# EXPLORING APACHE PIG

# HIVE BASICS

# Assignment Session - 4

**Task 1.1**
Write a program to implement wordcount using Pig.

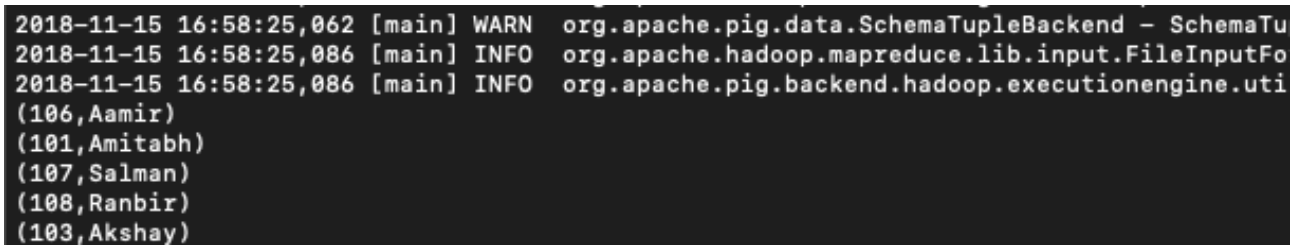
```
~/Documents/ACADGILD_BIG_DATA_ENGINEERING — -bash    ...space-sts-3.9.0.RELEASE/map-reduce-music/target — acadgild@localhost:~/manish — ssh acadgild@192.168.0.105
[acadgild@localhost manish]$ pig -f wordcount.pig
18/11/15 12:24:08 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/11/15 12:24:09 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/11/15 12:24:09 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-11-15 12:24:09,272 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-11-15 12:24:09,272 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/manish/pig_1542264849261.log
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2018-11-15 12:24:10,651 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2018-11-15 12:24:11,358 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-11-15 12:24:11,856 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-11-15 12:24:11,860 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-15 12:24:11,861 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:8020
2018-11-15 12:24:13,356 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-wordcount.pig-a18dfa86-b9af-4a38-ac91-d43eef38050c
2018-11-15 12:24:13,357 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-11-15 12:24:14,962 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-15 12:24:15,265 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 4893
50752
2018-11-15 12:24:15,740 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY
2018-11-15 12:24:15,866 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-15 12:24:15,877 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-11-15 12:24:16,007 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOpt
imizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-11-15 12:24:16,324 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-11-15 12:24:16,400 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2018-11-15 12:24:16,476 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-23
2018-11-15 12:24:16,527 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 3
2018-11-15 12:24:16,527 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 3
2018-11-15 12:24:16,639 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-15 12:24:16,756 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-11-15 12:24:17,734 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2018-11-15 12:24:17,753 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2018-11-15 12:24:17,758 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-11-15 12:24:17,758 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2018-11-15 12:24:17,764 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2018-11-15 12:24:17,769 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.InputSizeReducerEstimator
2018-11-15 12:24:17,799 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=692
2018-11-15 12:24:17,799 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2018-11-15 12:24:17,799 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2018-11-15 12:24:17,805 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2018-11-15 12:24:17,855 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.submit.replication is deprecated. Instead, use mapreduce.client.submit.file.replication
2018-11-15 12:24:19,430 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/acadgild/install/pig/pig-0.16.0/pig-0.16.0-core-h2.jar to Distri
butedCache through /tmp/temp-635210648/tmp-1491463771/pig-0.16.0-core-h2.jar
2018-11-15 12:24:19,508 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/acadgild/install/pig/pig-0.16.0/lib/automaton-1.11-8.jar to Dist
ributedCache through /tmp/temp-635210648/tmp-927440511/automaton-1.11-8.jar
2018-11-15 12:24:19,589 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/acadgild/install/pig/pig-0.16.0/lib/antlr-runtime-3.4.jar to Dis
tributedCache through /tmp/temp-635210648/tmp1524051080/antlr-runtime-3.4.jar
2018-11-15 12:24:19,695 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/acadgild/install/pig/pig-0.16.0/lib/joda-time-2.9.3.jar to Distr
ibutedCache through /tmp/temp-635210648/tmp-1407731388/joda-time-2.9.3.jar
2018-11-15 12:24:19,766 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
```

```
2018-11-15 12:27:09,610 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(data,9)
(with,4)
(to,4)
(and,3)
(Big,3)
(a,2)
(higher,2)
(are,2)
(or,2)
(is,2)
(data-processing,1)
(visualization.,1)
(traditional,1)
(statistical,1)
(information,1)
(application,1)
(originally,1)
(complexity,1)
(challenges,1)
(attributes,1)
(attributed,1)
(associated,1)
(adequately,1)
(velocity.,1)
(updating,,1)
(transfer,,1)
(querying,,1)
(discovery,1)
(concepts:,1)
(capturing,1)
(analysis.,1)
(veracity,1)
(variety,,1)
(storage,,1)
(software,1)
(sharing,,1)
(concepts,1)
(columns),1)
(volume,,1)
(source.,1)
(search.,1)
(privacy,1)
(include,1)
(greater,1)
(complex,1)
(value.,1)
(power,,1)
((rows),1)
((i.e.,,1)
(with.,1)
(three,1)
(refer,1)
(rate.,1)
```

## Task 1.2

(a) Top 5 employees (employee id and employee name) with highest rating. (In case two employees have same rating, employee with name coming first in dictionary should get preference)

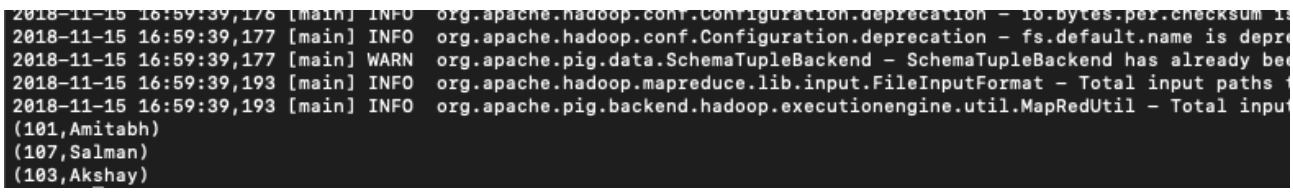
```
employees = load 'employee_details.txt' Using PigStorage(',') as
(empID:int,name:Chararray,salary:int,departmentID:int);
topEmp = order employees by salary desc, name asc;
emp = foreach topEmp generate empID, name;
top5 = limit emp 5;
DUMP top5;
```

(b) Top 3 employees (employee id and employee name) with highest salary, whose employee id is an odd number. (In case two employees have same salary, employee with name coming first in dictionary should get preference)

```
employees = load 'employee_details.txt' Using PigStorage(',') as
(empID:int,name:Chararray,salary:int,departmentID:int);
filterEmp = filter employees by empID%2!=0;
topEmp = order filterEmp by salary desc, name asc;
emp = foreach topEmp generate empID, name;
top3 = limit emp 3;
DUMP top3;
```

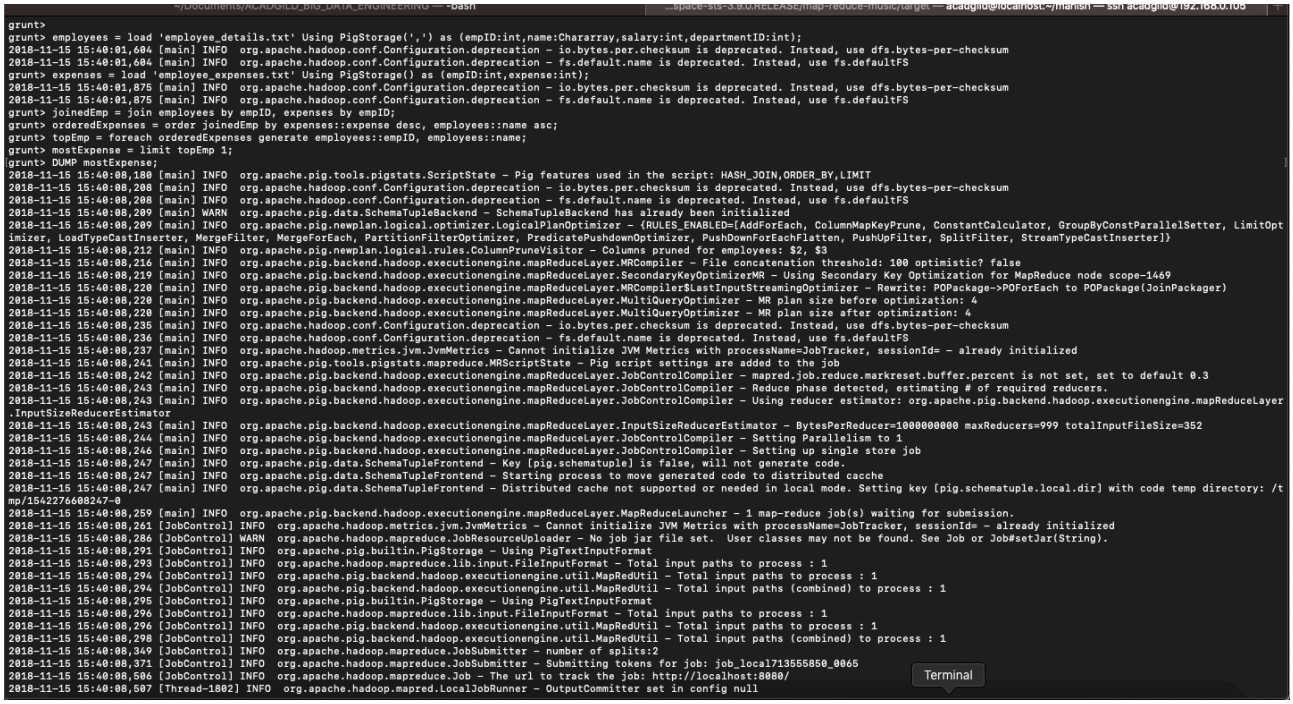
```
grunt> employees = load 'employee_details.txt' Using PigStorage(',') as (empID:int,name:Chararray,salary:int,departmentID:int);
2018-11-15 15:38:43,631 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 15:38:43,631 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> filterEmp = filter employees by empID%2!=0;
grunt> topEmp = order filterEmp by salary desc, name asc;
grunt> emp = foreach topEmp generate empID, name;
grunt> top3 = limit emp 3;
grunt> DUMP top3;
2018-11-15 15:38:46,152 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,FILTER,LIMIT
2018-11-15 15:38:46,209 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 15:38:46,210 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-15 15:38:46,210 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-11-15 15:38:46,210 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOpt
imizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-11-15 15:38:46,211 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for employees: $3
2018-11-15 15:38:46,219 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-11-15 15:38:46,221 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-1379
2018-11-15 15:38:46,221 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4
2018-11-15 15:38:46,221 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 4
2018-11-15 15:38:46,242 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 15:38:46,243 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-15 15:38:46,243 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-11-15 15:38:46,245 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2018-11-15 15:38:46,245 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-11-15 15:38:46,248 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2018-11-15 15:38:46,249 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2018-11-15 15:38:46,249 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2018-11-15 15:38:46,249 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.dir] with code temp directory: /t
mp/1542276526249-0
2018-11-15 15:38:46,260 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2018-11-15 15:38:46,268 [JobControl] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-11-15 15:38:46,286 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2018-11-15 15:38:46,289 [JobControl] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2018-11-15 15:38:46,291 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-11-15 15:38:46,291 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2018-11-15 15:38:46,292 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2018-11-15 15:38:46,326 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2018-11-15 15:38:46,351 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_local2127085431_0061
2018-11-15 15:38:46,459 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8080/
2018-11-15 15:38:46,460 [Thread-1692] INFO  org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter set in config null
2018-11-15 15:38:46,464 [Thread-1692] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-11-15 15:38:46,464 [Thread-1692] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.perc
ent
2018-11-15 15:38:46,465 [Thread-1692] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 15:38:46,465 [Thread-1692] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-15 15:38:46,465 [Thread-1692] INFO  org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter is org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigOutputCommitter
2018-11-15 15:38:46,469 [Thread-1692] INFO  org.apache.hadoop.mapred.LocalJobRunner - Waiting for map tasks
2018-11-15 15:38:46,469 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - Starting task: attempt_local2127085431_0061_m_000000_0
2018-11-15 15:38:46,477 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task -  Using ResourceCalculatorProcessTree : [ ]
2018-11-15 15:38:46,479 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.MapTask - Processing split: Number of splits :1
Total Length = 273
Input split[0]:
   Length = 273
```

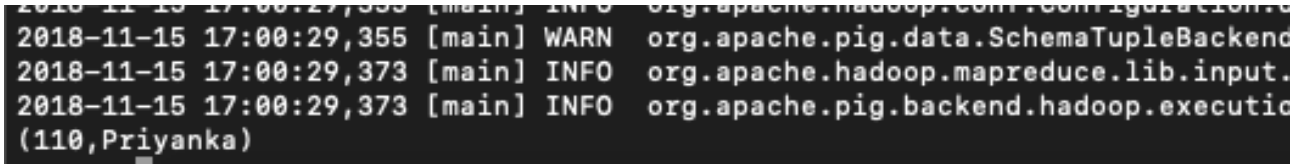


```
2018-11-15 16:59:39,176 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
2018-11-15 16:59:39,177 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depr
2018-11-15 16:59:39,177 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already be
2018-11-15 16:59:39,193 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths
2018-11-15 16:59:39,193 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total inpu
(101,Amitabh)
(107,Salman)
(103,Akshay)
```

(c) Employee (employee id and employee name) with maximum expense (In case two employees have same expense, employee with name coming first in dictionary should get preference)

```
employees = load 'employee_details.txt' Using PigStorage(',') as
(empID:int,name:Chararray,salary:int,departmentID:int);
expenses = load 'employee_expenses.txt' Using PigStorage() as (empID:int,expense:int);
joinedEmp = join employees by empID, expenses by empID;
orderedExpenses = order joinedEmp by expenses::expense desc, employees::name asc;
topEmp = foreach orderedExpenses generate employees::empID, employees::name;
mostExpense = limit topEmp 1;
DUMP mostExpense;
```

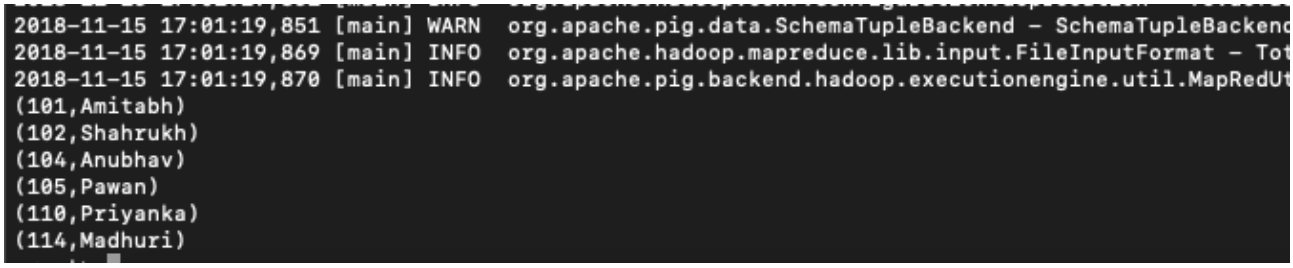


(d) List of employees (employee id and employee name) having entries in employee_expenses file.

```
employees = load 'employee_details.txt' Using PigStorage(',') as
(empID:int,name:Chararray,salary:int,departmentID:int);
expenses = load 'employee_expenses.txt' Using PigStorage() as (empID:int,expense:int);
joinedEmp = join employees by empID, expenses by empID;
emp = foreach orderedExpenses generate employees::empID, employees::name;
distinctEmp = DISTINCT emp;
DUMP distinctEmp;
```

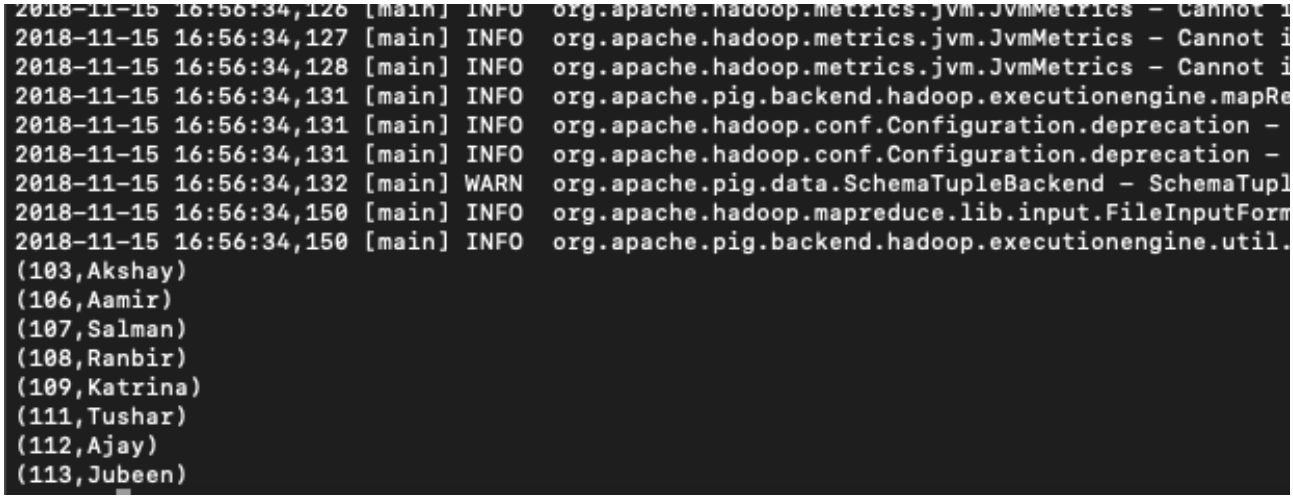
```
grunt>
grunt> employees = load 'employee_details.txt' Using PigStorage(',') as (empID:int,name:Chararray,salary:int,departmentID:int);
2018-11-15 15:42:37,697 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 15:42:37,697 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> expenses = load 'employee_expenses.txt' Using PigStorage() as (empID:int,expense:int);
2018-11-15 15:42:37,879 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 15:42:37,880 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> joinedEmp = join employees by empID, expenses by empID;
grunt> emp = foreach orderedExpenses generate employees::empID, employees::name;
grunt> distinctEmp = DISTINCT emp;
grunt> DUMP distinctEmp;
2018-11-15 15:42:40,786 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,ORDER_BY,DISTINCT
2018-11-15 15:42:40,811 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 15:42:40,812 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-15 15:42:40,812 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-11-15 15:42:40,812 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOpt
imizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-11-15 15:42:40,813 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for employees: $2, $3
2018-11-15 15:42:40,815 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-11-15 15:42:40,822 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-1561
2018-11-15 15:42:40,822 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2018-11-15 15:42:40,822 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4
2018-11-15 15:42:40,822 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 4
2018-11-15 15:42:40,849 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 15:42:40,853 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-15 15:42:40,854 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-11-15 15:42:40,856 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2018-11-15 15:42:40,856 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-11-15 15:42:40,856 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2018-11-15 15:42:40,856 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.InputSizeReducerEstimator
2018-11-15 15:42:40,857 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=352
2018-11-15 15:42:40,857 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2018-11-15 15:42:40,862 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2018-11-15 15:42:40,862 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2018-11-15 15:42:40,865 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2018-11-15 15:42:40,866 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.dir] with code temp directory: /t
mp/1542276760862-0
2018-11-15 15:42:40,886 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2018-11-15 15:42:40,887 [JobControl] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-11-15 15:42:40,899 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2018-11-15 15:42:40,903 [JobControl] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2018-11-15 15:42:40,903 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-11-15 15:42:40,903 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2018-11-15 15:42:40,905 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2018-11-15 15:42:40,905 [JobControl] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2018-11-15 15:42:40,917 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-11-15 15:42:40,917 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2018-11-15 15:42:40,917 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2018-11-15 15:42:40,950 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:2
2018-11-15 15:42:40,971 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_local1015542616_0069
2018-11-15 15:42:41,103 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8080/
2018-11-15 15:42:41,103 [Thread-1919] INFO  org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter set in config null
2018-11-15 15:42:41,108 [Thread-1919] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
```



```
2018-11-15 17:01:19,851 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend
2018-11-15 17:01:19,869 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Tot
2018-11-15 17:01:19,870 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUt
(101,Amitabh)
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
```

(e) List of employees (employee id and employee name) having no entry in employee_expenses file.

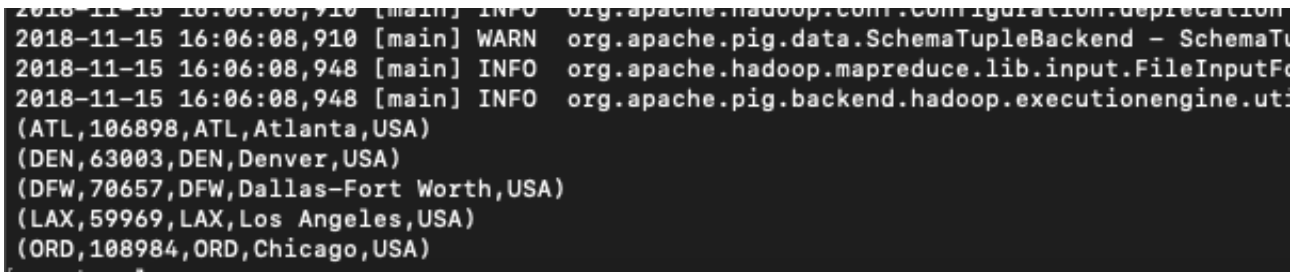
```
employees = load 'employee_details.txt' Using PigStorage(',') as
(empID:int,name:Chararray,salary:int,departmentID:int);
expenses = load 'employee_expenses.txt' Using PigStorage() as (empID:int,expense:int);
joinedEmpLeft = join employees by empID LEFT OUTER, expenses by empID;
filterEmp = filter joinedEmpLeft by expenses::expense is null;
emp = foreach filterEmp generate employees::empID, employees::name;
distinctEmp = DISTINCT emp;
DUMP distinctEmp;
```

### Task 1.3 Aviation Data Analysis

1. Find out the top 5 most visited destinations.

```
REGISTER '/home/acadgild/manish/piggybank.jar';
A = load '/home/acadgild/manish/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray)
$18 as dest;
C = filter B by dest is not null;
D = group C by dest;
E = foreach D generate group, COUNT(C.dest);
F = order E by $1 DESC;
Result = LIMIT F 5;
A1 = load '/home/acadgild/manish/airports.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
joined_table = join Result by $0, A2 by dest;
dump joined_table;
```

2. Which month has seen the most number of cancellations due to bad weather?

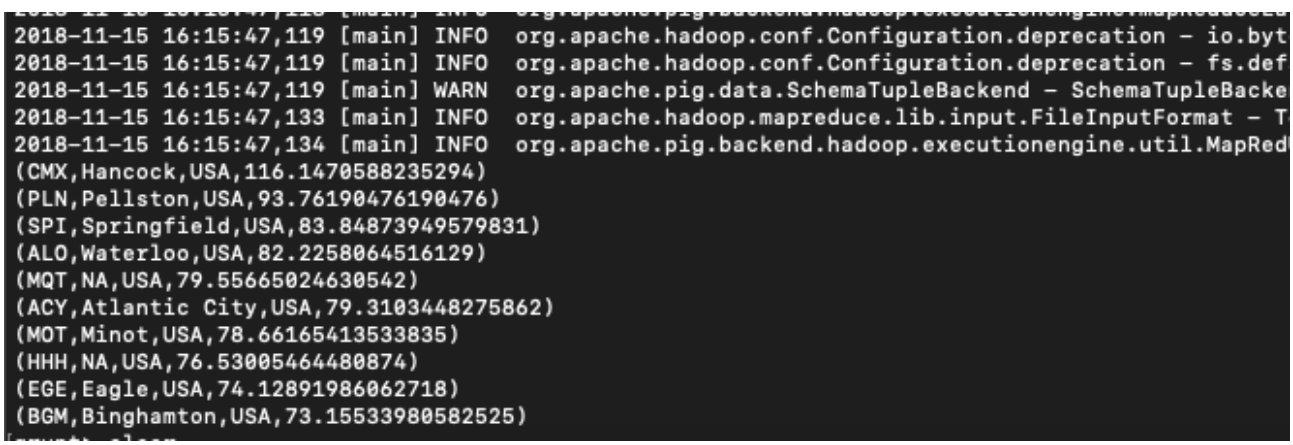
```
REGISTER '/home/acadgild/manish/piggybank.jar';
A = load '/home/acadgild/manish/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23
as cancel_code;
C = filter B by cancelled == 1 AND cancel_code =='B';
D = group C by month;
E = foreach D generate group, COUNT(C.cancelled);
F = order E by $1 DESC;
Result = limit F 1;
dump Result;
```

```
grunt> REGISTER '/home/acadgild/manish/piggybank.jar';
2018-11-15 16:11:05,350 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 16:11:05,350 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> A = load '/home/acadgild/manish/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-11-15 16:11:05,575 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 16:11:05,575 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
grunt> C = filter B by cancelled == 1 AND cancel_code =='B';
grunt> D = group C by month;
grunt> E = foreach D generate group, COUNT(C.cancelled);
grunt> F= order E by $1 DESC;
grunt> Result = limit F 1;
grunt> dump Result;
2018-11-15 16:11:08,787 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY,FILTER,LIMIT
2018-11-15 16:11:08,843 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 16:11:08,843 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-15 16:11:08,844 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-11-15 16:11:08,844 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOpt
imizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-11-15 16:11:08,855 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-11-15 16:11:08,858 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2018-11-15 16:11:08,863 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-2076
2018-11-15 16:11:08,863 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4
2018-11-15 16:11:08,863 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 4
2018-11-15 16:11:08,894 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 16:11:08,898 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-15 16:11:08,900 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-11-15 16:11:08,910 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2018-11-15 16:11:08,910 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-11-15 16:11:08,910 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2018-11-15 16:11:08,910 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.InputSizeReducerEstimator
2018-11-15 16:11:08,910 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=247963212
2018-11-15 16:11:08,910 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2018-11-15 16:11:08,914 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2018-11-15 16:11:08,918 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2018-11-15 16:11:08,919 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2018-11-15 16:11:08,919 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.dir] with code temp directory: /t
mp/1542278468914-0
2018-11-15 16:11:08,951 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2018-11-15 16:11:08,953 [JobControl] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-11-15 16:11:08,960 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2018-11-15 16:11:08,971 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-11-15 16:11:08,971 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2018-11-15 16:11:08,971 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 8
2018-11-15 16:11:09,003 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:8
2018-11-15 16:11:09,028 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_local1169861982_0092
2018-11-15 16:11:09,143 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8080/
2018-11-15 16:11:09,144 [Thread-2554] INFO  org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter set in config null
2018-11-15 16:11:09,149 [Thread-2554] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-11-15 16:11:09,149 [Thread-2554] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.perc
ent
2018-11-15 16:11:09,149 [Thread-2554] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-15 16:11:09,150 [Thread-2554] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
```

```
2018-11-15 16:11:32,141 [main] INFO  org.apache.pig.bac
2018-11-15 16:11:32,141 [main] INFO  org.apache.hadoop.
2018-11-15 16:11:32,141 [main] INFO  org.apache.hadoop.
2018-11-15 16:11:32,146 [main] WARN  org.apache.pig.dat
2018-11-15 16:11:32,163 [main] INFO  org.apache.hadoop.
2018-11-15 16:11:32,163 [main] INFO  org.apache.pig.bac
(12,250)
```

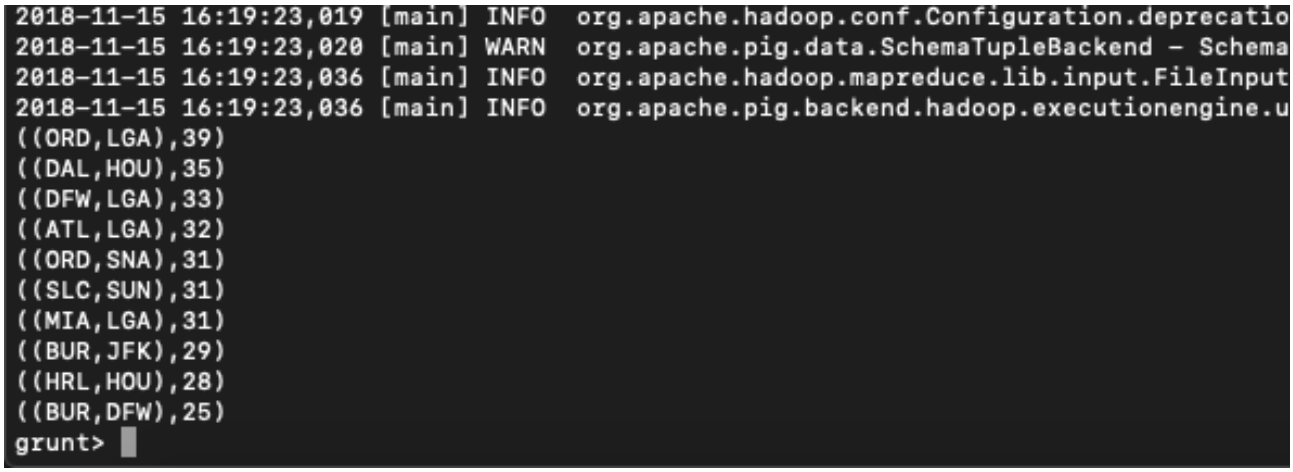
3. Top ten origins with the highest AVG departure delay

```
REGISTER '/home/acadgild/manish/piggybank.jar';
A = load '/home/acadgild/manish/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B1 = foreach A generate (int)$16 as dep_delay, (chararray)$17 as origin;
C1 = filter B1 by (dep_delay is not null) AND (origin is not null);
D1 = group C1 by origin;
E1 = foreach D1 generate group, AVG(C1.dep_delay);
Result = order E1 by $1 DESC;
Top_ten = limit Result 10;
Lookup = load '/home/acadgild/manish/airports.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as
country;
Joined = join Lookup1 by origin, Top_ten by $0;
Final = foreach Joined generate $0,$1,$2,$4;
Final_Result = ORDER Final by $3 DESC;
dump Final_Result;
```

4. Which route (origin & destination) has seen the maximum diversion?

```
REGISTER '/home/acadgild/manish/piggybank.jar';
A = load '/home/acadgild/manish/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;
C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);
D = GROUP C by (origin,dest);
E = FOREACH D generate group, COUNT(C.diversion);
F = ORDER E BY $1 DESC;
Result = limit F 10;
dump Result;
```

**Task 2.2**

Create a database named 'custom'.
Create a table named temperature_data inside custom having below fields:
1. date (mm-dd-yyyy) format
2. zip code
3. temperature
The table will be loaded from comma-delimited file.
Load the dataset.txt (which is ',' delimited) in the table.

```
hive> create database custom location ‘/home/hive';
hive> use custom;
hive> create table temperature_data_temp (temp_date string, zip_code integer, temperature integer)
row format delimited fields terminated by ',' location ‘/temp';
hive> load data local inpath '/home/acadgild/manish/temperature_data.txt' into table
custom.temperature_data_temp;
hive> create table temperature_data ( temp_date date, zip_code integer, temperature integer) row
format delimited fields terminated by ',' location ‘/home/hive';
hive> insert into table custom.temperature_data select from_unixtime(unix_timestamp(temp_date, 'dd-
MM-yyyy')), zip_code, temperature from custom.temperature_data_temp;
hive> describe custom.temperature_data;
```

```
hive> describe custom.temperature_data;
OK
temp_date               date
zip_code                int
temperature             int
Time taken: 0.19 seconds, Fetched: 3 row(s)
```

```
hive> select * from custom.temperature_data;
OK
1990-01-10      123112  10
1991-02-14      283901  11
1990-03-10      381920  15
1991-01-10      302918  22
1990-02-12      384902  9
1991-01-10      123112  11
1990-02-14      283901  12
1991-03-10      381920  16
1990-01-10      302918  23
1991-02-12      384902  10
1993-01-10      123112  11
1994-02-14      283901  12
1993-03-10      381920  16
1994-01-10      302918  23
1991-02-12      384902  10
1991-01-10      123112  11
1990-02-14      283901  12
1991-03-10      381920  16
1990-01-10      302918  23
1991-02-12      384902  10
Time taken: 0.514 seconds, Fetched: 20 row(s)
```

**Task 2.2**

**1. Fetch date and temperature from temperature_data where zip code is greater than 300000 and less than 399999.**

```
hive> select * from custom.temperature_data where zip_code > 300000 and zip_code < 399999;
```

```
hive> select * from custom.temperature_data where zip_code > 300000 and zip_code < 399999;
OK
1990-03-10      381920  15
1991-01-10      302918  22
1990-02-12      384902  9
1991-03-10      381920  16
1990-01-10      302918  23
1991-02-12      384902  10
1993-03-10      381920  16
1994-01-10      302918  23
1991-02-12      384902  10
1991-03-10      381920  16
1990-01-10      302918  23
1991-02-12      384902  10
Time taken: 0.986 seconds, Fetched: 12 row(s)
```

2. Calculate maximum temperature corresponding to every year from temperature_data table.

```
hive> select date_format(temp_date, "yyyy"), max(temperature) from custom.temperature_data group by date_format(temp_date, "yyyy");
```

```
hive> select date_format(temp_date, "yyyy"), max(temperature) from custom.temperature_data group by date_format(temp_date, "yyyy");
WARNING: 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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181115181947_6232d27e-1e88-4697-a8a2-19ab9c0fe5c3
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_1542275989468_0010, Tracking URL = http://localhost:8088/proxy/application_1542275989468_0010/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1542275989468_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-11-15 18:20:04,637 Stage-1 map = 0%,  reduce = 0%
2018-11-15 18:20:18,840 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.28 sec
2018-11-15 18:20:34,225 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.64 sec
MapReduce Total cumulative CPU time: 7 seconds 640 msec
Ended Job = job_1542275989468_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.64 sec   HDFS Read: 9162 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 640 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 49.423 seconds, Fetched: 4 row(s)
```

3. Calculate maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table.

```
hive> Select year, max(temperature) from ( select year ,zip_code, temperature, dense_rank() over (PARTITION by year order by temperature) as rank from (select date_format(temp_date, "yyyy") as year ,zip_code, temperature from custom.temperature_data) temp ) temp1 where rank > 2 group by year;
```

```
hive> Select year, max(temperature) from (
    > select year ,zip_code, temperature, dense_rank() over (PARTITION by year order by temperature) as rank from (se
ure_data) temp ) temp1 where rank > 2 group by year;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a differe
Query ID = acadgild_20181115184708_677a426c-300d-4f9e-9ff0-b16958aa263c
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_1542275989468_0014, Tracking URL = http://localhost:8088/proxy/application_1542275989468_0014/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1542275989468_0014
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-11-15 18:47:25,848 Stage-1 map = 0%,  reduce = 0%
2018-11-15 18:47:37,800 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.26 sec
2018-11-15 18:47:54,300 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.59 sec
MapReduce Total cumulative CPU time: 7 seconds 860 msec
Ended Job = job_1542275989468_0014
Launching Job 2 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_1542275989468_0015, Tracking URL = http://localhost:8088/proxy/application_1542275989468_0015/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1542275989468_0015
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-11-15 18:48:17,436 Stage-2 map = 0%,  reduce = 0%
2018-11-15 18:48:29,685 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.03 sec
2018-11-15 18:48:45,073 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.42 sec
MapReduce Total cumulative CPU time: 5 seconds 420 msec
Ended Job = job_1542275989468_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.86 sec   HDFS Read: 10868 HDFS Write: 142 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.42 sec   HDFS Read: 5687 HDFS Write: 127 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 280 msec
OK
1990    23
1991    22
```

4. Create a view on the top of last query, name it temperature_data_vw.

```
hive> create view temperature_data_vw as Select year, max(temperature) from ( select year ,zip_code,
temperature, dense_rank() over (PARTITION by year order by temperature) as rank from (select
date_format(temp_date, "yyyy") as year ,zip_code, temperature from custom.temperature_data) temp )
temp1 where rank > 2 group by year;
```

5. Export contents from temperature_data_vw to a file in local file system, such that each file is '|' delimited.

```
hive> insert overwrite local directory '/home/acadgild/manish/hive_export' row format delimited
fields terminated by '|' select * from temperature_data_vw;
```

```
[[acadgild@localhost ~]$ cd /home/acadgild/manish/hive_export/
[[acadgild@localhost hive_export]$ ls -ltr
total 4
-rw-r--r--. 1 acadgild acadgild 16 Nov 15 18:59 000000_0
[[acadgild@localhost hive_export]$ cat 000000_0
1990|23
1991|22
[acadgild@localhost hive_export]$
```

```
Time taken: 21.852 seconds
hive> insert overwrite local directory '/home/acadgild/manish/hive_export' row format delimited fields terminated by '|' select * from temperature_data_vw;
WARNING: 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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181115185726_a5419d10-6c65-403f-bd05-998e972d0a83
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_1542275989468_0016, Tracking URL = http://localhost:8088/proxy/application_1542275989468_0016/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1542275989468_0016
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-11-15 18:57:44,083 Stage-1 map = 0%,  reduce = 0%
2018-11-15 18:57:56,965 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.72 sec
2018-11-15 18:58:13,280 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 7.82 sec
2018-11-15 18:58:14,350 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.49 sec
MapReduce Total cumulative CPU time: 8 seconds 490 msec
Ended Job = job_1542275989468_0016
Launching Job 2 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_1542275989468_0017, Tracking URL = http://localhost:8088/proxy/application_1542275989468_0017/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1542275989468_0017
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-11-15 18:58:35,610 Stage-2 map = 0%,  reduce = 0%
2018-11-15 18:58:47,331 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.0 sec
2018-11-15 18:59:01,821 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.23 sec
MapReduce Total cumulative CPU time: 5 seconds 230 msec
Ended Job = job_1542275989468_0017
Moving data to local directory /home/acadgild/manish/hive_export
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.49 sec   HDFS Read: 10898 HDFS Write: 142 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.23 sec   HDFS Read: 5314 HDFS Write: 16 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 720 msec
OK
Time taken: 97.775 seconds
hive>
```