Assignment for Session 7: EXPLORING APACHE PIG

- Prachi Mohite

Task 1

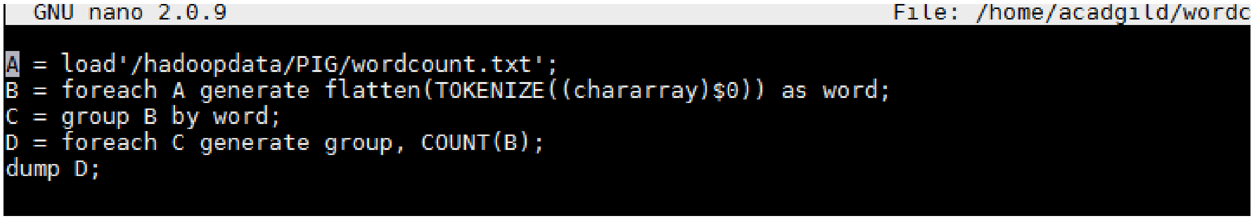
Write a program to implement wordcount using Pig.

Solution Approach

- Load the file in Pig relation separated with tab (space)
  - This can be achieved with built in function Load
  - With this load function, the default function named as ‘PigStorage’ is used to load the data as structured text files. With this function we can pass the field_delimeter. The default is tab ‘/t’.
- Iterate through the Pig relation to get the array of words i.e. bag of words.
  - This can be achieved using TOKENIZE built in function of PIG
  - Use the TOKENIZE function to split a string of words (all words in a single tuple) into a bag of words (each word in a single tuple). The following characters are considered to be word separators: space, double quote("), coma(,) parenthesis(()), star(*). Here the separator is space.

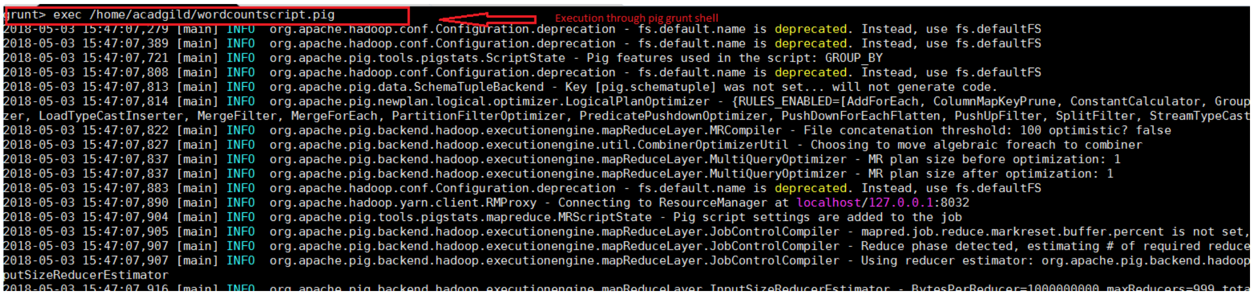
  - The nested bag of tuples can be un-nested using FLATTEN operator.
- Once bag of words is loaded in relation, group those words to get the word count
  - This can be achieved using GROUP BY Operator and COUNT built in function
  - COUNT - the COUNT function to compute the number of elements in a bag. COUNT requires a preceding GROUP ALL statement for global counts and a GROUP BY statement for group counts.
  - GROUP - Groups the data in one or more relations.

Pig Script



```
  GNU nano 2.0.9                                      File: /home/acadgild/wordc

A = load'/hadoopdata/PIG/wordcount.txt';
B = foreach A generate flatten(TOKENIZE((chararray)$0)) as word;
C = group B by word;
D = foreach C generate group, COUNT(B);
dump D;
```

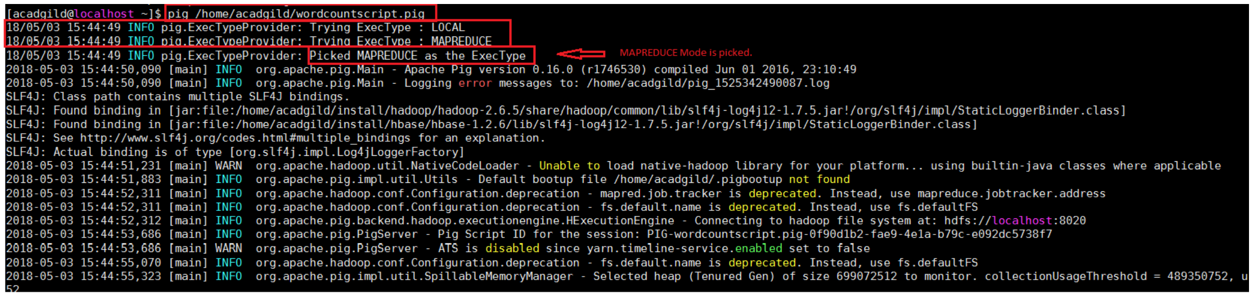
- Once pig script is ready , can be executed from pig grunt shell with exec command. For this need to run pig shell in required mode
  - Pig – runs in HDFS mode MAPREDUCE Mode
  - Pig – x Local – runs in local mode



Execution through pig grunt shell

```
grunt> exec /home/acadgild/wordcountscript.pig
2018-05-03 15:47:07,279 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 15:47:07,389 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 15:47:07,721 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2018-05-03 15:47:07,808 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 15:47:07,813 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-05-03 15:47:07,814 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, Group
zer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCast
2018-05-03 15:47:07,822 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-05-03 15:47:07,827 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2018-05-03 15:47:07,837 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2018-05-03 15:47:07,837 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2018-05-03 15:47:07,883 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 15:47:07,890 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-03 15:47:07,904 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2018-05-03 15:47:07,905 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set,
2018-05-03 15:47:07,907 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reduce
2018-05-03 15:47:07,907 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop
putSizeReducerEstimator
2018-05-03 15:47:07,916 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 tota
```

- The pig script can be executed directly from command line
  - By mentioning the pig - pig script name , here if we are not mentioning the execution mode, pig with try executing for LOCAL mode first and then will go for MAPREDUCE Mode

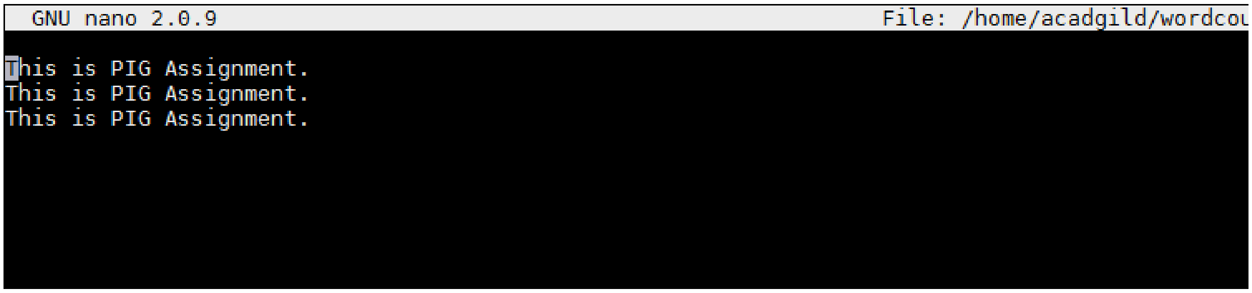


```
[acadgild@localhost ~]$ pig /home/acadgild/wordcountscript.pig
18/05/03 15:44:49 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/05/03 15:44:49 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/05/03 15:44:49 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-05-03 15:44:50,090 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-05-03 15:44:50,090 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1525342490087.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]
2018-05-03 15:44:51,231 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2018-05-03 15:44:51,883 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-05-03 15:44:52,311 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-05-03 15:44:52,311 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 15:44:52,312 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:8020
2018-05-03 15:44:53,686 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-wordcountscript.pig-0f90d1b2-fae9-4e1a-b79c-e092dc5738f7
2018-05-03 15:44:53,686 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-05-03 15:44:55,070 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 15:44:55,323 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. collectionUsageThreshold = 489350752, u
52
```

MAPREDUCE Mode is picked.

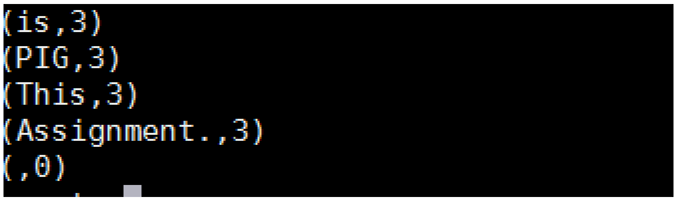
- Modes can be mentioned explicitly
- $ pig -x mapreduce Sample_script.pig
- $ pig -x local Sample_script.pig

Input

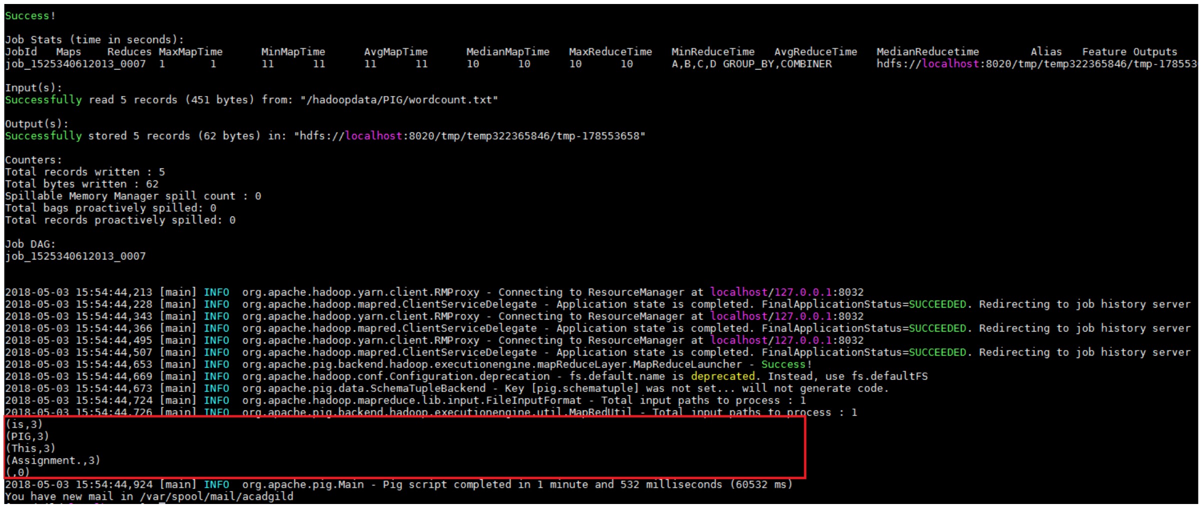


```
  GNU nano 2.0.9                                          File: /home/acadgild/wordcou

This is PIG Assignment.
This is PIG Assignment.
This is PIG Assignment.
```

Output after execution of Pig Script



```
(is,3)
(PIG,3)
(This,3)
(Assignment.,3)
(,0)
```



```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1525340612013_0007  1       1       11      11      11      11      10      10      10      10      A,B,C,D GROUP_BY,COMBINER       hdfs://localhost:8020/tmp/temp322365846/tmp-178553

Input(s):
Successfully read 5 records (451 bytes) from: "/hadoopdata/PIG/wordcount.txt"

Output(s):
Successfully stored 5 records (62 bytes) in: "hdfs://localhost:8020/tmp/temp322365846/tmp-178553658"

Counters:
Total records written : 5
Total bytes written : 62
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1525340612013_0007


2018-05-03 15:54:44,213 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-03 15:54:44,228 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-03 15:54:44,343 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-03 15:54:44,366 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-03 15:54:44,495 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-03 15:54:44,507 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-03 15:54:44,653 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-03 15:54:44,669 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 15:54:44,673 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-05-03 15:54:44,724 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-03 15:54:44,726 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(is,3)
(PIG,3)
(This,3)
(Assignment.,3)
(,0)
2018-05-03 15:54:44,924 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute and 532 milliseconds (60532 ms)
You have new mail in /var/spool/mail/acadgild
```

Task 2

We have employee_details and employee_expenses files. Use local mode while running Pig and write Pig Latin script to get below results:

employee_details (EmpID,Name,Salary,DepartmentID)
https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee_details.txt
employee_expenses(EmpID,Expence)
https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee_expenses.txt

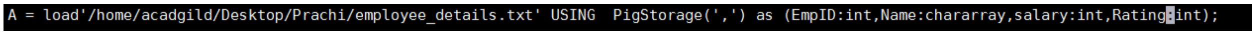
(a) Top 5 employees (employee id and employee name) with highest rating. (In case two employees have same rating, employee with name coming first in dictionary should get preference)

Solution Approach

- Create a script (pig latin) to
- Load the files in pig relation with the schema
    - Schemas enable you to assign names to fields and declare types for fields. Schemas are optional but we encourage you to use them whenever possible;

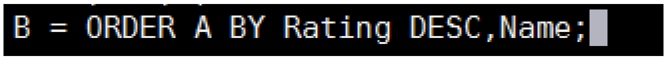
type declarations result in better parse-time error checking and more efficient code execution.

  - Schemas are defined with the LOAD, STREAM, and FOREACH operators using the AS clause



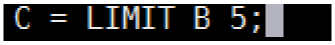
```
A = load'/home/acadgild/Desktop/Prachi/employee_details.txt' USING  PigStorage(',') as (EmpID:int,Name:chararray,salary:int,Rating:int);
```

- Using ORDER BY Relational Operator for Rating and Name get the desired output
  - Sorts a relation based on one or more fields.



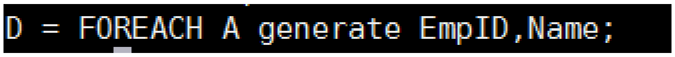
```
B = ORDER A BY Rating DESC,Name;
```

- To get top 5 , limit the output with LIMIT – the relational Operator



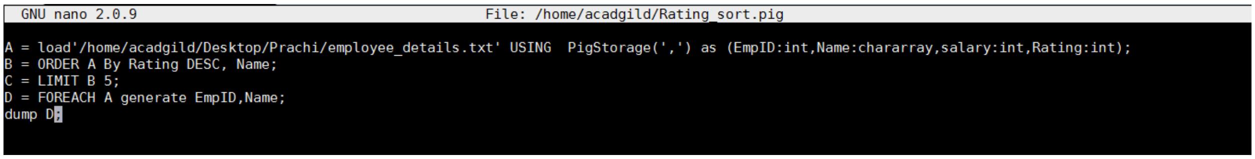
```
C = LIMIT B 5;
```

- Selecting the Employee id and employee name with below



```
D = FOREACH A generate EmpID,Name;
```

Pig script



```
  GNU nano 2.0.9                              File: /home/acadgild/Rating_sort.pig

A = load'/home/acadgild/Desktop/Prachi/employee_details.txt' USING  PigStorage(',') as (EmpID:int,Name:chararray,salary:int,Rating:int);
B = ORDER A By Rating DESC, Name;
C = LIMIT B 5;
D = FOREACH A generate EmpID,Name;
dump D;
```

Input files placed at

- /home/acadgild/Desktop/Prachi/employee_details.txt
- /home/acadgild/Desktop/Prachi/employee_expenses.txt

Local Mode of Pig Script

We can run Apache Pig in two modes, namely, Local Mode and HDFS mode.

Local Mode

In this mode, all the files are installed and run from your local host and local file system. There is no need of Hadoop or HDFS. This mode is generally used for testing purpose.

MapReduce Mode

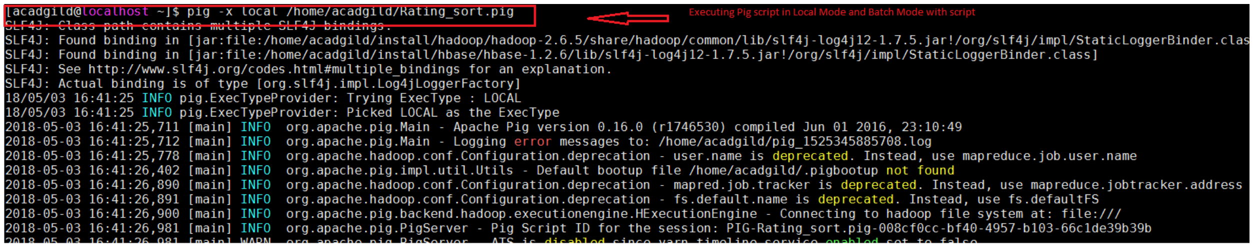
MapReduce mode is where we load or process the data that exists in the Hadoop File System (HDFS) using Apache Pig. In this mode, whenever we execute the Pig Latin statements to process the data, a MapReduce job is invoked in the back-end to perform a particular operation on the data that exists in the HDFS.

Apache Pig Execution Mechanisms

Apache Pig scripts can be executed in three ways, namely, interactive mode, batch mode, and embedded mode.

- Interactive Mode (Grunt shell) − You can run Apache Pig in interactive mode using the Grunt shell. In this shell, you can enter the Pig Latin statements and get the output (using Dump operator).
- Batch Mode (Script) − You can run Apache Pig in Batch mode by writing the Pig Latin script in a single file with .pig extension.
- Embedded Mode (UDF) − Apache Pig provides the provision of defining our own functions (User Defined Functions) in programming languages such as Java, and using them in our script.

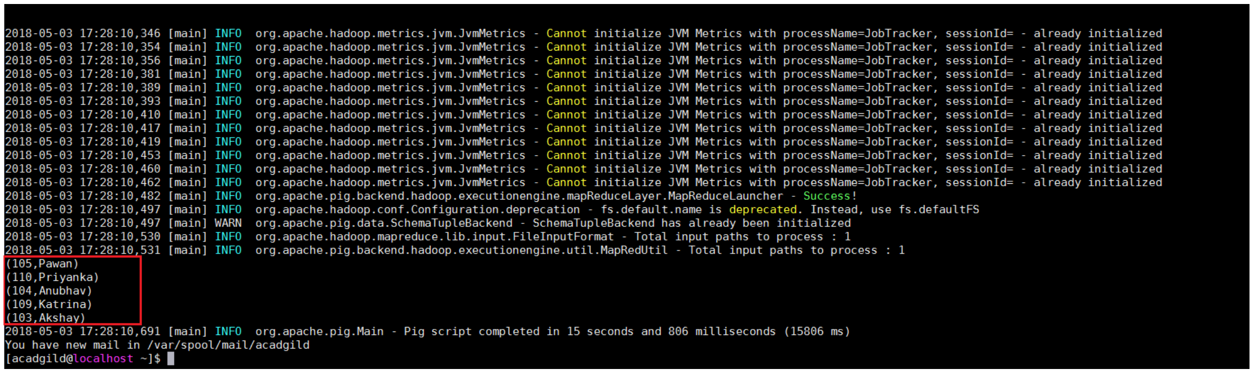
Execution of Script



Executing Pig script in Local Mode and Batch Mode with script

```
[acadgild@localhost ~]$ pig -x local /home/acadgild/Rating_sort.pig
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.clas
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]
18/05/03 16:41:25 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/05/03 16:41:25 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-05-03 16:41:25,711 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-05-03 16:41:25,712 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1525345885708.log
2018-05-03 16:41:25,778 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2018-05-03 16:41:26,402 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-05-03 16:41:26,890 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-05-03 16:41:26,891 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 16:41:26,900 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2018-05-03 16:41:26,981 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-Rating_sort.pig-008cf0cc-bf40-4957-b103-66c1de39b39b
2018-05-03 16:41:26,981 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn timeline service enabled set to false
```

Output of script



```
2018-05-03 17:28:10,346 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:28:10,354 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:28:10,356 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:28:10,381 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:28:10,389 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:28:10,393 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:28:10,410 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:28:10,417 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:28:10,419 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:28:10,453 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:28:10,460 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:28:10,462 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:28:10,482 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-03 17:28:10,497 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 17:28:10,497 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-03 17:28:10,530 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-03 17:28:10,531 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(105,Pawan)
(110,Priyanka)
(104,Anubhav)
(109,Katrina)
(103,Akshay)
2018-05-03 17:28:10,691 [main] INFO  org.apache.pig.Main - Pig script completed in 15 seconds and 806 milliseconds (15806 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ 
```

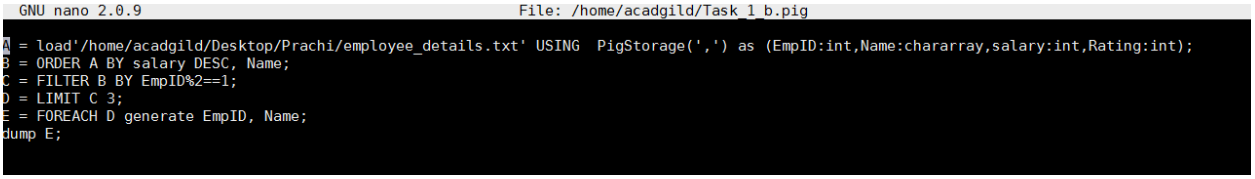
Task 2

(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)

Solution Approach –

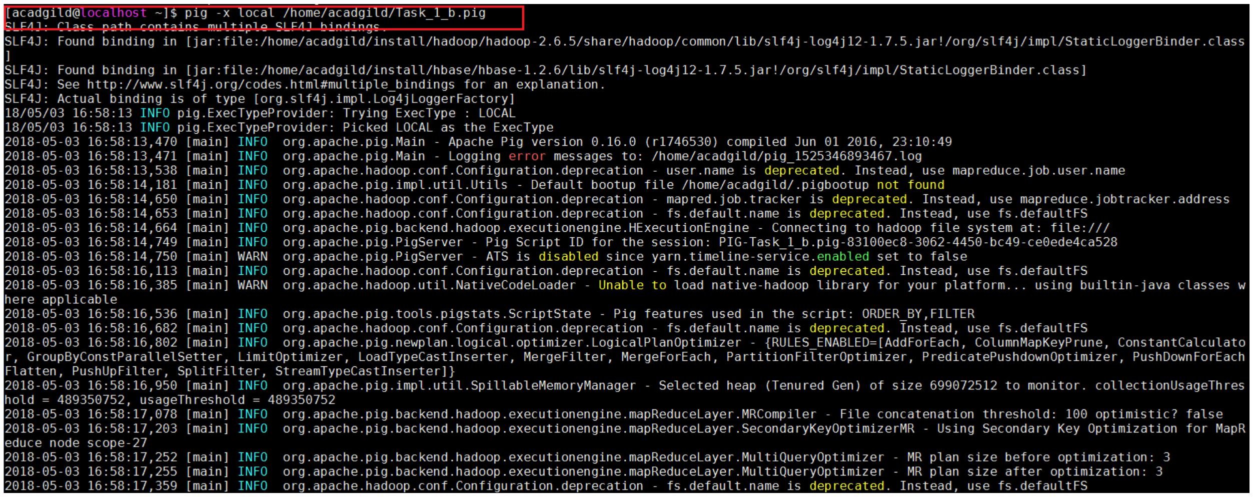
- Load the employee_details.txt in a relation with schema
- Order the relation by Salary in descending order and name in ascending order using ORDER BY
- To get the odd employee number use mathematical operation mod (%), if EmpID % 2 is equals to 1 then the number is odd. The condition can be checked using FILTER operator.
- We have to get top 3 highly paid employees, so we have to get TOP 3 , as relation is in descending order of salary we can limit the output to 3 by using LIMIT operator.

Input Script



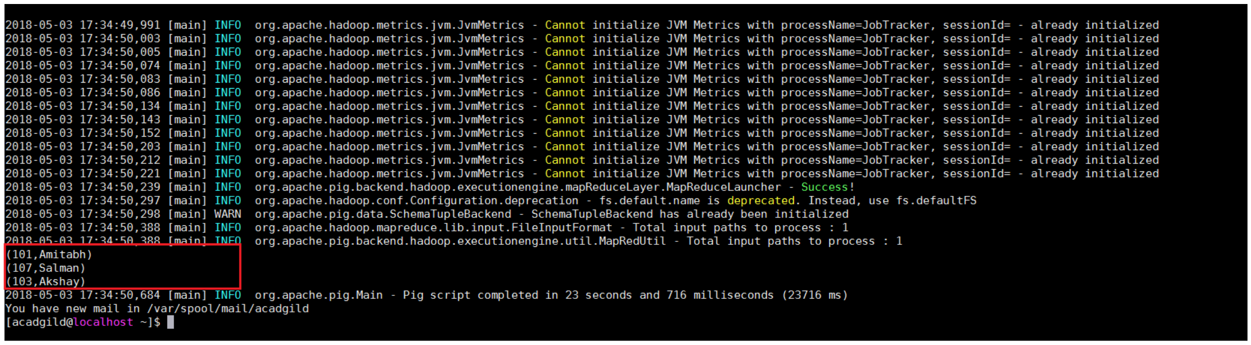
```
  GNU nano 2.0.9                              File: /home/acadgild/Task_1_b.pig

A = load'/home/acadgild/Desktop/Prachi/employee_details.txt' USING  PigStorage(',') as (EmpID:int,Name:chararray,salary:int,Rating:int);
B = ORDER A BY salary DESC, Name;
C = FILTER B BY EmpID%2==1;
D = LIMIT C 3;
E = FOREACH D generate EmpID, Name;
dump E;
```

Execution of Script



```
[acadgild@localhost ~]$ pig -x local /home/acadgild/Task_1_b.pig
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]
18/05/03 16:58:13 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/05/03 16:58:13 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-05-03 16:58:13,470 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-05-03 16:58:13,471 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1525346893467.log
2018-05-03 16:58:13,538 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2018-05-03 16:58:14,181 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-05-03 16:58:14,650 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-05-03 16:58:14,653 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 16:58:14,664 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2018-05-03 16:58:14,749 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-Task_1_b.pig-83100ec8-3062-4450-bc49-ce0ede4ca528
2018-05-03 16:58:14,750 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-05-03 16:58:16,113 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 16:58:16,385 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
2018-05-03 16:58:16,536 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,FILTER
2018-05-03 16:58:16,682 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 16:58:16,802 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculato
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEach
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-05-03 16:58:16,950 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. collectionUsageThres
hold = 489350752, usageThreshold = 489350752
2018-05-03 16:58:17,078 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-05-03 16:58:17,203 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapR
educe node scope-27
2018-05-03 16:58:17,252 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 3
2018-05-03 16:58:17,255 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 3
2018-05-03 16:58:17,359 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```

Output



```
2018-05-03 17:34:49,991 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:34:50,003 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:34:50,005 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:34:50,074 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:34:50,083 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:34:50,086 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:34:50,134 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:34:50,143 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:34:50,152 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:34:50,203 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:34:50,212 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:34:50,221 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-03 17:34:50,239 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-03 17:34:50,297 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 17:34:50,298 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-03 17:34:50,388 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-03 17:34:50,388 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Amitabh)
(107,Salman)
(103,Akshay)
2018-05-03 17:34:50,684 [main] INFO  org.apache.pig.Main - Pig script completed in 23 seconds and 716 milliseconds (23716 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ 
```

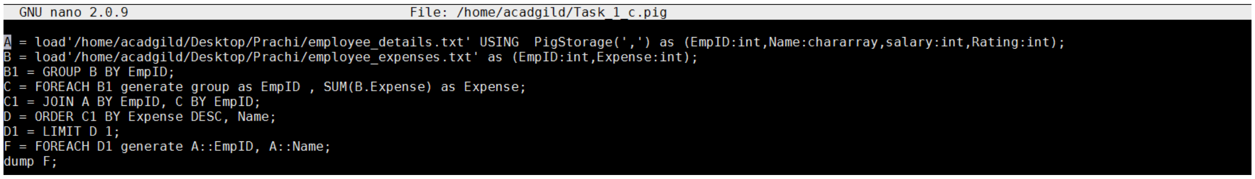
Task 2

(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)

Solution Approach –

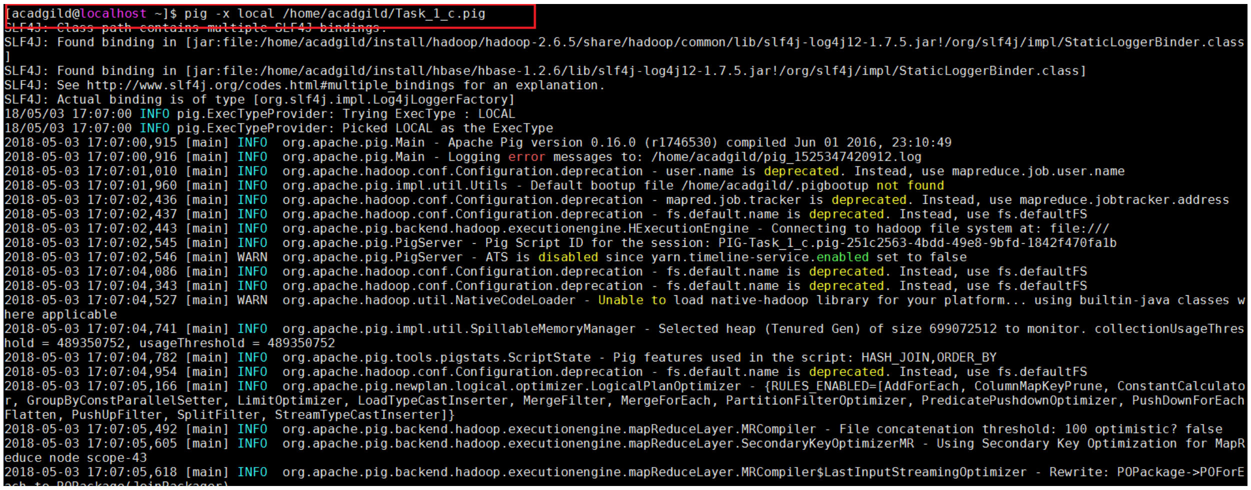
- Load the employee_details.txt and employee_expense in different relations with schema and appropriate field delimiter
  - Employee_details should be loaded using comma
  - Employee_expense should be loaded with tab (which is default delimiter, so no need to specify)
- As expenses have multiple entries for a employee, we need to group employee expenses (using GROUP BY) on employee number and get the SUM of expenses using SUM operator.
- The result of grouped data and employee details need to be clubbed using JOIN operator.
  - Here JOIN will indicate the INNER JOIN where result will be data from both the tables having same employee id .
  - INNER JOIN is done on employee id from both the tables
- As we have to get maximum expense done by employee, order the relation on expenses in descending order
- As we have to get only one employee, LIMIT the output to the 1.
- Generate the required columns – employee ID and Employee Name

Input Script



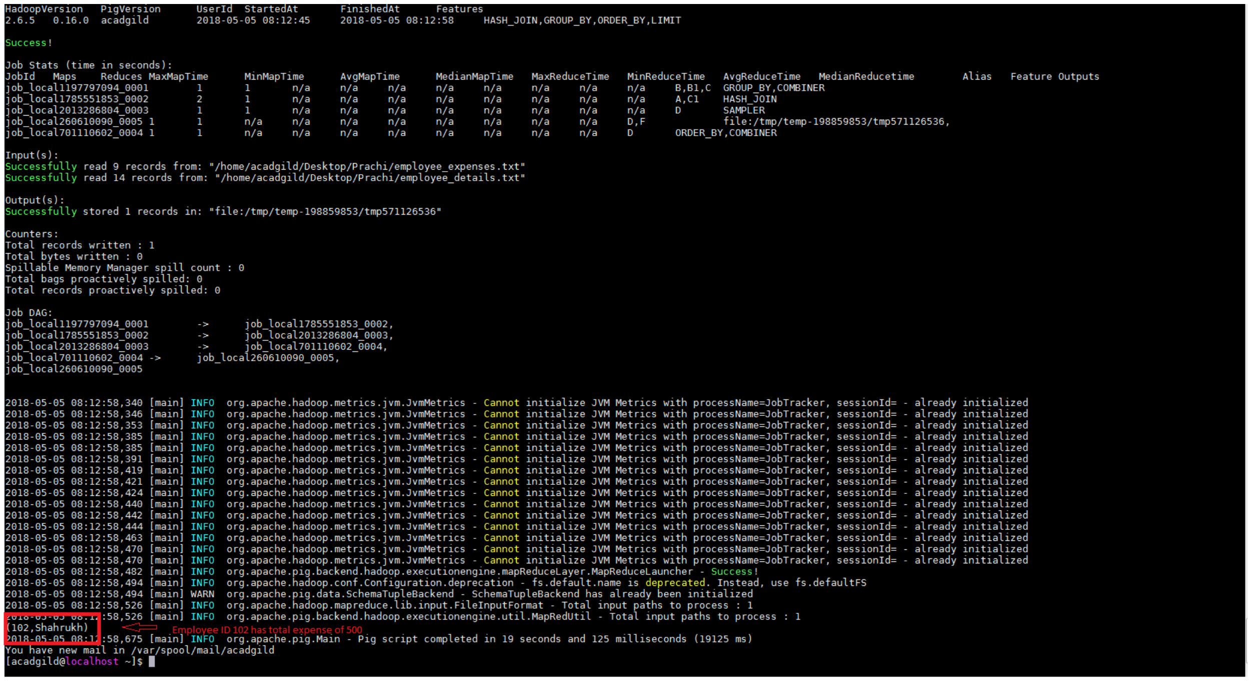
```
  GNU nano 2.0.9                              File: /home/acadgild/Task_1_c.pig

A = load'/home/acadgild/Desktop/Prachi/employee_details.txt' USING  PigStorage(',') as (EmpID:int,Name:chararray,salary:int,Rating:int);
B = load'/home/acadgild/Desktop/Prachi/employee_expenses.txt' as (EmpID:int,Expense:int);
B1 = GROUP B BY EmpID;
C = FOREACH B1 generate group as EmpID , SUM(B.Expense) as Expense;
C1 = JOIN A BY EmpID, C BY EmpID;
D = ORDER C1 BY Expense DESC, Name;
D1 = LIMIT D 1;
F = FOREACH D1 generate A::EmpID, A::Name;
dump F;
```

Execution of script



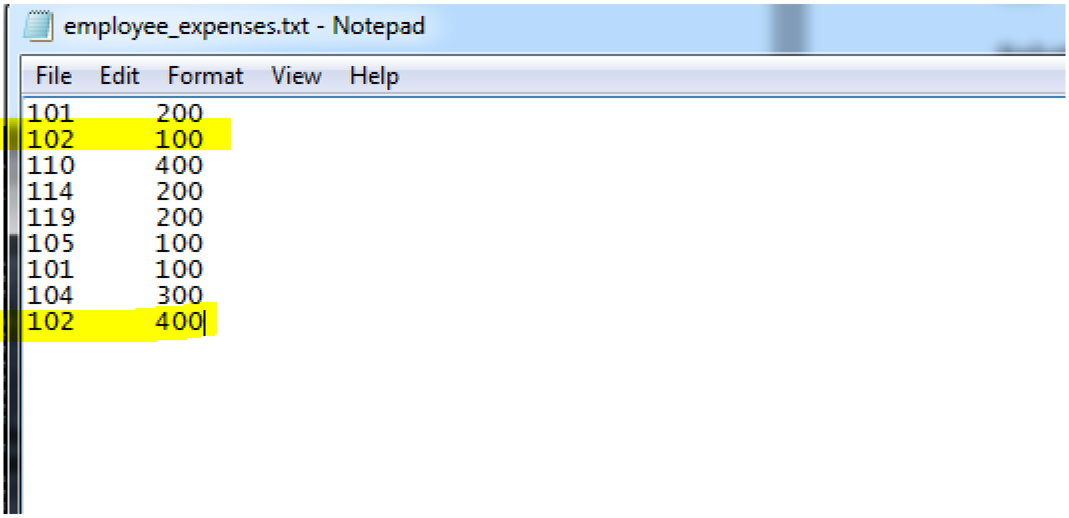
```
[acadgild@localhost ~]$ pig -x local /home/acadgild/Task_1_c.pig
~~SLF4J: Class path contains multiple SLF4J bindings.~~
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class
]
SLF4J: Found 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]
18/05/03 17:07:00 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/05/03 17:07:00 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-05-03 17:07:00,915 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-05-03 17:07:00,916 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1525347420912.log
2018-05-03 17:07:01,010 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2018-05-03 17:07:01,960 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-05-03 17:07:02,436 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-05-03 17:07:02,437 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 17:07:02,443 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2018-05-03 17:07:02,545 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-Task_1_c.pig-251c2563-4bdd-49e8-9bfd-1842f470fa1b
2018-05-03 17:07:02,546 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-05-03 17:07:04,086 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 17:07:04,343 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 17:07:04,527 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
2018-05-03 17:07:04,741 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. collectionUsageThres
hold = 489350752, usageThreshold = 489350752
2018-05-03 17:07:04,782 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,ORDER_BY
2018-05-03 17:07:04,954 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 17:07:05,166 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculato
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEach
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-05-03 17:07:05,492 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-05-03 17:07:05,605 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapR
educe node scope-43
2018-05-03 17:07:05,618 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForE
ach to POPackage(JoinPackager)
```

Output of script



```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.5   0.16.0  acadgild        2018-05-05 08:12:45     2018-05-05 08:12:58     HASH_JOIN,GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local1197797094_0001        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     B,B1,C  GROUP_BY,COMBINER
job_local1785551853_0002        2       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     A,C1    HASH_JOIN
job_local2013286804_0003        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     D       SAMPLER
job_local260610090_0005 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     D,F             file:/tmp/temp-198859853/tmp571126536,
job_local701110602_0004 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     D       ORDER_BY,COMBINER

Input(s):
Successfully read 9 records from: "/home/acadgild/Desktop/Prachi/employee_expenses.txt"
Successfully read 14 records from: "/home/acadgild/Desktop/Prachi/employee_details.txt"

Output(s):
Successfully stored 1 records in: "file:/tmp/temp-198859853/tmp571126536"

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

Job DAG:
job_local1197797094_0001        ->      job_local1785551853_0002,
job_local1785551853_0002        ->      job_local2013286804_0003,
job_local2013286804_0003        ->      job_local701110602_0004,
job_local701110602_0004 ->      job_local260610090_0005,
job_local260610090_0005


2018-05-05 08:12:58,340 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,346 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,353 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,385 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,385 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,391 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,419 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,421 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,424 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,440 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,442 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,444 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,463 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,470 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,470 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-05-05 08:12:58,482 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-05 08:12:58,494 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-05 08:12:58,494 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-05 08:12:58,526 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-05 08:12:58,526 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(102,Shahrukh)
2018-05-05 08:12:58,675 [main] INFO  org.apache.pig.Main - Pig script completed in 19 seconds and 125 milliseconds (19125 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Employee ID 102 has total expense of 500



```
101	200
102	100
110	400
114	200
119	200
105	100
101	100
104	300
102	400
```

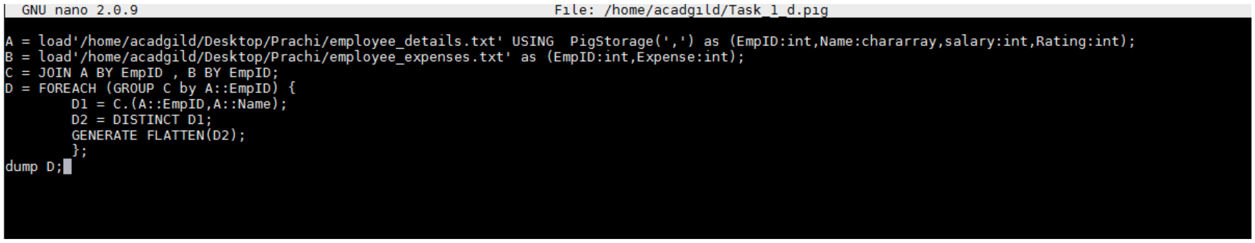
Task 2

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

Solution Approach –

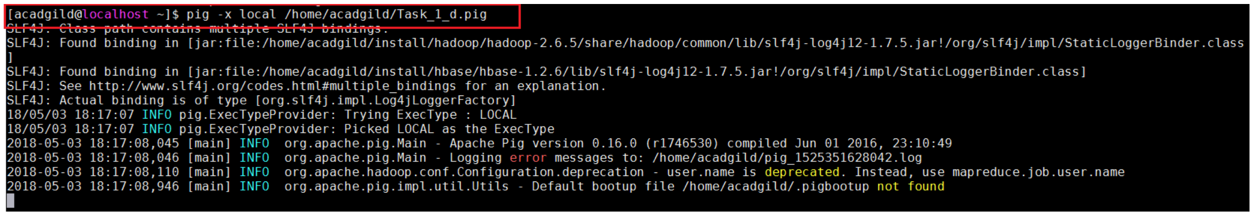
- Load the employee_details.txt and employee_expense in different relations with schema and appropriate field delimiter
    - Employee_details should be loaded using comma
    - Employee_expense should be loaded with tab (which is default delimiter, so no need to specify)
- To get the employees belonging to expense table we can perform
    - INNER JOIN and then get the DISTINCT employee using DISTINCT Operator
- Here we have used FOREACH block

Pig Script



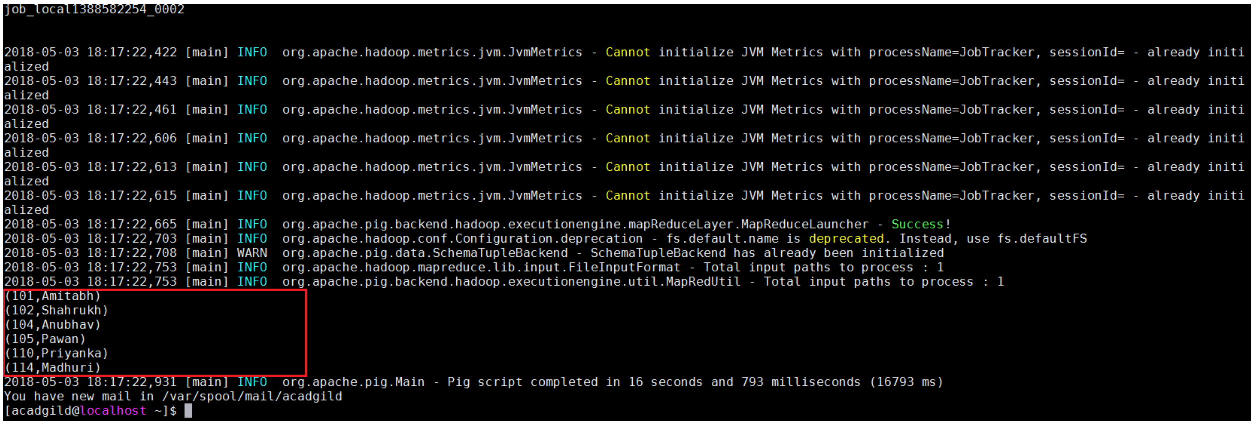
```
A = load'/home/acadgild/Desktop/Prachi/employee_details.txt' USING  PigStorage(',') as (EmpID:int,Name:chararray,salary:int,Rating:int);
B = load'/home/acadgild/Desktop/Prachi/employee_expenses.txt' as (EmpID:int,Expense:int);
C = JOIN A BY EmpID , B BY EmpID;
D = FOREACH (GROUP C by A::EmpID) {
        D1 = C.(A::EmpID,A::Name);
        D2 = DISTINCT D1;
        GENERATE FLATTEN(D2);
        };
dump D;
```

Execution



```
[acadgild@localhost ~]$ pig -x local /home/acadgild/Task_1_d.pig
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]
18/05/03 18:17:07 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/05/03 18:17:07 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-05-03 18:17:08,045 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-05-03 18:17:08,046 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1525351628042.log
2018-05-03 18:17:08,110 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2018-05-03 18:17:08,946 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
```

Output



```
job_local1388582254_0002


2018-05-03 18:17:22,422 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 18:17:22,443 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 18:17:22,461 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 18:17:22,606 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 18:17:22,613 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 18:17:22,615 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 18:17:22,665 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-03 18:17:22,703 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 18:17:22,708 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-03 18:17:22,753 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-03 18:17:22,753 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Amitabh)
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
2018-05-03 18:17:22,931 [main] INFO  org.apache.pig.Main - Pig script completed in 16 seconds and 793 milliseconds (16793 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Task 2

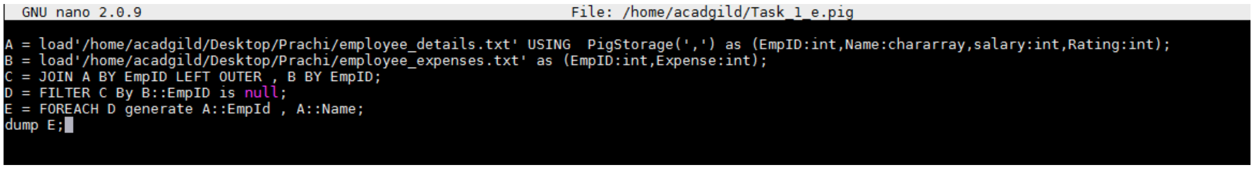
(e) List of employees (employee id and employee name) having no entry in employee_expenses file.

Solution Approach –

- Load the employee_details.txt and employee_expense in different relations with schema and appropriate field delimiter
  - Employee_details should be loaded using comma
  - Employee_expense should be loaded with tab (which is default delimiter, so no need to specify)
- To get employees, not having entries in expense table we have to perform
  - OUTER JOIN
    - LEFT OUTER JOIN – all rows of table which on left hand side of join
    - RIGHT OUTER JOIN – all rows of table which on right hand side of join
  - Here Employee_details will be at left hand side so we will perform LEFT OUTER JOIN

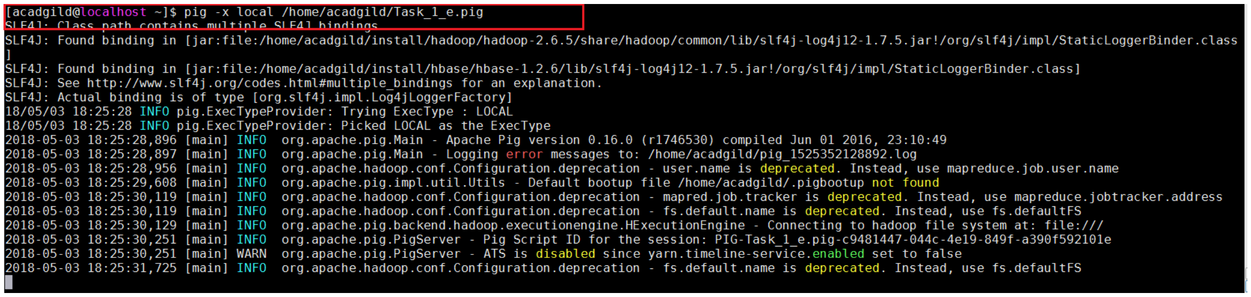
- Once joined check if expenses are null using “is null” condition and get the employees not having entries in employee_expenses.

PIG Script



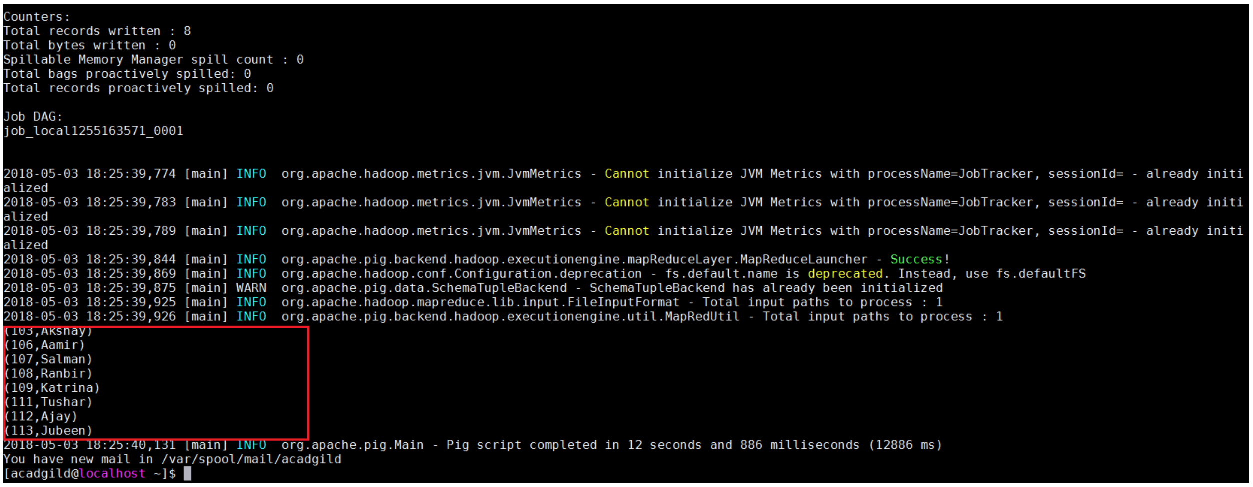
```
  GNU nano 2.0.9                                File: /home/acadgild/Task_1_e.pig

A = load'/home/acadgild/Desktop/Prachi/employee_details.txt' USING  PigStorage(',') as (EmpID:int,Name:chararray,salary:int,Rating:int);
B = load'/home/acadgild/Desktop/Prachi/employee_expenses.txt' as (EmpID:int,Expense:int);
C = JOIN A BY EmpID LEFT OUTER , B BY EmpID;
D = FILTER C By B::EmpID is null;
E = FOREACH D generate A::EmpId , A::Name;
dump E;
```

Execution



```
[acadgild@localhost ~]$ pig -x local /home/acadgild/Task_1_e.pig
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]
18/05/03 18:25:28 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/05/03 18:25:28 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-05-03 18:25:28,896 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-05-03 18:25:28,897 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1525352128892.log
2018-05-03 18:25:28,956 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2018-05-03 18:25:29,608 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-05-03 18:25:30,119 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-05-03 18:25:30,119 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 18:25:30,129 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2018-05-03 18:25:30,251 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-Task_1_e.pig-c9481447-044c-4e19-849f-a390f592101e
2018-05-03 18:25:30,251 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-05-03 18:25:31,725 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```

Output



```
Counters:
Total records written : 8
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1255163571_0001


2018-05-03 18:25:39,774 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 18:25:39,783 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 18:25:39,789 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 18:25:39,844 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-03 18:25:39,869 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 18:25:39,875 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-03 18:25:39,925 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-03 18:25:39,926 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(103,Akshay)
(106,Aamir)
(107,Salman)
(108,Ranbir)
(109,Katrina)
(111,Tushar)
(112,Ajay)
(113,Jubeen)
2018-05-03 18:25:40,131 [main] INFO  org.apache.pig.Main - Pig script completed in 12 seconds and 886 milliseconds (12886 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Task 3

Implement the use case present in below blog link and share the complete steps along with screenshot(s) from your end.

Blog Link

- https://acadgild.com/blog/aviation-data-analysis-using-apache-pig/

Data files can be downloaded from below links

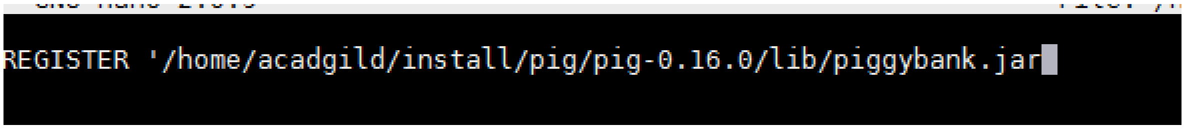
- DelyedFlights.csv
- Airport-Data.csv

Problem Statement 1

Find out the top 5 most visited destinations

Solution Approach

- Here we have to load data from Delayed flights into one relation
  - To achieve the same we have to use User Define functions to read data from CSV files
  - The *Piggy Bank* is a place for Pig users to share their functions
  - These UDFs are available at place '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar
  - We need to register these jar file using REGITSER statement

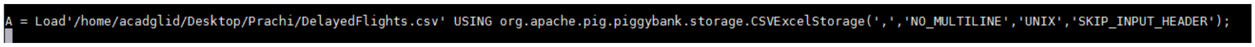


```
REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar
```

Why to use UDF functions from piggybanj.jar

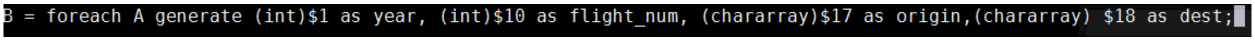
- These UDFs help to read columns headers and data
- Helps to deal with double quotes
- Also if any column is having values encoded in curly brackets , these UDFs helps to read\load them as single column instead of treating it as different columns.

Load data of Delayed Flights into a relation by skipping the column names



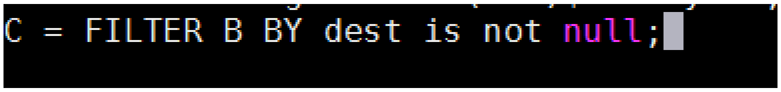
```
A = Load'/home/acadglid/Desktop/Prachi/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
```

Once data is loaded generate the columns names as required



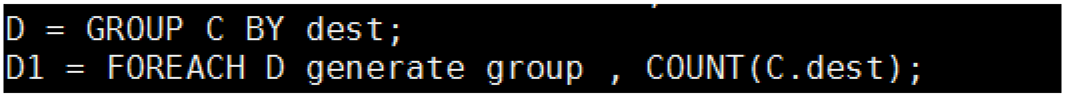
```
B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18 as dest;
```

Need to filter data where destination is not provided (i.e. dest column is not null)



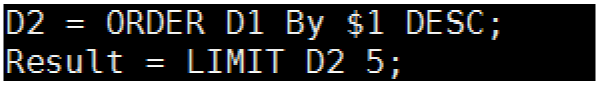
```
C = FILTER B BY dest is not null;
```

Now in order to get mostly visited destinations we need to group the destinations and check for their count.



```
D = GROUP C BY dest;
D1 = FOREACH D generate group , COUNT(C.dest);
```

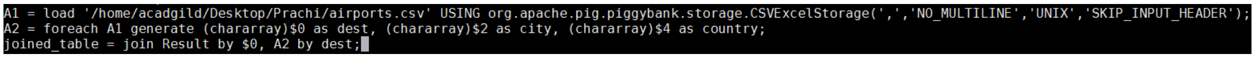
Get the top 5 visited places by using Order by and LIMIT



```
D2 = ORDER D1 By $1 DESC;
Result = LIMIT D2 5;
```

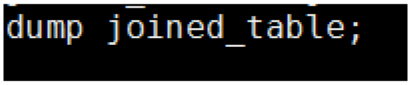
To get the names of places join to airports csv

- Load the airports CSV
- Generate the required columns
- Join on airport names



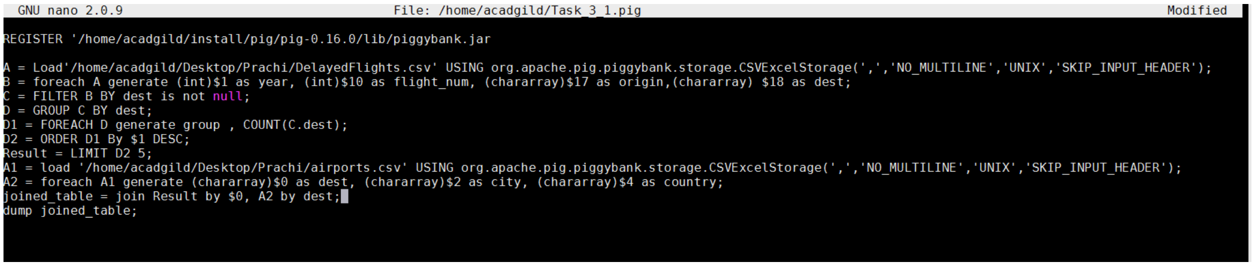
```
A1 = load '/home/acadgild/Desktop/Prachi/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
joined_table = join Result by $0, A2 by dest;
```

Dump the result



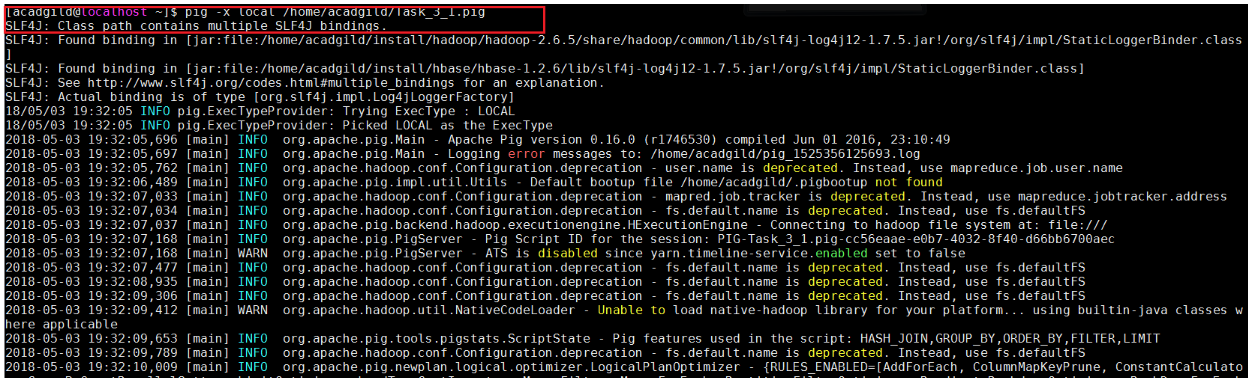
```
dump joined_table;
```

PIG Script



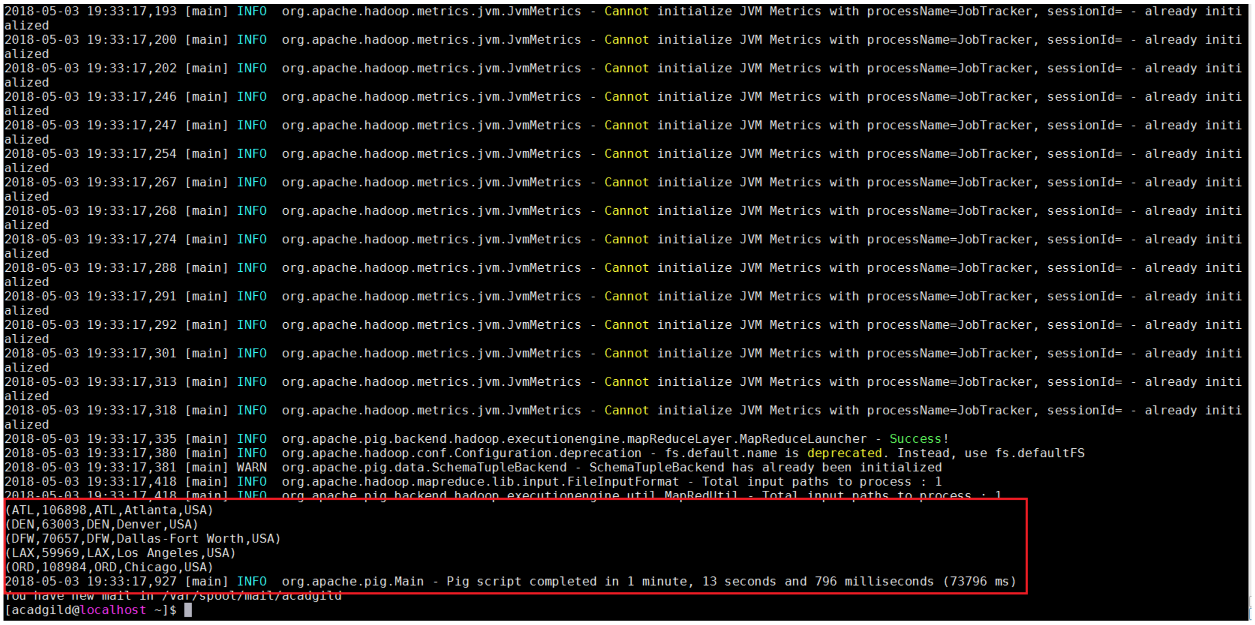
```
  GNU nano 2.0.9                    File: /home/acadgild/Task_3_1.pig                                        Modified

REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar

A = Load'/home/acadgild/Desktop/Prachi/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;
D1 = FOREACH D generate group , COUNT(C.dest);
D2 = ORDER D1 By $1 DESC;
Result = LIMIT D2 5;
A1 = load '/home/acadgild/Desktop/Prachi/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;
```

Execution



```
[acadgild@localhost ~]$ pig -x local /home/acadgild/Task_3_1.pig
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]
18/05/03 19:32:05 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/05/03 19:32:05 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-05-03 19:32:05,696 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-05-03 19:32:05,697 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1525356125693.log
2018-05-03 19:32:05,762 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2018-05-03 19:32:06,489 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-05-03 19:32:07,033 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-05-03 19:32:07,034 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:32:07,037 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2018-05-03 19:32:07,168 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-Task_3_1.pig-cc56eaae-e0b7-4032-8f40-d66bb6700aec
2018-05-03 19:32:07,168 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-05-03 19:32:07,477 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:32:08,935 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:32:09,306 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:32:09,412 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
2018-05-03 19:32:09,653 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,GROUP_BY,ORDER_BY,FILTER,LIMIT
2018-05-03 19:32:09,789 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:32:10,009 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculato
```

Output



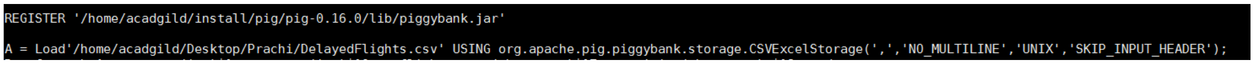
```
2018-05-03 19:33:17,193 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,200 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,202 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,246 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,247 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,254 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,267 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,268 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,274 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,288 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,291 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,292 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,301 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,313 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,318 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:33:17,335 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-03 19:33:17,380 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:33:17,381 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-03 19:33:17,418 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-03 19:33:17,418 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(ATL,106898,ATL,Atlanta,USA)
(DEN,63003,DEN,Denver,USA)
(DFW,70657,DFW,Dallas-Fort Worth,USA)
(LAX,59969,LAX,Los Angeles,USA)
(ORD,108984,ORD,Chicago,USA)
2018-05-03 19:33:17,927 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 13 seconds and 796 milliseconds (73796 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Problem Statement 2

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

Solution Approach

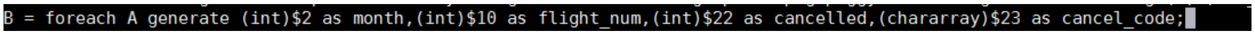
- Load the data from DelayedFlights.csv with help of UDF CSVExcelStorage written in piggybank jar



```
REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar'

A = Load'/home/acadgild/Desktop/Prachi/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
```

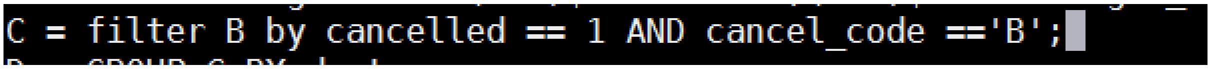
- Once data loaded generate the required columns to get the data
- Month (2)
- Cancellation Code (23)
- Flight number (10)

- Is flight Cancelled (22)



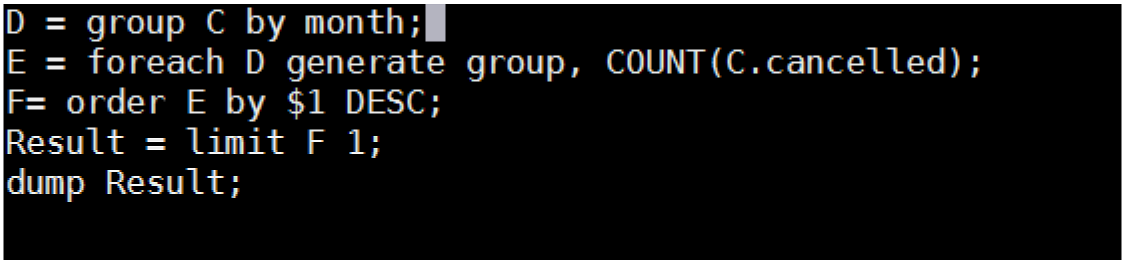
```
B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
```

Once required columns are fetched , filter the data based on Cancellation Code = Weather (b)



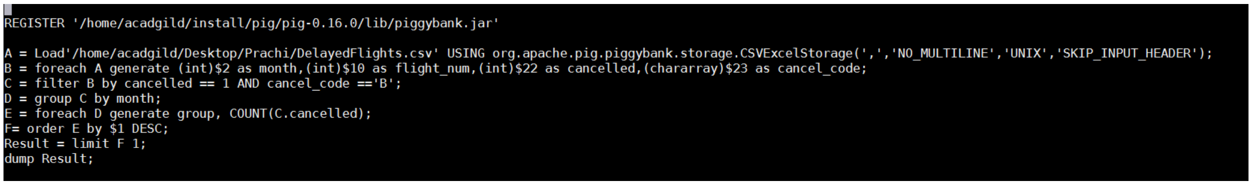
```
C = filter B by cancelled == 1 AND cancel_code =='B';
```

- Group the cancelled flight data by month



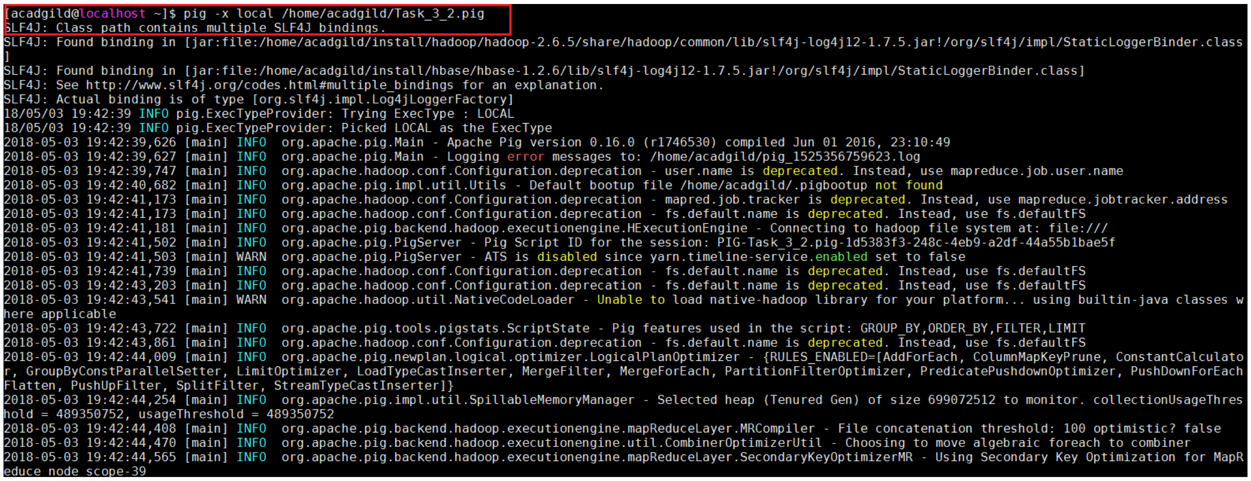
```
D = group C by month;
E = foreach D generate group, COUNT(C.cancelled);
F= order E by $1 DESC;
Result = limit F 1;
dump Result;
```

PIG script



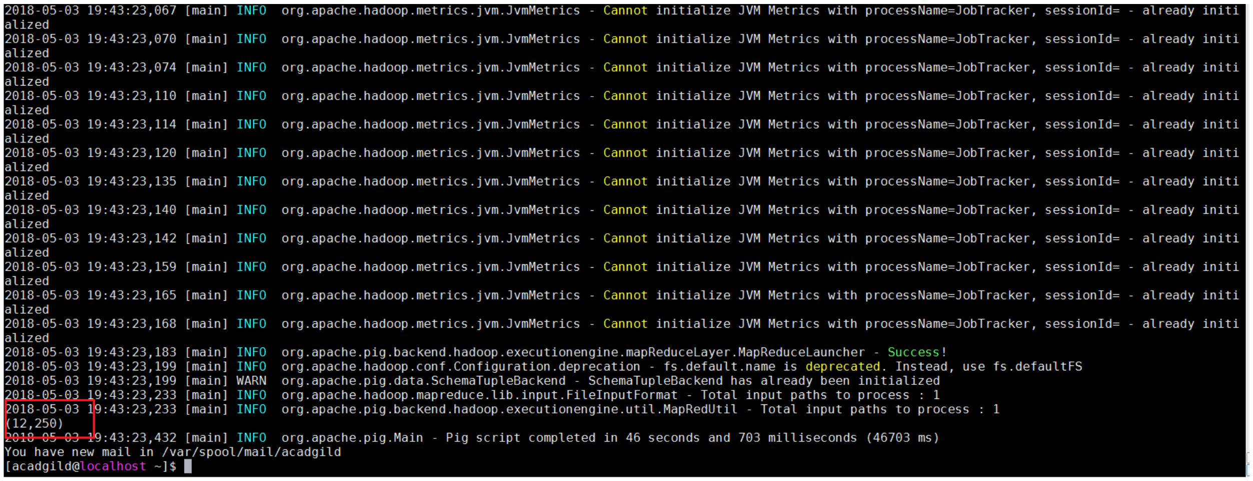
```
REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar'

A = Load'/home/acadgild/Desktop/Prachi/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;
```

Execution



```
[acadgild@localhost ~]$ pig -x local /home/acadgild/Task_3_2.pig
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]
18/05/03 19:42:39 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/05/03 19:42:39 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-05-03 19:42:39,626 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-05-03 19:42:39,627 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1525356759623.log
2018-05-03 19:42:39,747 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2018-05-03 19:42:40,682 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-05-03 19:42:41,173 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-05-03 19:42:41,173 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:42:41,181 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2018-05-03 19:42:41,502 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-Task_3_2.pig-1d5383f3-248c-4eb9-a2df-44a55b1bae5f
2018-05-03 19:42:41,503 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-05-03 19:42:41,739 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:42:43,203 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:42:43,541 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
2018-05-03 19:42:43,722 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY,FILTER,LIMIT
2018-05-03 19:42:43,861 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:42:44,009 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculato
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEach
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-05-03 19:42:44,254 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. collectionUsageThres
hold = 489350752, usageThreshold = 489350752
2018-05-03 19:42:44,408 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-05-03 19:42:44,470 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2018-05-03 19:42:44,565 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapR
educe node scope-39
```

Output



```
2018-05-03 19:43:23,067 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:43:23,070 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:43:23,074 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:43:23,110 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:43:23,114 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:43:23,120 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:43:23,135 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:43:23,140 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:43:23,142 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:43:23,159 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:43:23,165 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:43:23,168 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:43:23,183 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-03 19:43:23,199 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:43:23,199 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-03 19:43:23,233 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-03 19:43:23,233 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
2018-05-03 19:43:23,432 [main] INFO  org.apache.pig.Main - Pig script completed in 46 seconds and 703 milliseconds (46703 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ 
```

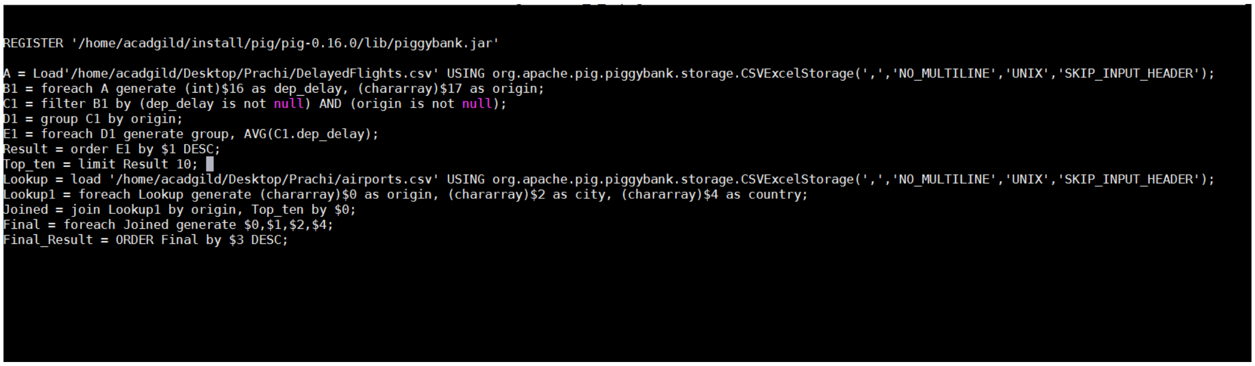
Problem Statement 3

Top ten origins with the highest AVG departure delay

Solution Approach

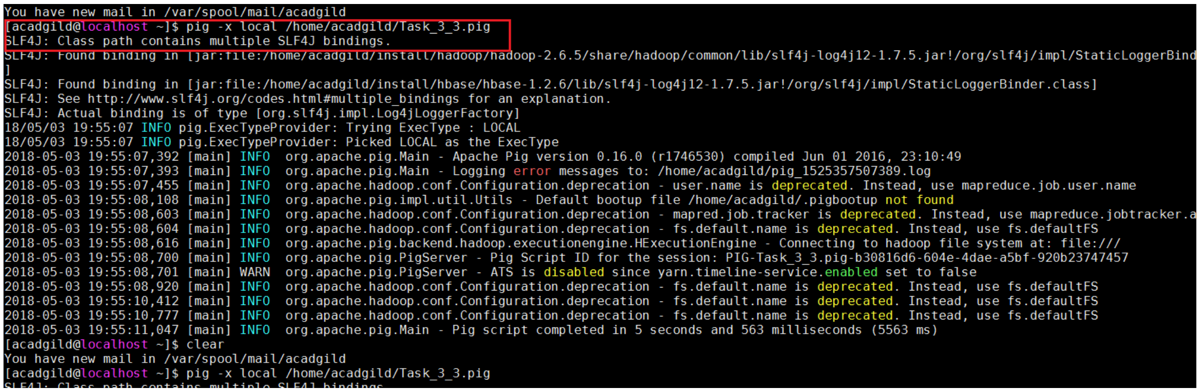
- Load Delayedflights.csv in a relation usinf UDF CSVExcelStorage written in piggybank
- Once data is loaded , generate the required columns as
  - Origin (17)
  - Delayed time (16)
- Filter data where origin is also not null and delayed time should not be null
- Group the filter data on based of Origin Name
- Get the average of all delayed times - This can be achieved using AVG function
- Get the details of origin by joining the airports.csv data

PIG Script



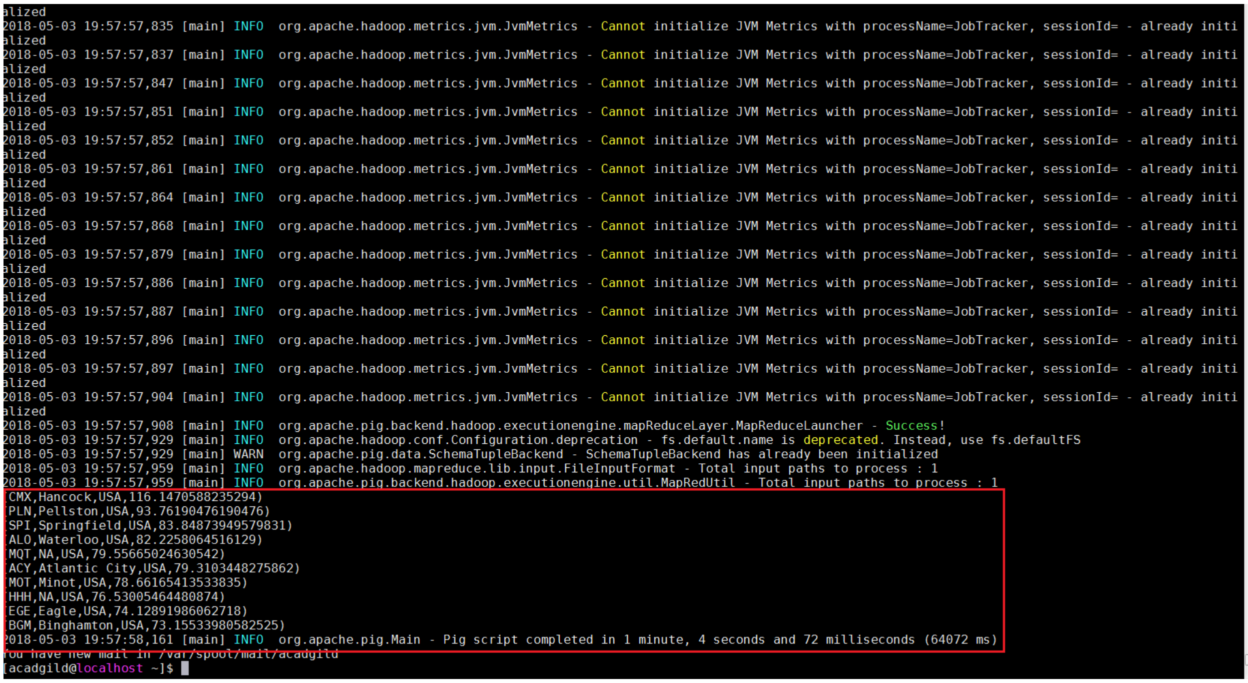
```
REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar'

A = Load'/home/acadgild/Desktop/Prachi/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/Desktop/Prachi/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;
```

Execution



```
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ pig -x local /home/acadgild/Task_3_3.pig
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/StaticLoggerBind
]
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]
18/05/03 19:55:07 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/05/03 19:55:07 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-05-03 19:55:07,392 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-05-03 19:55:07,393 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1525357507389.log
2018-05-03 19:55:07,455 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2018-05-03 19:55:08,108 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-05-03 19:55:08,603 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.a
2018-05-03 19:55:08,604 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:55:08,616 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2018-05-03 19:55:08,700 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-Task_3_3.pig-b30816d6-604e-4dae-a5bf-920b23747457
2018-05-03 19:55:08,701 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-05-03 19:55:08,920 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:55:10,412 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:55:10,777 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:55:11,047 [main] INFO  org.apache.pig.Main - Pig script completed in 5 seconds and 563 milliseconds (5563 ms)
[acadgild@localhost ~]$ clear
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ pig -x local /home/acadgild/Task_3_3.pig
SLF4J: Class path contains multiple SLF4J bindings.
```

Output



```
alized
2018-05-03 19:57:57,835 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,837 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,847 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,851 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,852 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,861 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,864 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,868 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,879 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,886 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,887 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,896 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,897 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,904 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 19:57:57,908 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-03 19:57:57,929 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 19:57:57,929 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-03 19:57:57,959 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-03 19:57:57,959 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(CMX,Hancock,USA,116.1470588235294)
(PLN,Pellston,USA,93.76190476190476)
(SPI,Springfield,USA,83.84873949579831)
(ALO,Waterloo,USA,82.2258064516129)
(MQT,NA,USA,79.55665024630542)
(ACY,Atlantic City,USA,79.3103448275862)
(MOT,Minot,USA,78.66165413533835)
(HHH,NA,USA,76.53005464480874)
(EGE,Eagle,USA,74.12891986062718)
(BGM,Binghamton,USA,73.15533980582525)
2018-05-03 19:57:58,161 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 4 seconds and 72 milliseconds (64072 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

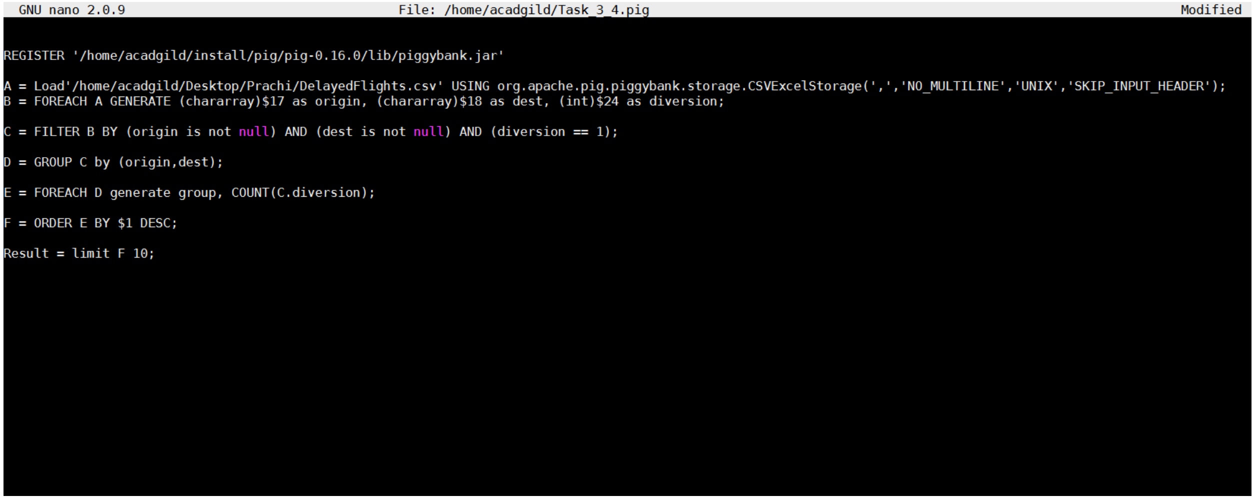
Problem Statement 4

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

Solution Approach

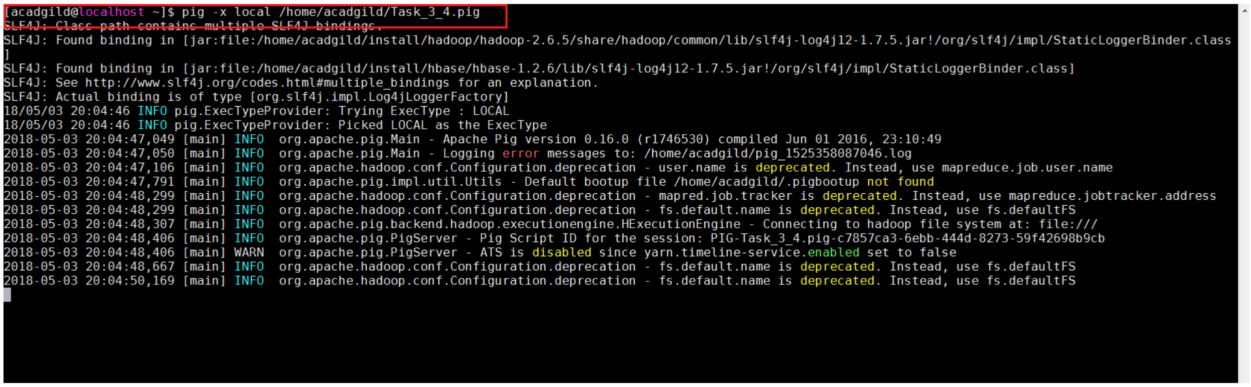
- Load Delayedflights.csv in a relation usinf UDF CSVExcelStorage written in piggybank
- Once data is loaded , generate the required columns as
    - Diverted (24) (yes = 1 , no = 0)
    - Origin (17)
    - Destination (18)
- Filter data where Destination , origin and diverted should not be null
- Then group data on basis of destination and origin and got the count of diversions

PIG Script



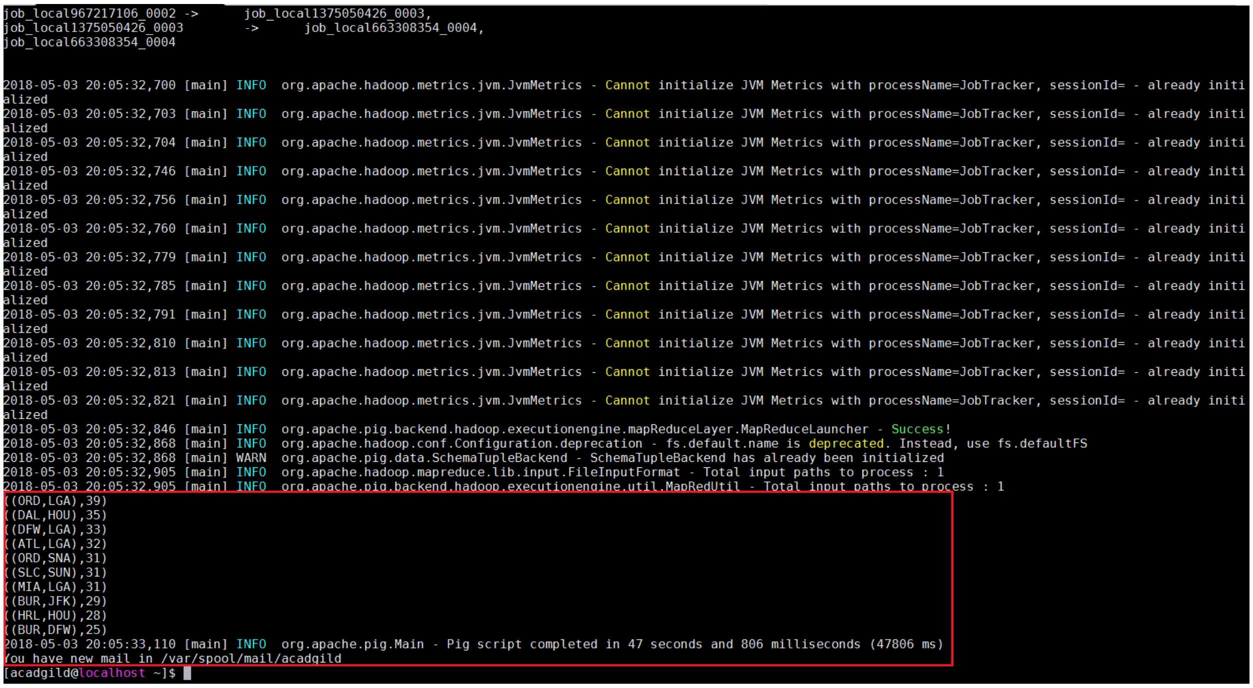
```
  GNU nano 2.0.9                          File: /home/acadgild/Task_3_4.pig                                                 Modified


REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar'

A = Load'/home/acadgild/Desktop/Prachi/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;
```

Execution



```
[acadgild@localhost ~]$ pig -x local /home/acadgild/Task_3_4.pig
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]
18/05/03 20:04:46 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/05/03 20:04:46 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-05-03 20:04:47,049 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-05-03 20:04:47,050 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1525358087046.log
2018-05-03 20:04:47,106 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2018-05-03 20:04:47,791 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-05-03 20:04:48,299 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-05-03 20:04:48,299 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 20:04:48,307 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2018-05-03 20:04:48,406 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-Task_3_4.pig-c7857ca3-6ebb-444d-8273-59f42698b9cb
2018-05-03 20:04:48,406 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-05-03 20:04:48,667 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 20:04:50,169 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```

Output



```
job_local967217106_0002 ->      job_local1375050426_0003,
job_local1375050426_0003        ->      job_local663308354_0004,
job_local663308354_0004


2018-05-03 20:05:32,700 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 20:05:32,703 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 20:05:32,704 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 20:05:32,746 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 20:05:32,756 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 20:05:32,760 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 20:05:32,779 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 20:05:32,785 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 20:05:32,791 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 20:05:32,810 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 20:05:32,813 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 20:05:32,821 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2018-05-03 20:05:32,846 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-03 20:05:32,868 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-03 20:05:32,868 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-03 20:05:32,905 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-03 20:05:32,905 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
((ORD,LGA),39)
((DAL,HOU),35)
((DFW,LGA),33)
((ATL,LGA),32)
((ORD,SNA),31)
((SLC,SUN),31)
((MIA,LGA),31)
((BUR,JFK),29)
((HRL,HOU),28)
((BUR,DFW),25)
2018-05-03 20:05:33,110 [main] INFO  org.apache.pig.Main - Pig script completed in 47 seconds and 806 milliseconds (47806 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ 
```