ADVANCE BIG DATA LAB ANALYTICS
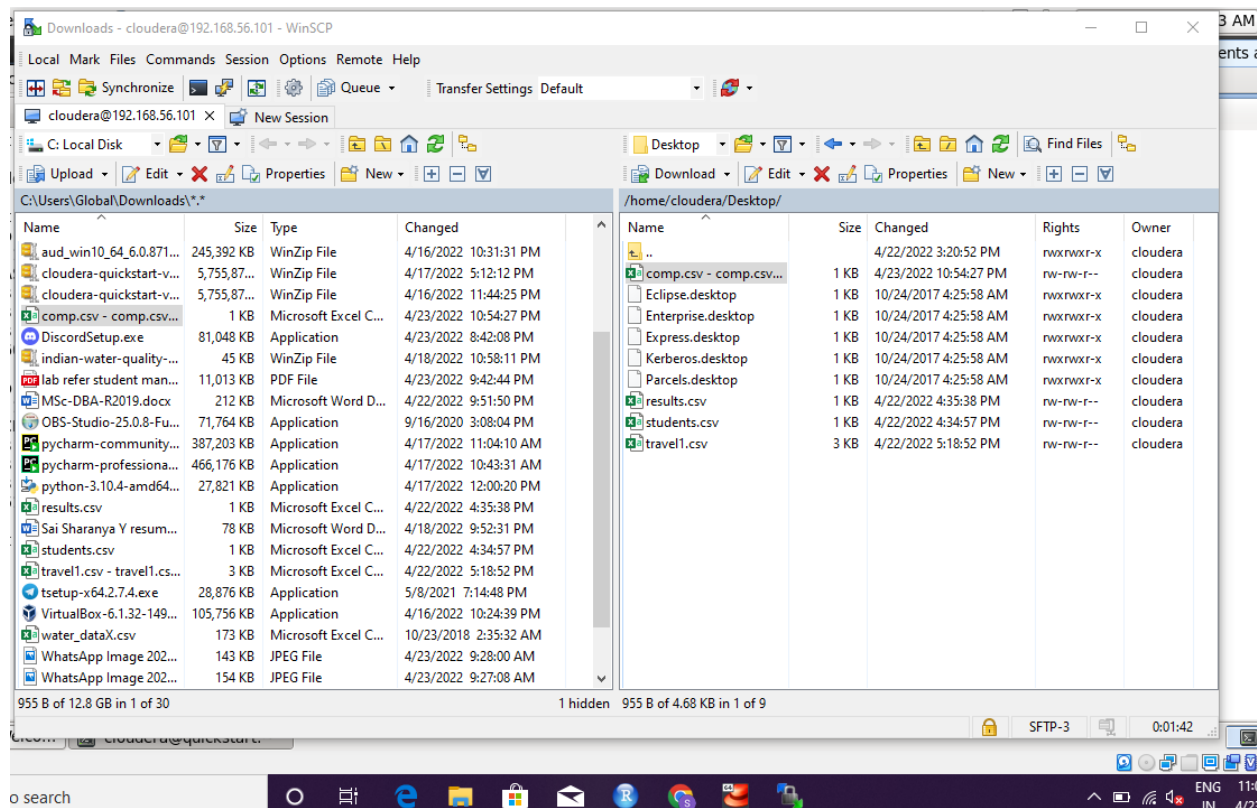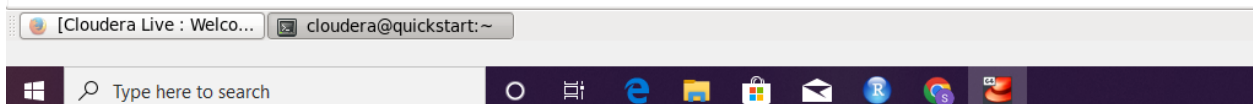## EXP - 4

19BTRCR043

SAI SHARANYA Y

To build a script which produces a report listing each Company & State and number of complaints raised by them.

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /Companydata
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 9 items
drwxr-xr-x   - cloudera supergroup          0 2022-04-23 10:25 /Companydata
drwxr-xr-x   - cloudera supergroup          0 2022-04-22 09:18 /TravelLab
drwxrwxrwx   - hdfs     supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - hbase    supergroup          0 2022-04-22 02:59 /hbase
drwxr-xr-x   - cloudera supergroup          0 2022-04-23 09:46 /salarydata
drwxr-xr-x   - solr     solr                0 2017-10-23 09:18 /solr
drwxrwxrwt   - hdfs     supergroup          0 2022-04-23 09:07 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Desktop/comp.csv  /Companydata
[cloudera@quickstart ~]$ hdfs dfs -ls /Companydata
Found 1 items
-rw-r--r--   1 cloudera supergroup        955 2022-04-23 10:26 /Companydata/comp.csv
[cloudera@quickstart ~]$
```

```
grunt>
grunt>
grunt>
grunt> compdata = load '/Companydata/comp.csv' using PigStorage (',') as (company_name : chararray, state : chararray, complaint : chararray);
2022-04-23 10:28:52,077 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-23 10:28:52,077 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
grunt> dump compdata;
2022-04-23 10:29:05,827 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2022-04-23 10:29:05,828 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachCol
umnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownFo
rEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2022-04-23 10:29:05,835 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-04-23 10:29:05,837 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-04-23 10:29:05,837 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-04-23 10:29:05,869 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-04-23 10:29:05,871 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-04-23 10:29:05,873 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is no
t set, set to default 0.3
2022-04-23 10:29:06,333 [DataStreamer for file /tmp/temp1844820160/tmp-554985555/hive-hbase-handler-1.1.0-cdh5.13.0.jar] WARN  org.apache.hadoop.hdfs.DFSClient - Caugh
t exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
```

```
HadoopVersion  PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.13.0 0.12.0-cdh5.13.0    root    2022-04-23 10:29:05    2022-04-23 10:29:36    UNKNOWN

Success!

Job Stats (time in seconds):
JobId    Maps   Reduces MaxMapTime    MinMapTIme    AvgMapTime    MedianMapTime    MaxReduceTime    MinReduceTime    AvgReduceTime    MedianReducetime    Alias F
eature  Outputs
job_1650620943421_0012 1    0    8    8    8    8    n/a    n/a    n/a    n/a    compdata    MAP_ONLY    hdfs://quickstart.cloudera:8020
/tmp/temp1844820160/tmp806086029,

Input(s):
Successfully read 19 records (1326 bytes) from: "/Companydata/comp.csv"

Output(s):
Successfully stored 19 records (820 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp1844820160/tmp806086029"

Counters:
Total records written : 19
Total bytes written : 820
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1650620943421_0012
```

```
2022-04-23 10:29:36,199 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-04-23 10:29:36,199 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-23 10:29:36,199 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-23 10:29:36,200 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-04-23 10:29:36,207 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-04-23 10:29:36,207 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Experian Information Solutions Inc.,NJ,1509954)
(AES/PHEAA,MA,3475943)
("EQUIFAX, INC.",NY)
(WELLS FARGO & COMPANY,CA,1527601)
(DISCOVER BANK,NV,1816726)
("BANK OF AMERICA, NATIONAL ASSOCIATION",PA)
("BANK OF AMERICA, NATIONAL ASSOCIATION",MD)
(SYNCHRONY FINANCIAL,CA,3054861)
("TRANSUNION INTERMEDIATE HOLDINGS, INC.",VA)
("Cascade Capital, LLC",FL)
(CAPITAL ONE FINANCIAL CORPORATION,MD,3108212)
(BANCO POPULAR DE PUERTO RICO,PR,2107177)
(TD BANK US HOLDING COMPANY,NJ,2874325)
("CITIBANK, N.A.",FL)
("CARRINGTON MORTGAGE SERVICES, LLC",ND)
(Experian Information Solutions Inc.,GA,2266132)
("WINN LAW GROUP, APC",CA)
("TRANSUNION INTERMEDIATE HOLDINGS, INC.",NC)
(Experian Information Solutions Inc.,IL,3119951)
grunt>
```

```
(Experian Information Solutions Inc.,IL,3119951)
grunt> x = group compdata by (company_name, state);
grunt> dump x;
2022-04-23 10:33:19,333 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2022-04-23 10:33:19,334 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEa
umnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushD
rEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2022-04-23 10:33:19,340 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-04-23 10:33:19,341 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-04-23 10:33:19,341 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-04-23 10:33:19,368 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-04-23 10:33:19,370 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-04-23 10:33:19,373 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent
t set, set to default 0.3
2022-04-23 10:33:19,374 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of requ
reducers.
2022-04-23 10:33:19,374 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.bac
hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2022-04-23 10:33:19,376 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReduce
9 totalInputFileSize=955
2022-04-23 10:33:19,376 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2022-04-23 10:33:19,692 [DataStreamer for file /tmp/temp1844820160/tmp713470568/hive-hcatalog-pig-adapter-1.1.0-cdh5.13.0.jar] WARN  org.apache.hadoop.hdfs.DFSCli
 Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
```

```
HadoopVersion    PigVersion        UserId  StartedAt        FinishedAt        Features
2.6.0-cdh5.13.0  0.12.0-cdh5.13.0    root  2022-04-23 10:33:19    2022-04-23 10:34:09    GROUP_BY

Success!

Job Stats (time in seconds):
JobId    Maps    Reduces MaxMapTime    MinMapTIme    AvgMapTime    MedianMapTime    MaxReduceTime    MinReduceTime    AvgReduceTime    MedianReducetime        Alias F
eature  Outputs
job_1650620943421_0013  1      1      15      15      15      15      10      10      10      10      compdata,x    GROUP_BY        hdfs://quickstart.cloudera:8020
/tmp/temp1844820160/tmp-325428902,

Input(s):
Successfully read 19 records (1326 bytes) from: "/Companydata/comp.csv"

Output(s):
Successfully stored 17 records (1445 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp1844820160/tmp-325428902"

Counters:
Total records written : 17
Total bytes written : 1445
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1650620943421_0013
```

[Cloudera Live : Welco...] [root@quickstart:~]

Type here to search   ENG 11:24 PM IN 4/23/2022

```
Job DAG:
job_1650620943421_0013


2022-04-23 10:34:09,286 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-04-23 10:34:09,287 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-23 10:34:09,287 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-23 10:34:09,287 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-04-23 10:34:09,294 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-04-23 10:34:09,294 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(("EQUIFAX, INC."),{("EQUIFAX, INC.",NY)})
(("CITIBANK, N.A."),{("CITIBANK, N.A.",FL)})
((AES/PHEAA,MA),{(AES/PHEAA,MA,3475943)})
((DISCOVER BANK,NV),{(DISCOVER BANK,NV,1816726)})
(("WINN LAW GROUP, APC"),{("WINN LAW GROUP, APC",CA)})
(("BANK OF AMERICA, NATIONAL ASSOCIATION"),{("BANK OF AMERICA, NATIONAL ASSOCIATION",MD),("BANK OF AMERICA, NATIONAL ASSOCIATION",PA)})
(("Cascade Capital, LLC"),{("Cascade Capital, LLC",FL)})
((SYNCHRONY FINANCIAL,CA),{(SYNCHRONY FINANCIAL,CA,3054861)})
((WELLS FARGO & COMPANY,CA),{(WELLS FARGO & COMPANY,CA,1527601)})
((TD BANK US HOLDING COMPANY,NJ),{(TD BANK US HOLDING COMPANY,NJ,2874325)})
((BANCO POPULAR DE PUERTO RICO,PR),{(BANCO POPULAR DE PUERTO RICO,PR,2107177)})
(("CARRINGTON MORTGAGE SERVICES, LLC"),{("CARRINGTON MORTGAGE SERVICES, LLC",ND)})
(("TRANSUNION INTERMEDIATE HOLDINGS, INC."),{("TRANSUNION INTERMEDIATE HOLDINGS, INC.",NC),("TRANSUNION INTERMEDIATE HOLDINGS, INC.",VA)})
((CAPITAL ONE FINANCIAL CORPORATION,MD),{(CAPITAL ONE FINANCIAL CORPORATION,MD,3108212)})
((Experian Information Solutions Inc.,GA),{(Experian Information Solutions Inc.,GA,2266132)})
((Experian Information Solutions Inc.,IL),{(Experian Information Solutions Inc.,IL,3119951)})
((Experian Information Solutions Inc.,NJ),{(Experian Information Solutions Inc.,NJ,1509954)})
grunt>
```

[Cloudera Live : Welco...] [root@quickstart:~]

ENG 11:25 PM

```
((Experian Information Solutions Inc.,NJ),{(Experian Information Solutions Inc.,NJ,1509954)})
grunt> y = foreach x generate group, COUNT(compdata.complaint);
grunt> dump y;
2022-04-23 10:42:39,937 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2022-04-23 10:42:39,938 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachCol
umnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownFo
rEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2022-04-23 10:42:39,956 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-04-23 10:42:39,957 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosing to move algebraic foreach to combiner
2022-04-23 10:42:39,961 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-04-23 10:42:39,961 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-04-23 10:42:40,004 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-04-23 10:42:40,006 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-04-23 10:42:40,012 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is no
t set, set to default 0.3
2022-04-23 10:42:40,012 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required
reducers.
2022-04-23 10:42:40,012 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.
hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2022-04-23 10:42:40,014 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=99
9 totalInputFileSize=955
2022-04-23 10:42:40,014 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2022-04-23 10:42:40,356 [DataStreamer for file /tmp/temp1844820160/tmp-2023882259/jdo-api-3.0.1.jar] WARN  org.apache.hadoop.hdfs.DFSClient - Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
```

```
HadoopVersion  PigVersion     UserId  StartedAt      FinishedAt      Features
2.6.0-cdh5.13.0 0.12.0-cdh5.13.0    root    2022-04-23 10:42:40    2022-04-23 10:43:20    GROUP_BY

Success!

Job Stats (time in seconds):
JobID   Maps   Reduces MaxMapTime    MinMapTIme    AvgMapTime    MedianMapTime  MaxReduceTime  MinReduceTime   AvgReduceTime  MedianReducetime         Alias F
eature  Outputs
job_1650620943421_0014  1      1      7      7      7      7      7      7      compdata,x,y    GROUP_BY,COMBINER        hdfs://quickstart.cloud
era:8020/tmp/temp1844820160/tmp-1698350341,

Input(s):
Successfully read 19 records (1326 bytes) from: "/Companydata/comp.csv"

Output(s):
Successfully stored 17 records (620 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp1844820160/tmp-1698350341"

Counters:
Total records written : 17
Total bytes written : 620
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1650620943421_0014
```
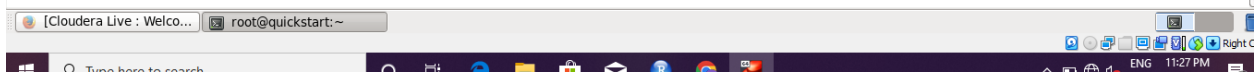
```
job_1650620943421_0014
2022-04-23 10:43:20,684 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-04-23 10:43:20,685 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-23 10:43:20,686 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-23 10:43:20,686 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-04-23 10:43:20,702 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-04-23 10:43:20,702 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(("EQUIFAX, INC."),1)
(("CITIBANK, N.A."),1)
((AES/PHEAA,MA),1)
((DISCOVER BANK,NV),1)
(("WINN LAW GROUP, APC"),1)
(("BANK OF AMERICA, NATIONAL ASSOCIATION"),2)
(("Cascade Capital, LLC"),1)
((SYNCHRONY FINANCIAL,CA),1)
((WELLS FARGO & COMPANY,CA),1)
((TD BANK US HOLDING COMPANY,NJ),1)
((BANCO POPULAR DE PUERTO RICO,PR),1)
(("CARRINGTON MORTGAGE SERVICES, LLC"),1)
(("TRANSUNION INTERMEDIATE HOLDINGS, INC."),2)
((CAPITAL ONE FINANCIAL CORPORATION,MD),1)
((Experian Information Solutions Inc.,GA),1)
((Experian Information Solutions Inc.,IL),1)
((Experian Information Solutions Inc.,NJ),1)
grunt>
```

((Experian Information Solutions Inc.,NJ),1)
grunt> z = ORDER y BY $1 DESC;
grunt> dump z;
2022-04-23 10:44:49,587 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY
2022-04-23 10:44:49,588 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachCol
umnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownFo
rEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2022-04-23 10:44:49,600 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-04-23 10:44:49,607 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosing to move algebraic foreach to combiner
2022-04-23 10:44:49,610 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 3
2022-04-23 10:44:49,610 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 3
2022-04-23 10:44:49,637 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-04-23 10:44:49,639 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-04-23 10:44:49,661 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is no
t set, set to default 0.3
2022-04-23 10:44:49,664 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required
reducers.
2022-04-23 10:44:49,664 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.
hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2022-04-23 10:44:49,666 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=99
9 totalInputFileSize=955
2022-04-23 10:44:49,666 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2022-04-23 10:44:50,013 [DataStreamer for file /tmp/temp1844820160/tmp589044177/hive-hbase-handler-1.1.0-cdh5.13.0.jar] WARN org.apache.hadoop.hdfs.DFSClient - Caught
 exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)

[Cloudera Live : Welco...] [root@quickstart:~]

ENG 11:28 PM

Type here to search

Success!

Job Stats (time in seconds):
JobId   Maps   Reduces MaxMapTime   MinMapTIme   AvgMapTime   MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime   Alias F
eature Outputs
job_1650620943421_0015 1   1   7   7   7   7   8   8   8   8   compdata,x,y   GROUP_BY,COMBINER
job_1650620943421_0016 1   1   6   6   6   6   8   8   8   8   z   SAMPLER
job_1650620943421_0017 1   1   7   7   7   7   8   8   8   8   z   ORDER_BY   hdfs://quickstart.cloudera:8020/tmp/tem
p1844820160/tmp-1261696629,

Input(s):
Successfully read 19 records (1326 bytes) from: "/Companydata/comp.csv"

Output(s):
Successfully stored 17 records (620 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp1844820160/tmp-1261696629"

Counters:
Total records written : 17
Total bytes written : 620
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1650620943421_0015 ->   job_1650620943421_0016,
job_1650620943421_0016 ->   job_1650620943421_0017,
job_1650620943421_0017

[Cloudera Live : Welco...] [root@quickstart:~]

```
        at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
2022-04-23 10:46:57,298 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to
job history server
2022-04-23 10:46:57,583 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to
job history server
2022-04-23 10:46:57,760 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-04-23 10:46:57,761 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-23 10:46:57,761 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-23 10:46:57,761 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-04-23 10:46:57,766 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-04-23 10:46:57,766 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(("TRANSUNION INTERMEDIATE HOLDINGS, INC."),2)
(("BANK OF AMERICA, NATIONAL ASSOCIATION"),2)
((Experian Information Solutions Inc.,NJ),1)
((Experian Information Solutions Inc.,IL),1)
((Experian Information Solutions Inc.,GA),1)
((CAPITAL ONE FINANCIAL CORPORATION,MD),1)
(("CARRINGTON MORTGAGE SERVICES, LLC"),1)
((BANCO POPULAR DE PUERTO RICO,PR),1)
((TD BANK US HOLDING COMPANY,NJ),1)
((WELLS FARGO & COMPANY,CA),1)
((SYNCHRONY FINANCIAL,CA),1)
(("Cascade Capital, LLC"),1)
(("WINN LAW GROUP, APC"),1)
((DISCOVER BANK,NV),1)
((AES/PHEAA,MA),1)
(("CITIBANK, N.A."),1)
(("EQUIFAX, INC."),1)
grunt>
```

```
grunt> store z into '/Companydata/required';
2022-04-23 10:49:25,760 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY
2022-04-23 10:49:25,761 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachCol
umnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownFo
rEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2022-04-23 10:49:25,769 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-04-23 10:49:25,776 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosing to move algebraic foreach to combiner
2022-04-23 10:49:25,779 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 3
2022-04-23 10:49:25,780 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 3
2022-04-23 10:49:25,809 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-04-23 10:49:25,811 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-04-23 10:49:25,818 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is no
t set, set to default 0.3
2022-04-23 10:49:25,818 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required
reducers.
2022-04-23 10:49:25,818 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.
hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2022-04-23 10:49:25,820 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=99
9 totalInputFileSize=955
2022-04-23 10:49:25,820 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2022-04-23 10:49:26,123 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job6946078999961941446.jar
2022-04-23 10:49:30,096 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job6946078999961941446.jar created
2022-04-23 10:49:30,102 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2022-04-23 10:49:30,103 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2022-04-23 10:49:30,103 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2022-04-23 10:49:30,103 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2022-04-23 10:49:30,119 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2022-04-23 10:49:30,119 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
```