-03 23:09:31,236 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.75 sec
2021-03-03 23:09:57,799 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.41 sec
MapReduce Total cumulative CPU time: 9 seconds 410 msec
Ended Job = job_1614834423086_0003
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_1614834423086_0004, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1614834423086_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1614834423086_0004
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-03-03 23:10:31,658 Stage-2 map = 0%,  reduce = 0%
2021-03-03 23:10:54,250 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.19 sec
```

2021-03-03 23:11:20,633 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 8.14 sec

MapReduce Total cumulative CPU time: 8 seconds 140 msec

Ended Job = job_1614834423086_0004

Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/word_count1

Table default.word_count1 stats: [numFiles=1, numRows=134, totalSize=1281, rawDataSize=1147]

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1   Cumulative CPU: 9.41 sec   HDFS Read: 8815 HDFS Write: 3540 SUCCESS

Stage-Stage-2: Map: 1 Reduce: 1   Cumulative CPU: 8.14 sec   HDFS Read: 8085 HDFS Write: 1359 SUCCESS

Total MapReduce CPU Time Spent: 17 seconds 550 msec

OK

Time taken: 181.247 seconds

EXECUTING HIVE QUERY:

hive> SELECT *FROM word_count1;

OK

	2

(	1

(HDFS),	1

/həˈduːp/)	1

All	1

Apache	2

Distributed	1

File	1

Hadoop	6

It	3

MapReduce	2

System	1

The	1

This	2

a	10

access 1

across 1

advantage 1

allows 1

also 1

amounts 1

and 7

approach 1

architecture 1

are 3

as 1

assumption 1

automatically 1

be 3

big 1

blocks 1

built 1

by 1

cluster. 1

clusters 2

code 1

collection 1

commodity 1

common 2

computation 1

computation. 1

computer 1

computers 1

consists 1

conventional 1

core 1

data 6

dataset 1

designed 2

distributed 2

distributes 1

efficiently 1

facilitates 1

failures 1

faster 1

file 1

files 1

for 2

found 1

framework 1

framework.[6] 1

from 1

fundamental 1

handled 1

hardware 1

hardware, 1

hardware.[4][5] 1

has 1

have 1

high-speed 1

higher-end 1

in 4

into 2

involving 1

is 3

it 1

known 1

large 1

locality,[7] 1

manipulate 1

many 1

massive1

model. 2

modules 1

more 2

network 1

networking.[8][9] 1

nodes 3

occurrences 1

of 8

on 2

open-source 1

originally 1

packaged 1

parallel 1

parallel.1

part 1

part, 1

problems 1

process 1

processed 1

processing 2

programming 2

provides 1

relies 1

should 1

since 1

software 2

solve 1

splits 1

still 1

storage 2

supercomputer 1

system 1

takes 1

than 1

that 3

the 7

them 1

then 1

they 1

to 3

to. 1

transfers 1

use 1

use.[3] 1

using 2

utilities 1

via 1

was 1

where 2

which 2

with 1

would 1

Time taken: 0.153 seconds, Fetched: 134 row(s)