﻿hduser@hduser-VirtualBox:~$ hadoop version

Hadoop 2.9.1

Subversion https://github.com/apache/hadoop.git -r e30710aea4e6e55e69372929106cf119af06fd0e

Compiled by root on 2018-04-16T09:33Z

Compiled with protoc 2.5.0

From source with checksum 7d6d2b655115c6cc336d662cc2b919bd

This command was run using /usr/local/hadoop-2.9.1/share/hadoop/common/hadoop-common-2.9.1.jar

hduser@hduser-VirtualBox:~$ cd /usr/local/hadoop-2.9.1/sbin

hduser@hduser-VirtualBox:/usr/local/hadoop-2.9.1/sbin$

hduser@hduser-VirtualBox:/usr/local/hadoop-2.9.1/sbin$ ./start-all.sh

This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh

Starting namenodes on [localhost]

localhost: starting namenode, logging to /usr/local/hadoop-2.9.1/logs/hadoop-hduser-namenode-hduser-VirtualBox.out

localhost: starting datanode, logging to /usr/local/hadoop-2.9.1/logs/hadoop-hduser-datanode-hduser-VirtualBox.out

Starting secondary namenodes [0.0.0.0]

0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop-2.9.1/logs/hadoop-hduser-secondarynamenode-hduser-VirtualBox.out

starting yarn daemons

starting resourcemanager, logging to /usr/local/hadoop-2.9.1/logs/yarn-hduser-resourcemanager-hduser-VirtualBox.out

localhost: starting nodemanager, logging to /usr/local/hadoop-2.9.1/logs/yarn-hduser-nodemanager-hduser-VirtualBox.out

hduser@hduser-VirtualBox:/usr/local/hadoop-2.9.1/sbin$ jps

8785 NodeManager

8662 ResourceManager

8472 SecondaryNameNode

9035 Jps

8267 DataNode

8142 NameNode

[hduser@hduser-VirtualBox:/usr/local/hadoop-2.9.1/sbin$](mailto:hduser@hduser-VirtualBox:/usr/local/hadoop-2.9.1/sbin$) cd

hduser@hduser-VirtualBox:~$ hadoop fs -put /home/hduser/Downloads/sharedfolder/airlines

hduser@hduser-VirtualBox:~$ hadoop fs -put /home/hduser/Downloads/sharedfolder/flights

﻿hive> SHOW DATABASES;

OK

default

Time taken: 2.131 seconds, Fetched: 1 row(s)

hive> create database airflight\_db ;

OK

Time taken: 0.34 seconds

hive> SHOW DATABASES;

OK

airflight\_db

default

Time taken: 0.119 seconds, Fetched: 2 row(s)

﻿hive> USE airflight\_db ;

OK

Time taken: 0.169 seconds

hive> CREATE external TABLE IF NOT EXISTS airlines (AirlineCode INT,Description STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/hduser/airlines' ;

OK

Time taken: 0.84 seconds

hive> SHOW TABLES ;

OK

airlines

Time taken: 0.153 seconds, Fetched: 1 row(s)

﻿hive> CREATE external TABLE IF NOT EXISTS flights (fDate TIMESTAMP, AirlineCode INT,FlightNum INT,Origin STRING,Destination STRING,DepartureTime STRING,DepartureDelay INT,ArrivalTime STRING,ArrivalDelay INT,Airtime INT,Distance INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/hduser/flights' ;

OK

Time taken: 0.464 seconds

hive>

> SHOW TABLES ;

OK

airlines

flights

Time taken: 0.075 seconds, Fetched: 2 row(s)

﻿hive> SELECT COUNT(\*) FROM airlines ;

Query ID = hduser\_20221227161409\_ed57c1c4-c93f-4815-b6f3-e4d99e9d2cdb

Total jobs = 1

Launching Job 1 out of 1

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\_1672136289846\_0003, Tracking URL = http://hduser-VirtualBox:8088/proxy/application\_1672136289846\_0003/

Kill Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job\_1672136289846\_0003

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2022-12-27 16:14:30,179 Stage-1 map = 0%, reduce = 0%

2022-12-27 16:14:41,705 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.3 sec

2022-12-27 16:14:52,910 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.94 sec

MapReduce Total cumulative CPU time: 3 seconds 940 msec

Ended Job = job\_1672136289846\_0003

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.94 sec HDFS Read: 50781 HDFS Write: 5 SUCCESS

Total MapReduce CPU Time Spent: 3 seconds 940 msec

OK

1579

Time taken: 44.424 seconds, Fetched: 1 row(s)

﻿hive> SELECT COUNT(\*) FROM flights ;

Query ID = hduser\_20221227161535\_cfb826a3-b012-4bc0-8689-fab37441bb81

Total jobs = 1

Launching Job 1 out of 1

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\_1672136289846\_0004, Tracking URL = http://hduser-VirtualBox:8088/proxy/application\_1672136289846\_0004/

Kill Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job\_1672136289846\_0004

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2022-12-27 16:15:47,972 Stage-1 map = 0%, reduce = 0%

2022-12-27 16:15:56,561 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.09 sec

2022-12-27 16:16:05,016 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.4 sec

MapReduce Total cumulative CPU time: 3 seconds 400 msec

Ended Job = job\_1672136289846\_0004

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.4 sec HDFS Read: 31667840 HDFS Write: 7 SUCCESS

Total MapReduce CPU Time Spent: 3 seconds 400 msec

OK

476881

Time taken: 30.436 seconds, Fetched: 1 row(s)

hive>

* Find count of flights that had arrival delay

﻿hive> SELECT count(FlightNum) FROM flights WHERE ArrivalDelay < 0;

Query ID = hduser\_20221227162328\_0d7564f2-67ca-430c-bb1a-a7ca93672a96

Total jobs = 1

Launching Job 1 out of 1

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\_1672136289846\_0007, Tracking URL = http://hduser-VirtualBox:8088/proxy/application\_1672136289846\_0007/

Kill Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job\_1672136289846\_0007

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2022-12-27 16:23:43,874 Stage-1 map = 0%, reduce = 0%

2022-12-27 16:23:57,942 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.18 sec

2022-12-27 16:24:08,395 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.04 sec

MapReduce Total cumulative CPU time: 7 seconds 40 msec

Ended Job = job\_1672136289846\_0007

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.04 sec HDFS Read: 31668824 HDFS Write: 7 SUCCESS

Total MapReduce CPU Time Spent: 7 seconds 40 msec

OK

279709

Time taken: 40.555 seconds, Fetched: 1 row(s)

* Find count of flights that had departure delay

﻿hive> SELECT count(FlightNum) FROM flights WHERE DepartureDelay < 0;

Query ID = hduser\_20221227162635\_1f202fce-197d-4d7b-aa82-e613a11c8be6

Total jobs = 1

Launching Job 1 out of 1

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\_1672136289846\_0008, Tracking URL = http://hduser-VirtualBox:8088/proxy/application\_1672136289846\_0008/

Kill Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job\_1672136289846\_0008

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2022-12-27 16:26:48,416 Stage-1 map = 0%, reduce = 0%

2022-12-27 16:26:58,928 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.55 sec

2022-12-27 16:27:07,330 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.14 sec

MapReduce Total cumulative CPU time: 5 seconds 140 msec

Ended Job = job\_1672136289846\_0008

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.14 sec HDFS Read: 31668826 HDFS Write: 7 SUCCESS

Total MapReduce CPU Time Spent: 5 seconds 140 msec

OK

273610

Time taken: 32.662 seconds, Fetched: 1 row(s)

hive>

* find the average distance travelled by a flight

﻿hive> SELECT FlightNum, round(AVG(Distance),1) as avg\_distance FROM flights group by FlightNum;

Query ID = hduser\_20221227163308\_9444a681-524d-425c-95ff-9f2e160f392d

Total jobs = 1

Launching Job 1 out of 1

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\_1672136289846\_0010, Tracking URL = http://hduser-VirtualBox:8088/proxy/application\_1672136289846\_0010/

Kill Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job\_1672136289846\_0010

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2022-12-27 16:33:23,006 Stage-1 map = 0%, reduce = 0%

2022-12-27 16:33:33,629 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.32 sec

2022-12-27 16:33:43,010 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.54 sec

MapReduce Total cumulative CPU time: 5 seconds 540 msec

Ended Job = job\_1672136289846\_0010

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.54 sec HDFS Read: 31669120 HDFS Write: 70103 SUCCESS

Total MapReduce CPU Time Spent: 5 seconds 540 msec

OK

FlightNum avg\_distance

1 1731.7

2 1610.7

3 1479.2

4 1681.9

5 1683.2

6 1814.9

7 1823.8

8 1855.3

9 1333.9

10 1183.3

11 1451.3

12 1822.3

14 2703.0

15 1720.8

16 1291.9

17 1836.4

18 2274.4

19 1685.0

20 1434.5

………………….. Mentioned start and end of data as large number of rows

﻿7413 162.0

7414 134.8

7415 135.3

7416 190.0

7420 257.0

7423 133.5

7426 255.0

7427 255.0

7428 150.0

7429 150.0

7430 150.0

7431 150.0

7432 272.0

7438 200.0

8406 482.0

Time taken: 35.399 seconds, Fetched: 6366 row(s)

hive>

* List the data that belong to the airline - American Airlines Inc

﻿hive> SELECT a.Description, f.AirlineCode, f. FlightNum, f.Origin, f.Destination, f.DepartureTime, f.DepartureDelay, f.ArrivalTime , f.ArrivalDelay, f.Airtime, f.Distance

> FROM flights f, airlines a

> Where a.Description = 'American Airlines Inc.'

> And a. AirlineCode = f. AirlineCode;

﻿Query ID = hduser\_20221227165001\_b8c21c72-51eb-44b0-9370-d0a5fb257f00

Total jobs = 1

Execution log at: /tmp/hduser/hduser\_20221227165001\_b8c21c72-51eb-44b0-9370-d0a5fb257f00.log

2022-12-27 16:50:10 Starting to launch local task to process map join; maximum memory = 518979584

2022-12-27 16:50:13 Dump the side-table for tag: 1 with group count: 1 into file: file:/tmp/hduser/f887c47e-7c64-46d7-8f98-f682f40f73ed/hive\_2022-12-27\_16-50-01\_456\_2499338088983086958-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile21--.hashtable

2022-12-27 16:50:13 Uploaded 1 File to: file:/tmp/hduser/f887c47e-7c64-46d7-8f98-f682f40f73ed/hive\_2022-12-27\_16-50-01\_456\_2499338088983086958-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile21--.hashtable (280 bytes)

2022-12-27 16:50:13 End of local task; Time Taken: 3.006 sec.

Execution completed successfully

MapredLocal task succeeded

Launching Job 1 out of 1

Number of reduce tasks is set to 0 since there's no reduce operator

Starting Job = job\_1672136289846\_0013, Tracking URL = http://hduser-VirtualBox:8088/proxy/application\_1672136289846\_0013/

Kill Command = /usr/local/hadoop-2.9.1/bin/hadoop job -kill job\_1672136289846\_0013 …..

﻿American Airlines Inc. 19805 356 DFW ATL 1332 27 1644 34 108 731

American Airlines Inc. 19805 357 DFW SAT 1525 10 1629 945 247

American Airlines Inc. 19805 357 SAT DFW 1710 10 1810 -5 40 247

American Airlines Inc. 19805 358 ORD LGA 1402 2 1659 -6 96 733

American Airlines Inc. 19805 359 LGA ORD 1904 14 2054 24 120 733

American Airlines Inc. 19805 360 ORD LGA 1515 20 1807 796 733

American Airlines Inc. 19805 362 DFW MSY 1611 101 1735 105 66 447

………. Mentioned mid and end of data as large number of rows

﻿American Airlines Inc. 19805 1533 MIA BOS 1345 0 1637 -18 152 1258

American Airlines Inc. 19805 1534 DFW DTW 1429 14 1750 0114 986

American Airlines Inc. 19805 1534 DTW DFW 1839 9 2019 -6 142 986

American Airlines Inc. 19805 1535 DFW MCI 1731 1 1906 671 460

American Airlines Inc. 19805 1535 MCI DFW 1938 -2 2118 -2 79 460

American Airlines Inc. 19805 1536 DFW SFO 1754 -1 1933 -12 200 1464

American Airlines Inc. 19805 1537 DFW RDU 0723 -7 1104 -6 134 1061

American Airlines Inc. 19805 1537 RDU DFW 1146 -4 1405 10 167 1061

Time taken: 46.555 seconds, Fetched: 43256 row(s)