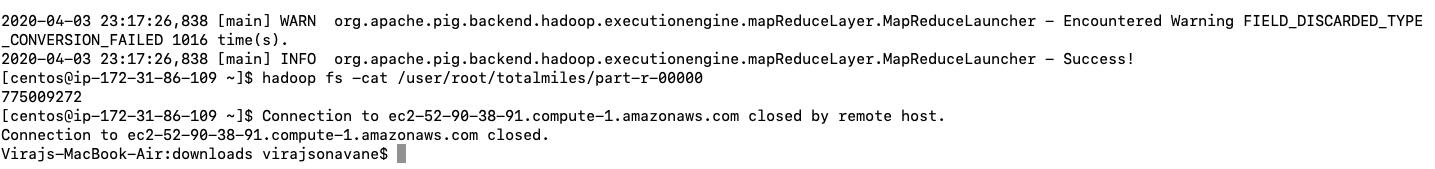
**Homework#6**

Viraj Sonavane



**In above Screenshot we are calculating total distance travelled by the flight.**

**Overall step followed while Installing Hadoop using AWS:**

Last login: Fri Apr 3 19:43:39 on ttys000

Virajs-MacBook-Air:~ virajsonavane$ ls

Applications Downloads Movies Public

Desktop Eclipse Music VirtualBox VMs

Documents Library Pictures \_\_c2j\_java\_\_

Virajs-MacBook-Air:~ virajsonavane$ cd downloads

Virajs-MacBook-Air:downloads virajsonavane$ sudo ssh -i "Vikey.pem" centos@ec2-52-90-38-91.compute-1.amazonaws.com

Password:

The authenticity of host 'ec2-52-90-38-91.compute-1.amazonaws.com (52.90.38.91)' can't be established.

ECDSA key fingerprint is SHA256:doOGiUAJpEIrMC9jCBqikuLoiEnPAvw4QPcPZHwjHGA.

Are you sure you want to continue connecting (yes/no)? yes

Warning: Permanently added 'ec2-52-90-38-91.compute-1.amazonaws.com,52.90.38.91' (ECDSA) to the list of known hosts.

Last login: Fri Apr 3 22:24:01 2020 from 132.241.160.76

[centos@ip-172-31-86-109 ~]$ sudo -u hdfs hadoop namenode -format

20/04/03 22:45:52 INFO namenode.NameNode: STARTUP\_MSG:

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

STARTUP\_MSG: Starting NameNode

STARTUP\_MSG: host = ip-172-31-86-109.ec2.internal/172.31.86.109

STARTUP\_MSG: args = [-format]

STARTUP\_MSG: version = 1.0.1

STARTUP\_MSG: build = -r ; compiled by 'jenkins' on Tue Mar 20 12:08:58 EDT 2012

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

20/04/03 22:45:52 INFO util.GSet: VM type = 64-bit

20/04/03 22:45:52 INFO util.GSet: 2% max memory = 19.33375 MB

20/04/03 22:45:52 INFO util.GSet: capacity = 2^21 = 2097152 entries

20/04/03 22:45:52 INFO util.GSet: recommended=2097152, actual=2097152

20/04/03 22:45:53 INFO namenode.FSNamesystem: fsOwner=hdfs

20/04/03 22:45:53 INFO namenode.FSNamesystem: supergroup=supergroup

20/04/03 22:45:53 INFO namenode.FSNamesystem: isPermissionEnabled=false

20/04/03 22:45:53 INFO namenode.FSNamesystem: dfs.block.invalidate.limit=100

20/04/03 22:45:53 INFO namenode.FSNamesystem: isAccessTokenEnabled=false accessKeyUpdateInterval=0 min(s), accessTokenLifetime=0 min(s)

20/04/03 22:45:53 INFO namenode.NameNode: Caching file names occuring more than 10 times

20/04/03 22:45:53 INFO common.Storage: Image file of size 110 saved in 0 seconds.

20/04/03 22:45:53 INFO common.Storage: Storage directory /var/lib/hadoop/cache/hadoop/dfs/name has been successfully formatted.

20/04/03 22:45:53 INFO namenode.NameNode: SHUTDOWN\_MSG:

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

SHUTDOWN\_MSG: Shutting down NameNode at ip-172-31-86-109.ec2.internal/172.31.86.109

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

[centos@ip-172-31-86-109 ~]$ for i in hadoop-namenode hadoop-datanode ; do sudo service $i start ; done

Starting hadoop-namenode (via systemctl): [ OK ]

Starting hadoop-datanode (via systemctl): [ OK ]

[centos@ip-172-31-86-109 ~]$ sudo -u hdfs hadoop fs -ls /

Found 1 items

drwxr-xr-x - mapred supergroup 0 2020-04-03 22:46 /var

[centos@ip-172-31-86-109 ~]$ sudo -u hdfs hadoop fs -mkdir /user/$USER

[centos@ip-172-31-86-109 ~]$ sudo -u hdfs hadoop fs -chown $USER /user/$USER

[centos@ip-172-31-86-109 ~]$ pig

2020-04-03 22:47:09,189 [main] INFO org.apache.pig.Main - Logging error messages to: /home/centos/pig\_1585968429186.log

2020-04-03 22:47:09,444 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:8020

2020-04-03 22:47:09,989 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021

grunt> fs -mkdir tmp

grunt> quit

[centos@ip-172-31-86-109 ~]$ cd

[centos@ip-172-31-86-109 ~]$ $scp -i ~/Downloads/Vikey.pem ~/Downloads/1987.csv centos@ec2-52-90-38-91.compute-1.amazonaws.com:~/data/

-bash: -i: command not found

[centos@ip-172-31-86-109 ~]$ scp -i ~/Downloads/Vikey.pem ~/Downloads/1987.csv centos@ec2-52-90-38-91.compute-1.amazonaws.com:~/data/

Warning: Identity file /home/centos/Downloads/Vikey.pem not accessible: No such file or directory.

The authenticity of host 'ec2-52-90-38-91.compute-1.amazonaws.com (172.31.86.109)' can't be established.

ECDSA key fingerprint is SHA256:doOGiUAJpEIrMC9jCBqikuLoiEnPAvw4QPcPZHwjHGA.

ECDSA key fingerprint is MD5:d5:8f:38:5f:e4:c4:a3:40:a8:f6:7f:9a:15:43:98:69.

Are you sure you want to continue connecting (yes/no)? yes

Warning: Permanently added 'ec2-52-90-38-91.compute-1.amazonaws.com,172.31.86.109' (ECDSA) to the list of known hosts.

Permission denied (publickey,gssapi-keyex,gssapi-with-mic).

lost connection

[centos@ip-172-31-86-109 ~]$ sudo scp -i ~/Downloads/Vikey.pem ~/Downloads/1987.csv centos@ec2-52-90-38-91.compute-1.amazonaws.com:~/data/

Warning: Identity file /home/centos/Downloads/Vikey.pem not accessible: No such file or directory.

The authenticity of host 'ec2-52-90-38-91.compute-1.amazonaws.com (172.31.86.109)' can't be established.

ECDSA key fingerprint is SHA256:doOGiUAJpEIrMC9jCBqikuLoiEnPAvw4QPcPZHwjHGA.

ECDSA key fingerprint is MD5:d5:8f:38:5f:e4:c4:a3:40:a8:f6:7f:9a:15:43:98:69.

Are you sure you want to continue connecting (yes/no)? yes

Warning: Permanently added 'ec2-52-90-38-91.compute-1.amazonaws.com,172.31.86.109' (ECDSA) to the list of known hosts.

Permission denied (publickey,gssapi-keyex,gssapi-with-mic).

lost connection

[centos@ip-172-31-86-109 ~]$ quit

-bash: quit: command not found

[centos@ip-172-31-86-109 ~]$ exit

logout

Connection to ec2-52-90-38-91.compute-1.amazonaws.com closed.

Virajs-MacBook-Air:downloads virajsonavane$ sudo ssh -i "Vikey.pem" centos@ec2-52-90-38-91.compute-1.amazonaws.com

Password:

Sorry, try again.

Password:

Sorry, try again.

Password:

sudo: 3 incorrect password attempts

Virajs-MacBook-Air:downloads virajsonavane$

Virajs-MacBook-Air:downloads virajsonavane$

Virajs-MacBook-Air:downloads virajsonavane$ scp -i ~/Downloads/Vikey.pem ~/Downloads/1987.csv centos@ec2-52-90-38-91.compute-1.amazonaws.com:~/data/

The authenticity of host 'ec2-52-90-38-91.compute-1.amazonaws.com (52.90.38.91)' can't be established.

ECDSA key fingerprint is SHA256:doOGiUAJpEIrMC9jCBqikuLoiEnPAvw4QPcPZHwjHGA.

Are you sure you want to continue connecting (yes/no)? yes

Warning: Permanently added 'ec2-52-90-38-91.compute-1.amazonaws.com,52.90.38.91' (ECDSA) to the list of known hosts.

@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

@ WARNING: UNPROTECTED PRIVATE KEY FILE! @

@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

Permissions 0644 for '/Users/virajsonavane/Downloads/Vikey.pem' are too open.

It is required that your private key files are NOT accessible by others.

This private key will be ignored.

Load key "/Users/virajsonavane/Downloads/Vikey.pem": bad permissions

centos@ec2-52-90-38-91.compute-1.amazonaws.com: Permission denied (publickey,gssapi-keyex,gssapi-with-mic).

lost connection

Virajs-MacBook-Air:downloads virajsonavane$ sudo scp -i ~/Downloads/Vikey.pem ~/Downloads/1987.csv centos@ec2-52-90-38-91.compute-1.amazonaws.com:~/data/

Password:

scp: /home/centos/data/: Is a directory

Virajs-MacBook-Air:downloads virajsonavane$ sudo scp -i ~/Downloads/Vikey.pem ~/Downloads/1987.csv centos@ec2-52-90-38-91.compute-1.amazonaws.com:~/

1987.csv 100% 121MB 1.1MB/s 01:54

Virajs-MacBook-Air:downloads virajsonavane$ hadoop fs -copyFromLocal 1987.csv /user/centos

-bash: hadoop: command not found

Virajs-MacBook-Air:downloads virajsonavane$ sudo ssh -i "Vikey.pem" centos@ec2-3-89-133-111.compute-1.amazonaws.com

Password:

^Z

[1]+ Stopped sudo ssh -i "Vikey.pem" centos@ec2-3-89-133-111.compute-1.amazonaws.com

Virajs-MacBook-Air:downloads virajsonavane$ sudo ssh -i "Vikey.pem" centos@ec2-52-90-38-91.compute-1.amazonaws.com

Last login: Fri Apr 3 22:45:29 2020 from 132.241.160.76

[centos@ip-172-31-86-109 ~]$ hadoop fs -copyFromLocal 1987.csv /user/centos

[centos@ip-172-31-86-109 ~]$ vi totalmiles.pig

[centos@ip-172-31-86-109 ~]$ pig totalmiles.pig

2020-04-03 23:16:21,669 [main] INFO org.apache.pig.Main - Logging error messages to: /home/centos/pig\_1585970181666.log

2020-04-03 23:16:21,971 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:8020

2020-04-03 23:16:22,501 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021

2020-04-03 23:16:22,995 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP\_BY

2020-04-03 23:16:23,150 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false

2020-04-03 23:16:23,158 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosing to move algebraic foreach to combiner

2020-04-03 23:16:23,183 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1

2020-04-03 23:16:23,183 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1

2020-04-03 23:16:23,270 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job

2020-04-03 23:16:23,297 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3

2020-04-03 23:16:23,301 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job8464260311784796783.jar

2020-04-03 23:16:25,348 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job8464260311784796783.jar created

2020-04-03 23:16:25,365 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job

2020-04-03 23:16:25,458 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.

\*\*\*\*hdfs://localhost:8020/user/centos/1987.csv

2020-04-03 23:16:25,774 [Thread-4] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1

2020-04-03 23:16:25,774 [Thread-4] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

2020-04-03 23:16:25,784 [Thread-4] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 2

2020-04-03 23:16:25,964 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete

2020-04-03 23:16:26,716 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job\_202004032243\_0001

2020-04-03 23:16:26,716 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More information at: http://localhost:50030/jobdetails.jsp?jobid=job\_202004032243\_0001

2020-04-03 23:16:48,129 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 10% complete

2020-04-03 23:16:51,167 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 20% complete

2020-04-03 23:16:54,217 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 26% complete

2020-04-03 23:16:57,272 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 40% complete

2020-04-03 23:17:00,295 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete

2020-04-03 23:17:26,816 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete

2020-04-03 23:17:26,821 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features

1.0.1 0.9.2 centos 2020-04-03 23:16:23 2020-04-03 23:17:26 GROUP\_BY

Success!

Job Stats (time in seconds):

JobId Maps Reduces MaxMapTime MinMapTIme AvgMapTime MaxReduceTime MinReduceTime AvgReduceTime Alias Feature Outputs

job\_202004032243\_0001 2 1 24 24 24 18 18 18 milage\_recs,records,tot\_miles GROUP\_BY,COMBINER /user/root/totalmiles,

Input(s):

Successfully read 1311827 records (127167761 bytes) from: "/user/centos/1987.csv"

Output(s):

Successfully stored 1 records (10 bytes) in: "/user/root/totalmiles"

Counters:

Total records written : 1

Total bytes written : 10

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

Job DAG:

job\_202004032243\_0001

2020-04-03 23:17:26,838 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD\_DISCARDED\_TYPE\_CONVERSION\_FAILED 1016 time(s).

2020-04-03 23:17:26,838 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!

[centos@ip-172-31-86-109 ~]$ hadoop fs -cat /user/root/totalmiles/part-r-00000

775009272

[centos@ip-172-31-86-109 ~]$ Connection to ec2-52-90-38-91.compute-1.amazonaws.com closed by remote host.

Connection to ec2-52-90-38-91.compute-1.amazonaws.com closed.

Virajs-MacBook-Air:downloads virajsonavane$