

CS451 Introduction to Parallel and Distributed Computing - Assignment 5

Batkhishig Dulamsurankhor - A20543498

November 22, 2024

Part 1

- 14) a. i. yarn node -list

```
2024 Fall — hadoop@bdulamsurankhor-1:~/hadoop-3.4.0/sbin — ssh cc@129.114.26.216 — 127x24
...— ssh cc@129.114.26.216 ...-3.4.0/etc/hadoop — zsh ...4.0/etc/hadoop — zsh ...4.0/etc/hadoop — zsh ...
Starting namenodes on [bdulamsurankhor-1.novalocal]
bdulamsurankhor-1.novalocal: Warning: Permanently added 'bdulamsurankhor-1.novalocal' (ED25519) to the list of known hosts.
Starting datanodes
bdulamsurankhor-4.novalocal: Warning: Permanently added 'bdulamsurankhor-4.novalocal' (ED25519) to the list of known hosts.
bdulamsurankhor-2.novalocal: Warning: Permanently added 'bdulamsurankhor-2.novalocal' (ED25519) to the list of known hosts.
bdulamsurankhor-3.novalocal: Warning: Permanently added 'bdulamsurankhor-3.novalocal' (ED25519) to the list of known hosts.
Starting secondary namenodes [bdulamsurankhor-1.novalocal]
[hadoop@bdulamsurankhor-1 sbin]$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
[hadoop@bdulamsurankhor-1 sbin]$ ./mr-jobhistory-daemon.sh start historyserver
WARNING: Use of this script to start the MR JobHistory daemon is deprecated.
WARNING: Attempting to execute replacement "mapred --daemon start" instead.
[hadoop@bdulamsurankhor-1 sbin]$ yarn node -list
2024-11-22 17:41:56,979 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at bdulamsurankhor-1.novalocal/10.56.3.239:8032
Total Nodes:4
Node-Id          Node-State Node-Http-Address      Number-of-Running-Containers
bdulamsurankhor-3.novalocal:43941      RUNNING bdulamsurankhor-3.novalocal:8042          0
bdulamsurankhor-2.novalocal:44357      RUNNING bdulamsurankhor-2.novalocal:8042          0
bdulamsurankhor-4.novalocal:42791      RUNNING bdulamsurankhor-4.novalocal:8042          0
bdulamsurankhor-1.novalocal:46219      RUNNING bdulamsurankhor-1.novalocal:8042          0
[hadoop@bdulamsurankhor-1 sbin]$ 
[hadoop@bdulamsurankhor-1 sbin]$
```

Part 2

- 4) Verify that the file was uploaded to HDFS by hadoop fs -ls /data

```
● ● ● 2024 Fall — hadoop@bdulamsurankhor-1:~ — ssh cc@129.114.26.216 — 127x24
...— ssh cc@129.114.26.216 ...-3.4.0/etc/hadoop -- zsh ....4.0/etc/hadoop -- zsh ... ....4.0/etc/hadoop -- zsh ...
[[hadoop@bdulamsurankhor-1 ~]$ hadoop fs -ls /data
Found 1 items
-rw-r--r-- 2 hadoop supergroup 231149003 2024-11-22 17:43 /data/bioproject.xml
[hadoop@bdulamsurankhor-1 ~]$ ]
```

5) time hadoop jar hadoop-3.4.0/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.0.jar wordcount /data/bioproject.xml /data/wordcount1

```
● ● ● 2024 Fall — hadoop@bdulamsurankhor-1:~ — ssh cc@129.114.26.216 — 127x24
...— ssh cc@129.114.26.216 ...-3.4.0/etc/hadoop -- zsh ....4.0/etc/hadoop -- zsh ... ....4.0/etc/hadoop -- zsh ...
java.io.IOException: Got error, status=ERROR, status message , ack with firstBadLink as 10.56.0.255:9866
    at org.apache.hadoop.hdfs.protocol.datatransfer.DataTransferProtoUtil.checkBlockOpStatus(DataTransferProtoUtil.java:128
)
    at org.apache.hadoop.hdfs.protocol.datatransfer.DataTransferProtoUtil.checkBlockOpStatus(DataTransferProtoUtil.java:104
)
    at org.apache.hadoop.hdfs.DataStreamer.createBlockOutputStream(DataStreamer.java:1921)
    at org.apache.hadoop.hdfs.DataStreamer.nextBlockOutputStream(DataStreamer.java:1822)
    at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:751)
2024-11-22 17:46:48,512 WARN hdfs.DataStreamer: Abandoning BP-925688237-10.56.3.239-1732315252639:blk_1073741844_1020
2024-11-22 17:46:48,516 WARN hdfs.DataStreamer: Excluding datanode DatanodeInfoWithStorage[10.56.0.255:9866,DS-fd764505-dd83-4a
f8-9dc4-4d286ad0724a,DISK]
2024-11-22 17:46:48,546 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1732315290494_0001
2024-11-22 17:46:48,546 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-11-22 17:46:48,800 INFO conf.Configuration: resource-types.xml not found
2024-11-22 17:46:48,800 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-11-22 17:46:49,278 INFO impl.YarnClientImpl: Submitted application application_1732315290494_0001
2024-11-22 17:46:49,351 INFO mapreduce.Job: The url to track the job: http://bdulamsurankhor-1.novalocal:8088/proxy/application
_1732315290494_0001/
2024-11-22 17:46:49,352 INFO mapreduce.Job: Running job: job_1732315290494_0001
2024-11-22 17:46:59,528 INFO mapreduce.Job: Job job_1732315290494_0001 running in uber mode : false
2024-11-22 17:46:59,530 INFO mapreduce.Job: map 0% reduce 0%
2024-11-22 17:47:19,702 INFO mapreduce.Job: map 46% reduce 0%
2024-11-22 17:47:25,778 INFO mapreduce.Job: map 77% reduce 0%
```

hadoop fs -du /data/wordcount1/

```
● ● ● 2024 Fall — hadoop@bdulamsurankhor-1:~ — ssh cc@129.114.26.216 — 127x35
...— ssh cc@129.114.26.216 ...-3.4.0/etc/hadoop -- zsh ....4.0/etc/hadoop -- zsh ... ....4.0/etc/hadoop -- zsh ... +
[[hadoop@bdulamsurankhor-1 ~]$ hadoop fs -du /data/wordcount1/
0          /data/wordcount1/_SUCCESS
20056175  40112350  /data/wordcount1/part-r-00000
[hadoop@bdulamsurankhor-1 ~]$ ]
```

```
hadoop fs -cat /data/wordcount1/part-r-00000 — grep arctic
```

```
● ● ● 2024 Fall — hadoop@bdulamsurankhor-1:~ — ssh cc@129.114.26.216 — 127x35
...— ssh cc@129.114.26.216 ...-3.4.0/etc/hadoop -- zsh ....4.0/etc/hadoop -- zsh ... ....4.0/etc/hadoop -- zsh ... +
antarctica</Title>      1
antarcticum      32
antarcticum</Name>      3
antarcticum</OrganismName>      3
antarcticus      31
antarcticus&lt;/i&gt;      4
antarcticus&lt;/i&gt;&lt;/b&gt;..      1
antarcticus).      1
antarcticus,      1
antarcticus</Name>      5
antarcticus</OrganismName>      5
arctic      21
arctica      27
arctica&lt;/I&gt;      2
arctica&lt;/i&gt;      3
arctica&lt;/i&gt;,      1
arctica.</Description>  2
arctica</Name>      5
arctica</OrganismName>  5
arcticus      31
arcticus&lt;/i&gt;      2
arcticus</Name>      4
arcticus</OrganismName>  4
holarctica      77
humans.Antarctic      1
palearctica      66
palearctica</Name>      1
sub-Antarctic      4
sub-arctic      4
subantarctic      1
subantarcticus      7
subantarcticus</Name>  1
subantarcticus</OrganismName>  1
subarctic      21
[hadoop@bdulamsurankhor-1 ~]$ ]
```

6) SELECT COUNT(*) FROM VehicleData;

```

2024 Fall — hadoop@bdulamsurankhor-1:~/apache-hive-4.0.1-bin — ssh cc@129.114.26.216 — 127x35
...— ssh cc@129.114.26.216      ...-3.4.0/etc/hadoop --zsh      ....4.0/etc/hadoop --zsh ...      ....4.0/etc/hadoop --zsh ...
) ~[hadoop-hdfs-client-3.4.0.jar:?]
    at org.apache.hadoop.hdfs.protocol.datatransfer.DataTransferProtoUtil.checkBlockOpStatus(DataTransferProtoUtil.java:104
) ~[hadoop-hdfs-client-3.4.0.jar:?]
    at org.apache.hadoop.hdfs.DataStreamer.createBlockOutputStream(DataStreamer.java:1921) ~[hadoop-hdfs-client-3.4.0.jar:?
]
    at org.apache.hadoop.hdfs.DataStreamer.nextBlockOutputStream(DataStreamer.java:1822) ~[hadoop-hdfs-client-3.4.0.jar:?
]
    at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:751) ~[hadoop-hdfs-client-3.4.0.jar:?]
24/11/22 18:00:03 [Thread-491]: WARN hdfs.DataStreamer: Abandoning BP-925688237-10.56.3.239-1732315252639:blk_1073741896_1072
24/11/22 18:00:03 [Thread-491]: WARN hdfs.DataStreamer: Excluding datanode DatanodeInfoWithStorage[10.56.0.255:9866,DS-fd764505-
dd83-4af8-9dc4-4d286ad0724a,DISK]
Starting Job = job_1732315290494_0002, Tracking URL = http://bdulamsurankhor-1.novalocal:8088/proxy/application_1732315290494_0
002/
Kill Command = /opt/hadoop/hadoop-3.4.0/bin/mapred job -kill job_1732315290494_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
24/11/22 18:00:14 [HiveServer2-Background-Pool: Thread-73]: WARN mapreduce.Counters: Group org.apache.hadoop.mapred.Task$Counte
r is deprecated. Use org.apache.hadoop.mapreduce.TaskCounter instead
2024-11-22 18:00:14,348 Stage-1 map = 0%, reduce = 0%
2024-11-22 18:00:21,762 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.62 sec
2024-11-22 18:01:22,024 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.62 sec
2024-11-22 18:02:23,006 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.62 sec
2024-11-22 18:03:23,889 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.62 sec
2024-11-22 18:03:34,207 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.25 sec
MapReduce Total cumulative CPU time: 5 seconds 250 msec
Ended Job = job_1732315290494_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.25 sec HDFS Read: 11782361 HDFS Write: 105 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 250 msec
+-----+
| _c0 |
+-----+
| 34175 |
+-----+
1 row selected (217.497 seconds)
0: jdbc:hive2://>
0: jdbc:hive2://>

```

SELECT MIN(barrels08), AVG(barrels08), MAX(barrels08) FROM VehicleData;

```

2024 Fall — hadoop@bdulamsurankhor-1:~/apache-hive-4.0.1-bin — ssh cc@129.114.26.216 — 127x35
...— ssh cc@129.114.26.216      ...-3.4.0/etc/hadoop --zsh      ....4.0/etc/hadoop --zsh ...      ....4.0/etc/hadoop --zsh ...
) ~[hadoop-hdfs-client-3.4.0.jar:?]
    at org.apache.hadoop.hdfs.protocol.datatransfer.DataTransferProtoUtil.checkBlockOpStatus(DataTransferProtoUtil.java:128
) ~[hadoop-hdfs-client-3.4.0.jar:?]
    at org.apache.hadoop.hdfs.protocol.datatransfer.DataTransferProtoUtil.checkBlockOpStatus(DataTransferProtoUtil.java:104
) ~[hadoop-hdfs-client-3.4.0.jar:?]
    at org.apache.hadoop.hdfs.DataStreamer.createBlockOutputStream(DataStreamer.java:1921) ~[hadoop-hdfs-client-3.4.0.jar:?
]
    at org.apache.hadoop.hdfs.DataStreamer.nextBlockOutputStream(DataStreamer.java:1822) ~[hadoop-hdfs-client-3.4.0.jar:?
]
    at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:751) ~[hadoop-hdfs-client-3.4.0.jar:?]
24/11/22 18:04:17 [Thread-70]: WARN hdfs.DataStreamer: Abandoning BP-925688237-10.56.3.239-1732315252639:blk_1073741944_1120
24/11/22 18:04:17 [Thread-70]: WARN hdfs.DataStreamer: Excluding datanode DatanodeInfoWithStorage[10.56.0.255:9866,DS-fd764505-
dd83-4af8-9dc4-4d286ad0724a,DISK]
Starting Job = job_1732315290494_0003, Tracking URL = http://bdulamsurankhor-1.novalocal:8088/proxy/application_1732315290494_0
003/
Kill Command = /opt/hadoop/hadoop-3.4.0/bin/mapred job -kill job_1732315290494_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
24/11/22 18:04:26 [HiveServer2-Background-Pool: Thread-100]: WARN mapreduce.Counters: Group org.apache.hadoop.mapred.Task$Counte
r is deprecated. Use org.apache.hadoop.mapreduce.TaskCounter instead
2024-11-22 18:04:26,241 Stage-1 map = 0%, reduce = 0%
2024-11-22 18:04:33,588 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.58 sec
2024-11-22 18:05:34,608 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.58 sec
2024-11-22 18:06:35,452 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.58 sec
2024-11-22 18:07:36,227 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.58 sec
2024-11-22 18:07:43,471 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.99 sec
MapReduce Total cumulative CPU time: 6 seconds 990 msec
Ended Job = job_1732315290494_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.99 sec HDFS Read: 11787022 HDFS Write: 136 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 990 msec
+-----+-----+-----+
| _c0   |     _c1    |     _c2    |
+-----+-----+-----+
| 0.059892 | 17.820177449476272 | 47.06831 |
+-----+-----+-----+
1 row selected (210.147 seconds)
0: jdbc:hive2://>

```

SELECT (barrels08/city08) FROM VehicleData;

```

● ● ● 2024 Fall — hadoop@bdulamsurankhor-1:~/apache-hive-4.0.1-bin — ssh cc@129.114.26.216 — 127x35
...— ssh cc@129.114.26.216 | ...-3.4.0/etc/hadoop --zsh | ....4.0/etc/hadoop --zsh ... | ....4.0/etc/hadoop --zsh ... | +
+-----+
| 0.8716353310479058 |
| 0.6822403953188941 |
| 0.550305407988214 |
| 0.6822403953188941 |
| 0.599298911271528 |
| 0.8716353310479058 |
| 0.9694411333869485 |
| 0.7875944438733553 |
| 0.7540551737735146 |
| 0.6539881115867978 |
| 1.5695714950561523 |
| 0.8716353310479058 |
| 0.8716353310479058 |
| 0.5732347654259723 |
| 0.4898370901743571 |
| 0.4225769409766564 |
| 2.99505615234375 |
| 1.0767260158763212 |
| 0.8716353310479058 |
| 0.8716353310479058 |
| 1.1440213918685913 |
| 2.4970455169677734 |
| 0.9155832926432291 |
| 0.550305407988214 |
| 0.6539881115867978 |
| 0.4898370901743571 |
| 0.6278285526093983 |
| 0.7875944438733553 |
| 0.7163524150848388 |
| 0.8716353310479058 |
| 0.8716353310479058 |
| 1.1440213918685913 |
+-----+
34,175 rows selected (1.872 seconds)
0: jdbc:hive2://> ■

```

INSERT OVERWRITE DIRECTORY 'ThreeColExtract' SELECT barrels08, city08, charge120 FROM VehicleData;

```

● ● ● 2024 Fall — hadoop@bdulamsurankhor-1:~/apache-hive-4.0.1-bin — ssh cc@129.114.26.216 — 127x35
...— ssh cc@129.114.26.216 | ...-3.4.0/etc/hadoop --zsh | ....4.0/etc/hadoop --zsh ... | ....4.0/etc/hadoop --zsh ... | +
+-----+
dd83-4af8-9dc4-4d286ad0724a,DISK
24/11/22 18:21:27 [Thread-35]: WARN hdfs.DataStreamer: Exception in createBlockOutputStream blk_1073742008_1184
java.io.IOException: Got error, status=ERROR, status message , ack with firstBadLink as 10.56.0.230:9866
        at org.apache.hadoop.hdfs.protocol.datatransfer.DataTransferProtoUtil.checkBlockOpStatus(DataTransferProtoUtil.java:128
) ~[hadoop-hdfs-client-3.4.0.jar:?]
        at org.apache.hadoop.hdfs.protocol.datatransfer.DataTransferProtoUtil.checkBlockOpStatus(DataTransferProtoUtil.java:104
) ~[hadoop-hdfs-client-3.4.0.jar:?]
        at org.apache.hadoop.hdfs.DataStreamer.createBlockOutputStream(DataStreamer.java:1921) ~[hadoop-hdfs-client-3.4.0.jar:?
]
        at org.apache.hadoop.hdfs.DataStreamer.nextBlockOutputStream(DataStreamer.java:1822) ~[hadoop-hdfs-client-3.4.0.jar:?]
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:751) ~[hadoop-hdfs-client-3.4.0.jar:?]
24/11/22 18:21:27 [Thread-35]: WARN hdfs.DataStreamer: Abandoning BP-925688237-10.56.3.239-1732315252639:blk_1073742008_1184
24/11/22 18:21:27 [Thread-35]: WARN hdfs.DataStreamer: Excluding datanode DatanodeInfoWithStorage[10.56.0.230:9866,DS-24440dc0-1bd2-41af-8cdc-cb60eac72317,DISK]
Starting Job = job_1732315290494_0005, Tracking URL = http://bdulamsurankhor-1.novalocal:8088/proxy/application_1732315290494_0
005/
Kill Command = /opt/hadoop/hadoop-3.4.0/bin/mapred job -kill job_1732315290494_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
24/11/22 18:21:37 [HiveServer2-Background-Pool: Thread-55]: WARN mapreduce.Counters: Group org.apache.hadoop.mapred.Task$Counte
r is deprecated. Use org.apache.hadoop.mapreduce.TaskCounter instead
2024-11-22 18:21:37,222 Stage-1 map = 0%, reduce = 0%
2024-11-22 18:21:44,675 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.26 sec
MapReduce Total cumulative CPU time: 3 seconds 260 msec
Ended Job = job_1732315290494_0005
Stage-3 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
Stage-4 is filtered out by condition resolver.
Moving data to directory hdfs://bdulamsurankhor-1.novalocal:9000/user/hadoop/ThreeColExtract/.hive-staging_hive_2024-11-22_18-2
1-18_363_2362118187097090045-1-ext-10000
Moving data to directory ThreeColExtract
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 3.26 sec HDFS Read: 11773125 HDFS Write: 627873 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 260 msec
No rows affected (27.602 seconds)
0: jdbc:hive2://> ■

```

```
hadoop fs -ls ThreeColExtract
```

The terminal window shows the command `hadoop fs -ls ThreeColExtract` being run. The output of the job is displayed, detailing the map and reduce steps, and the final success message. The job ID is `job_1732315290494_0005`. The log includes several WARN messages about abandoning blocks and excluding datanodes.

```
2024 Fall — hadoop@bdulamsurankhor-1:~/apache-hive-4.0.1-bin — ssh cc@129.114.26.216 — 127x35
...— ssh cc@129.114.26.216 ...-3.4.0/etc/hadoop --zsh ...4.0/etc/hadoop --zsh .......4.0/etc/hadoop --zsh ...
at org.apache.hadoop.hdfs.protocol.datatransfer.DataTransferProtoUtil.checkBlockOpStatus(DataTransferProtoUtil.java:104
) ~[hadoop-hdfs-client-3.4.0.jar:?:]
at org.apache.hadoop.hdfs.DataStreamer.createBlockOutputStream(DataStreamer.java:1921) ~[hadoop-hdfs-client-3.4.0.jar:?
]
at org.apache.hadoop.hdfs.DataStreamer.nextBlockOutputStream(DataStreamer.java:1822) ~[hadoop-hdfs-client-3.4.0.jar:?:]
at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:751) ~[hadoop-hdfs-client-3.4.0.jar:?:]
24/11/22 18:21:27 [Thread-35]: WARN hdfs.DataStreamer: Abandoning BP-925688237-10.56.3.239-1732315252639:blk_1073742008_1184
24/11/22 18:21:27 [Thread-35]: WARN hdfs.DataStreamer: Excluding datanode DatanodeInfoWithStorage[10.56.0.230:9866,DS-24440dc0-
1bd2-41af-8cdc-cb60eac72317,DISK]
Starting Job = job_1732315290494_0005, Tracking URL = http://bdulamsurankhor-1.novalocal:8088/proxy/application_1732315290494_0
005/
Kill Command = /opt/hadoop/hadoop-3.4.0/bin/mapred job -kill job_1732315290494_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
24/11/22 18:21:37 [HiveServer2-Background-Pool: Thread-55]: WARN mapreduce.Counters: Group org.apache.hadoop.mapred.Task$Counte
r is deprecated. Use org.apache.hadoop.mapreduce.TaskCounter instead
2024-11-22 18:21:37,222 Stage-1 map = 0%, reduce = 0%
2024-11-22 18:21:44,675 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.26 sec
MapReduce Total cumulative CPU time: 3 seconds 260 msec
Ended Job = job_1732315290494_0005
Stage-3 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
Stage-4 is filtered out by condition resolver.
Moving data to directory hdfs://bdulamsurankhor-1.novalocal:9000/user/hadoop/ThreeColExtract/.hive-staging_hive_2024-11-22_18-2
1-18_363_2362118187097090045-1-ext-10000
Moving data to directory ThreeColExtract
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 3.26 sec HDFS Read: 11773125 HDFS Write: 627873 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 260 msec
No rows affected (27.602 seconds)
0: jdbc:hive2://> !quit
Closing: 0: jdbc:hive2://
[[hadoop@bdulamsurankhor-1 apache-hive-4.0.1-bin]$ hadoop fs -ls ThreeColExtract
Found 1 items
-rw-r--r-- 2 hadoop supergroup 627873 2024-11-22 18:21 ThreeColExtract/000000_0
[hadoop@bdulamsurankhor-1 apache-hive-4.0.1-bin]$ ]]
```

Download data:

```
wget http://cdmgcsarprd01.dpu.depaul.edu/CSC555/SSBM1/lineorder.tbl
```

Load data:

```
LOAD DATA LOCAL INPATH '/opt/hadoop/lineorder.tbl' OVERWRITE INTO TABLE lineorder;
```

```

2024 Fall — hadoop@bdulamsurankhor-1:~/apache-hive-4.0.1-bin — ssh cc@129.114.26.216 — 127x50
...— ssh cc@129.114.26.216 ...14.26.216 -i cs451-bk.key ...ssh cc@129.114.26.216 ...ssh cc@129.114.26.216 + 
24/11/22 22:37:30 [9c84119d-0c05-43d0-8032-9a8388fe246c main]: WARN calcite.RelOptHiveTable: No Stats for default@lineorder, Columns: lo_quantity, lo_shipmode, lo_tax
No Stats for default@lineorder, Columns: lo_quantity, lo_shipmode, lo_tax
24/11/22 22:37:31 [HiveServer2-Background-Pool: Thread-55]: WARN ql.Driver: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez) or using Hive 1.X releases.
Query ID = hadoop_20241122223724_f3023152-e1b9-4928-9dbd-3dbc4f64fe9b
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 3
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=<nnumber>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
WARN : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez) or using Hive 1.X releases.
24/11/22 22:37:32 [HiveServer2-Background-Pool: Thread-55]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Starting Job = job_1732315290494_0014, Tracking URL = http://bdulamsurankhor-1.novalocal:8088/proxy/application_1732315290494_0014/
Kill Command = /opt/hadoop/hadoop-3.4.0/bin/mapred job -kill job_1732315290494_0014
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 3
24/11/22 22:37:45 [HiveServer2-Background-Pool: Thread-55]: WARN mapreduce.Counters: Group org.apache.hadoop.mapred.Task$Counter is deprecated. Use org.apache.hadoop.mapreduce.TaskCounter instead
2024-11-22 22:37:45,100 Stage-1 map = 0%, reduce = 0%
2024-11-22 22:37:57,782 Stage-1 map = 33%, reduce = 0%, Cumulative CPU 6.5 sec
2024-11-22 22:37:59,898 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 15.36 sec
2024-11-22 22:38:02,010 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 25.07 sec
2024-11-22 22:38:06,231 Stage-1 map = 100%, reduce = 33%, Cumulative CPU 28.92 sec
2024-11-22 22:38:07,279 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 34.0 sec
2024-11-22 22:38:08,327 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 38.84 sec
MapReduce Total cumulative CPU time: 38 seconds 840 msec
Ended Job = job_1732315290494_0014
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3 Reduce: 3 Cumulative CPU: 38.84 sec HDFS Read: 594399737 HDFS Write: 510 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 38 seconds 840 msec
+-----+
| lo_shipmode | _c1 |
+-----+
| TRUCK      | 4.005838235753945 |
| AIR         | 4.0023230935198615 |
| MAIL        | 4.001676047753917 |
| RAIL        | 4.000103382867211 |
| REG AIR     | 3.998940777810193 |
| SHIP        | 3.9935317643784702 |
| FOB         | 4.001717011413867 |
+-----+
7 rows selected (44.901 seconds)
0: jdbc:hive2://> 

```

7) Use the same vehicles file. Copy the vehicles.csv file to the HDFS if it is not already there.

VehicleData = LOAD '/data/vehicles.csv' USING PigStorage(',') AS (barrels08:FLOAT, barrelsA08:FLOAT, charge120:FLOAT, charge240:FLOAT, city08:FLOAT);

You can see the table description by

DESCRIBE VehicleData;

Verify that your data has loaded by running:

VehicleG = GROUP VehicleData ALL;

Count = FOREACH VehicleG GENERATE COUNT(VehicleData);

DUMP Count;

```

● ● ● 2024 Fall — hadoop@bdulamsurankhor-1:~ — ssh cc@129.114.26.216 — 127x50
...— ssh cc@129.114.26.216 ...14.26.216 -i cs451-bk.key ...ssh cc@129.114.26.216 ...ssh cc@129.114.26.216 +
```

Success!

Job Stats (time in seconds):

JobId	Maps	Reduces	MaxMapTime	MinMapTime	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime
ceTime	MedianReduceTime		Alias	Feature	Outputs				
job_1732315290494_0015	1	1	4	4	4	4	4	4	Count,VehicleData,Vehicle
leG	GROUP_BY,COMBINER		hdfs://bdulamsurankhor-1.novalocal:9000/tmp/temp734754683/tmp279684727,						

Input(s):
Successfully read 34175 records (11766965 bytes) from: "/data/vehicles.csv"

Output(s):
Successfully stored 1 records (9 bytes) in: "hdfs://bdulamsurankhor-1.novalocal:9000/tmp/temp734754683/tmp279684727"

Counters:

- Total records written : 1
- Total bytes written : 9
- Spillable Memory Manager spill count : 0
- Total bags proactively spilled: 0
- Total records proactively spilled: 0

Job DAG:
job_1732315290494_0015

2024-11-22 22:43:28,657 [main] INFO org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at bdulamsurankhor-1.novalocal/10.56.3.239:8032
2024-11-22 22:43:28,664 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2024-11-22 22:43:28,700 [main] INFO org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at bdulamsurankhor-1.novalocal/10.56.3.239:8032
2024-11-22 22:43:28,703 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2024-11-22 22:43:28,737 [main] INFO org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at bdulamsurankhor-1.novalocal/10.56.3.239:8032
2024-11-22 22:43:28,740 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2024-11-22 22:43:28,771 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 5 time(s).
2024-11-22 22:43:28,771 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-11-22 22:43:28,775 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
2024-11-22 22:43:28,776 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-11-22 22:43:28,799 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-11-22 22:43:28,799 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(34174)
grunt> █