

BIG DATA HADOOP & SPARK TRAINING

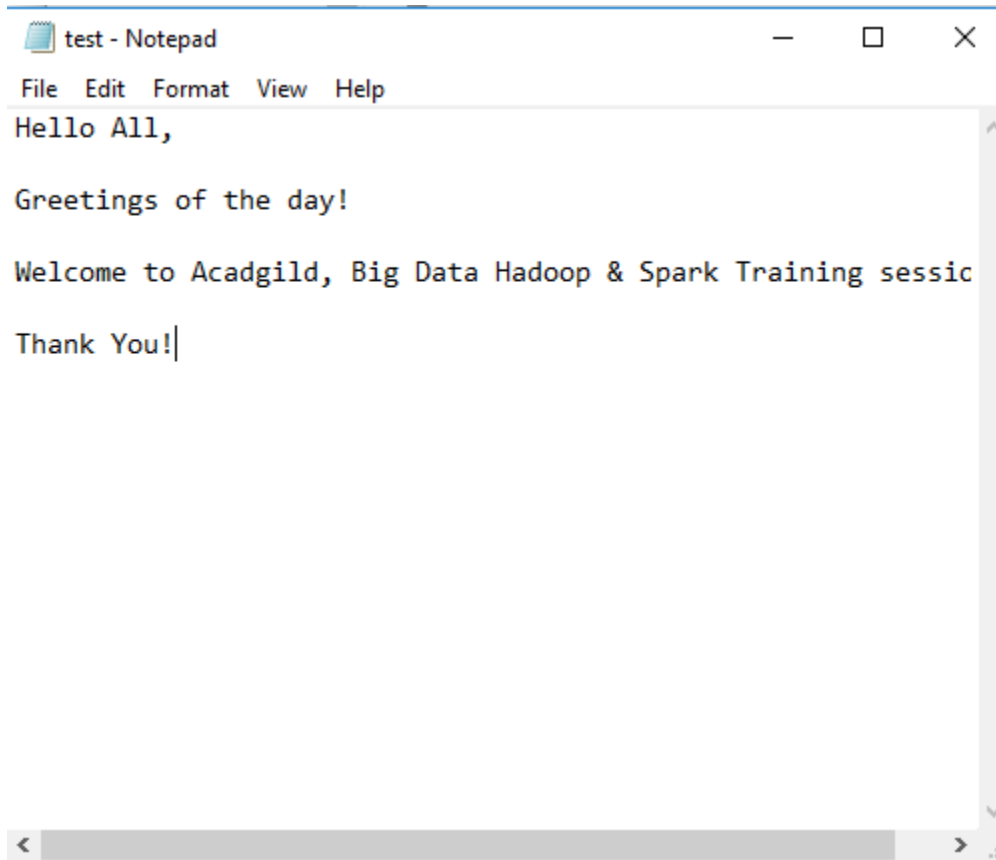
Assignment 7: Exploring Apache Pig. By Rashmi.K

Contents

Task 1: Write a program to implement word count using Pig.....	2
Task 2: We have employee_details and employee_expenses files. Use local mode while running Pig and	5
write Pig Latin script to get below results:.....	5
(a) Top 5 employees (employee id and employee name) with highest rating. (In case two.....	5
employees have same rating, employee with name coming first in dictionary should get	5
preference)	5
(b) Top 3 employees (employee id and employee name) with highest salary, whose employee id 8	8
is an odd number. (In case two employees have same salary, employee with name coming first ..	8
in dictionary should get preference)	8
(c) Employee (employee id and employee name) with maximum expense (In case two	11
employees have same expense, employee with name coming first in dictionary should get	11
preference)	11
(d) List of employees (employee id and employee name) having entries in employee_expenses.15	15
file.	15
(e) List of employees (employee id and employee name) having no entry in employee_expenses	18
file	18
 Task 3: Implement the use case present in below blog link and share the complete steps along with	
screenshot(s) from your end.....	22
Problem Statement 1: Find out the top 5 most visited destinations.	22
Problem Statement 2: Which month has seen the most number of cancellations due to bad	
weather?.....	26
Problem Statement 3:Top ten origins with the highest AVG departure delay	29
Problem Statement 4: Which route (origin & destination) has seen the maximum diversion?	33

Task 1: Write a program to implement word count using Pig.

Below screen shot represents the input file for the word count program “test.txt”:



Below is the pig script run in local mode to find the word count in the above file:

A = load '/hadoopdata/pig/test.txt';

This command loads the file “test.txt” from HDFS to the relation “A”.

B= foreach A generate flatten(TOKENIZE((chararray)\$0)) AS word;

This command takes each record from “A” and flattens(remove the elements from the bag) every word (*as word*) from the input file and save in the relation “B”.

C= group B by word;

This command groups each word from “B” and save it in the relation “C”.

D= foreach C generate group, COUNT(B);

This command calculates count of every word from the relation “B” and groups them to store in the relation “D”.

dump D;

This command executes the map reduce program to count the words in the input file.

```
grunt>
grunt> A = load '/hadoopdata/pig/test.txt';
2018-03-27 16:42:30,138 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B= foreach A generate flatten(TOKENIZE((chararray)$0)) AS word;
grunt> C= group B by word;
grunt> D= foreach C generate group, COUNT(B);
grunt> dump D;
2018-03-27 16:43:06,850 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2018-03-27 16:43:06,910 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-27 16:43:06,915 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
2018-03-27 16:43:06,918 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-27 16:43:06,936 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-03-27 16:43:06,940 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2018-03-27 16:43:06,958 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2018-03-27 16:43:06,959 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2018-03-27 16:43:06,989 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-27 16:43:06,991 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:43:06,999 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2018-03-27 16:43:07,004 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-03-27 16:43:07,005 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2018-03-27 16:43:07,005 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2018-03-27 16:43:07,008 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=10000000000 maxReducers=999 totalInputFileSize=116
2018-03-27 16:43:07,008 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelis
```

```

job_1522148259963_0005]
2018-03-27 16:43:48,591 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:43:48,603 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 16:43:48,811 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:43:48,828 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 16:43:48,933 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:43:48,948 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 16:43:49,098 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-27 16:43:49,098 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.5 0.16.0 acadgild 2018-03-27 16:43:06 2018-03-27 16:43:49 GROUP_BY

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime M
medianReduceTime Alias Feature Outputs
job_1522148259963_0005 1 1 7 7 7 7 8 8 8 8 A,B,C,D GROUP_BY,COMBINER h
dfs://localhost:8020/tmp/temp-893193075/tmp1566002081,

Input(s):
Successfully read 7 records (488 bytes) from: "/hadoopdata/pig/test.txt"

Output(s):
Successfully stored 19 records (238 bytes) in: "hdfs://localhost:8020/tmp/temp-893193075/tmp1566002081"

Counters:
Total records written : 19
Total bytes written : 238
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:

```

```

2018-03-27 16:43:49,105 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:43:49,115 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 16:43:49,305 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:43:49,340 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 16:43:49,446 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:43:49,461 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 16:43:49,578 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-03-27 16:43:49,582 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-27 16:43:49,583 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2018-03-27 16:43:49,601 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-03-27 16:43:49,601 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(G,1)
(of,1)
(to,1)
(All,1)
(Big,1)
(the,1)
(Data,1)
(You!,1)
(day!,1)
(Hello,1)
(Spark,1)
(Thank,1)
(Hadoop,1)
(Welcome,1)
(Acadgild,1)
(Training,1)
(Greetings,1)
(sessions.,1)
(,0)
grunt>

```

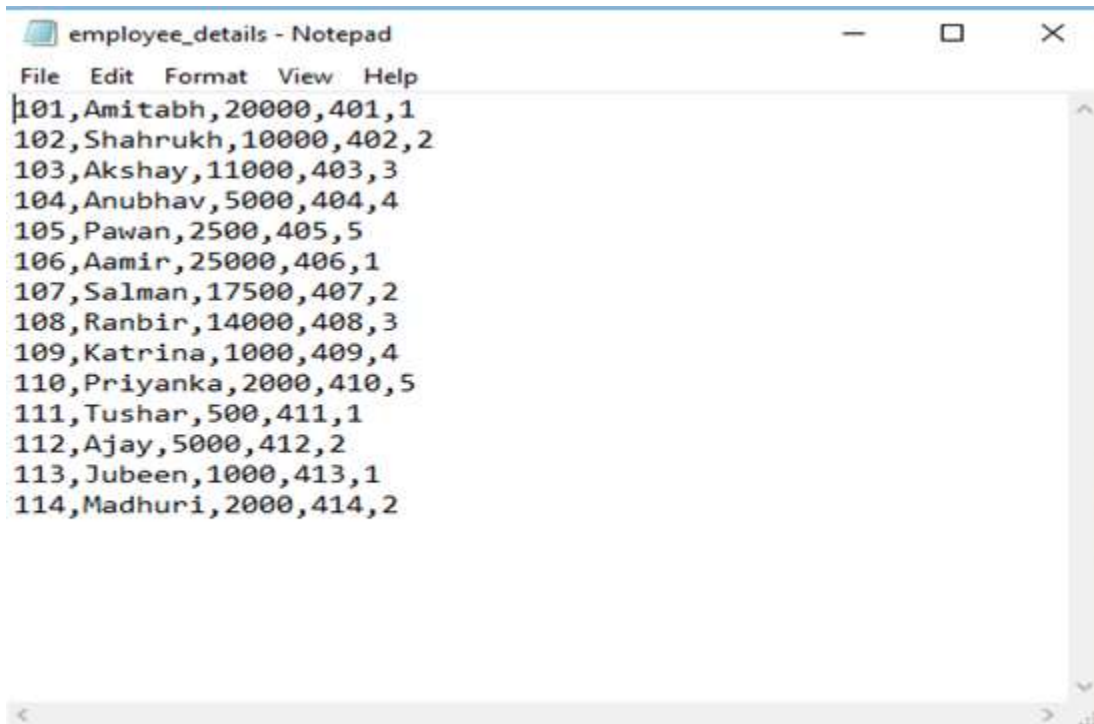
← output: word count in the file test.txt

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:

(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)

Below screen shot represents the input file “employeedetails.txt”:



Below is the pig script run in local mode to find the word count in the above file:

```
empl = LOAD '/hadoopdata/pig/employee_details1.txt' using PigStorage(',') as  
(eid:int,name:chararray,salary:int,dno:int,rating:int);
```

This command loads the file “employeedetails.txt” from HDFS to the relation “empl”.

```
grpname = GROUP empl by name;
```

This command groups all records of empl by name condition and saves it in the relation “grpname”.

```
top5 = FOREACH grpname {  
    sorted = ORDER empl by name ASC, rating DESC;  
    filtered = FILTER empl by rating >=4;  
    generate flatten(filtered);
```

```
};
```

This command takes each record of relation `grpname`

- sorts
 - Name is ascending order.
 - Rating in descending order.

So that if two employees have same rating, employee with name coming first in dictionary should get Preference.

- Filters the records if rating is equal to or greater than 4.
- Flattens, i.e. remove all records from filtered relation.

top = limit top5 5;

This command limits the output only to **top 5 records**, saves it in the relation “top”

dump top;

This command executes the mapreduce program to find **top 5 employees** with highest ratings from the input file.

```
grunt> empl = LOAD '/hadoopdata/pig/employee_details1.txt' using PigStorage(',') as (eid:int,name:chararray,salary:int,dno:int,rating:int);
2018-03-27 16:36:21,819 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> grpname = GROUP empl by name;
grunt> top5 = FOREACH grpname {
>> sorted = ORDER empl by name ASC, rating DESC;
>> filtered = FILTER empl by rating >=4;
>> generate flatten(filtered);
>> };
grunt> dump top5;
2018-03-27 16:36:49,149 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP BY
2018-03-27 16:36:49,257 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-27 16:36:49,271 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
2018-03-27 16:36:49,352 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-27 16:36:49,487 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 489350752
2018-03-27 16:36:49,611 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-03-27 16:36:49,683 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2018-03-27 16:36:49,684 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2018-03-27 16:36:49,757 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-27 16:36:49,852 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:36:50,392 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2018-03-27 16:36:50,415 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2018-03-27 16:36:50,415 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-03-27 16:36:50,415 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, us
```

```

tus=SUCCEEDED. Redirecting to job history server
2018-03-27 16:37:30,888 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:37:30,831 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 16:37:31,050 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:37:31,066 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 16:37:31,274 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-27 16:37:31,280 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion      UserId StartedAt      FinishedAt      Features
2.6.5 0.16.0 acadgild 2018-03-27 16:36:50 2018-03-27 16:37:31 GROUP_BY

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime M
edianReduceTime Alias Feature Outputs
job_1522148259963_0004 1 1 7 7 7 7 8 8 8 8 empl,filtered,grpname,top5 G
ROUP_BY hdfs://localhost:8020/tmp/temp-893193075/tmp-1833278322,

Input(s):
Successfully read 14 records (714 bytes) from: "/hadoopdata/pig/employee_details1.txt"

Output(s):
Successfully stored 4 records (95 bytes) in: "hdfs://localhost:8020/tmp/temp-893193075/tmp-1833278322"

Counters:
Total records written : 4
Total bytes written : 95
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1522148259963_0004

```

```

Input(s):
Successfully read 14 records (714 bytes) from: "/hadoopdata/pig/employee_details1.txt"

Output(s):
Successfully stored 4 records (95 bytes) in: "hdfs://localhost:8020/tmp/temp-893193075/tmp-1833278322"

Counters:
Total records written : 4
Total bytes written : 95
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

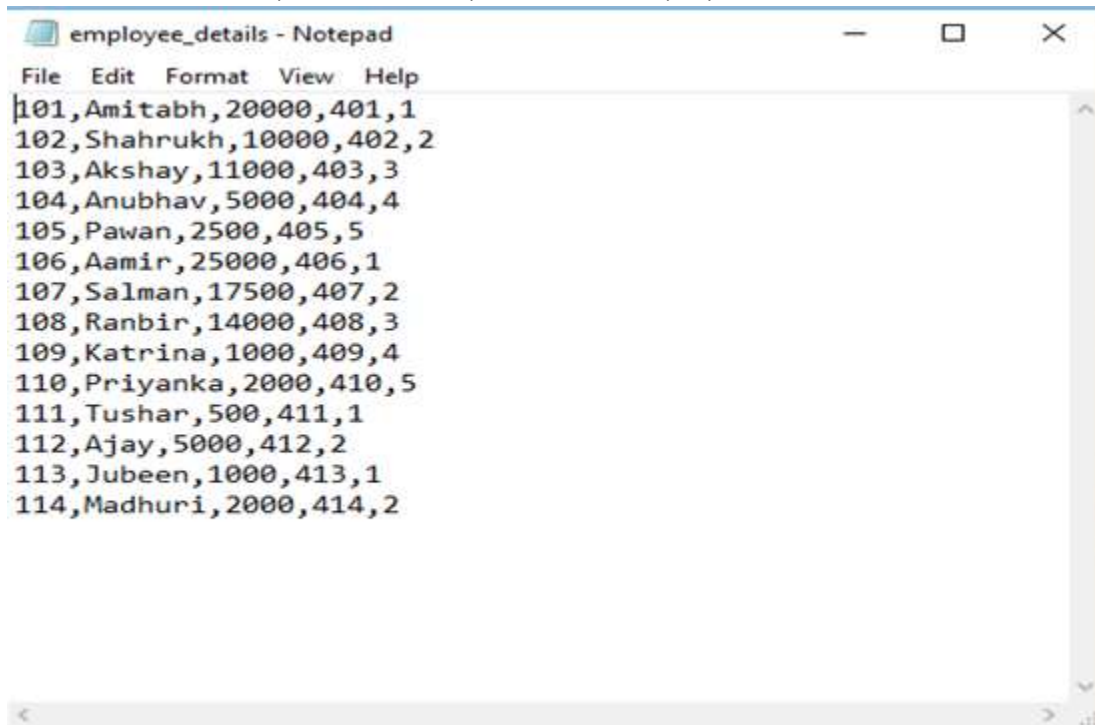
Job DAG:
job_1522148259963_0004

2018-03-27 16:37:31,286 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:37:31,302 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 16:37:31,413 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:37:31,433 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 16:37:31,545 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 16:37:31,557 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 16:37:31,675 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-03-27 16:37:31,687 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-27 16:37:31,690 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2018-03-27 16:37:31,740 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-03-27 16:37:31,740 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(105,Pawan,2500,405,5)
(104,Anubhav,5000,404,4)
(109,Katrina,1000,409,4)
(110,Priyanka,2000,410,5)
grunt> █

```


(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)

Below screen shot represents the input file for “employee_details.txt”:



```
File Edit Format View Help
101,Amitabh,20000,401,1
102,Shahrukh,10000,402,2
103,Akshay,11000,403,3
104,Anubhav,5000,404,4
105,Pawan,2500,405,5
106,Aamir,25000,406,1
107,Salman,17500,407,2
108,Ranbir,14000,408,3
109,Katrina,1000,409,4
110,Priyanka,2000,410,5
111,Tushar,500,411,1
112,Ajay,5000,412,2
113,Jubeen,1000,413,1
114,Madhuri,2000,414,2
```

Below is the pig script run in local mode:

```
empl = LOAD '/hadoopdata/pig/employee_details1.txt' using PigStorage(',') as  
(eid:int,name:chararray,salary:int,dno:int,rate:int);
```

This command loads the file “employee_details.txt” from HDFS to the relation “empl”.

```
A = FILTER empl by (eid % 2!= 0);
```

This command will filter all the records holding odd number of employee id and saves it in the relation “A”.

```
sorted = ORDER A by salary DESC,name ASC;
```

This command sorts each record in the relation “B” as following:

- Name is ascending order.
- Rating in descending order

```
top = limit sorted 3;
```

This command limits the output only to top **3 records**, saves it in the relation “top”

dump top;

This command executes the mapreduce program to find top **3 employees** with highest salary and odd number of employee id from the input file.

```
grunt> empl = LOAD '/hadoopdata/pig/employee_details1.txt' using PigStorage(',') as (eid:int,name:chararray,salary:int,dno:int,rating:int);
2018-03-27 19:19:35,958 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> A = FILTER empl by (eid % 2 != 0);
grunt> sorted = ORDER A by salary DESC,name ASC;
grunt> top = limit sorted 3;
grunt> dump top;
2018-03-27 19:20:16,717 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,FILTER,LIMIT
2018-03-27 19:20:16,755 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-27 19:20:16,757 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
2018-03-27 19:20:16,757 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-27 19:20:16,760 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-03-27 19:20:16,766 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-1231
2018-03-27 19:20:16,767 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4
2018-03-27 19:20:16,767 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 4
2018-03-27 19:20:16,778 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-27 19:20:16,780 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 19:20:16,783 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2018-03-27 19:20:16,783 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-03-27 19:20:16,783 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2018-03-27 19:20:16,853 [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 DistributedCache through /tmp/temp-574938762/tmp2074656725/pig-0.16.0-core-h2.jar
2018-03-27 19:20:16,867 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/ho
```

```

2018-03-27 19:22:44,995 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 19:22:45,057 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 19:22:45,067 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 19:22:45,124 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-27 19:22:45,133 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

```

```

HadoopVersion PigVersion      UserId StartedAt      FinishedAt      Features
2.6.5      0.16.0      acadgild      2018-03-27 19:20:16      2018-03-27 19:22:45      ORDER_BY,FILTER,LIMIT

```

Success!

Job Stats (time in seconds):

JobId	Maps	Reduces	MaxMapTime	MinMapTime	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime	M
job_1522148259963_0063	1	0	7	7	7	0	0	0	0	A,empl MAP_ONLY
job_1522148259963_0064	1	1	7	7	7	9	9	9	9	sorted SAMPLER
job_1522148259963_0065	1	1	7	7	7	8	8	8	8	sorted ORDER_BY,COMBINER
job_1522148259963_0066	1	1	7	7	7	8	8	8	8	sorted hdfs://localhost:8020/tmp/temp-574938762/tmp-1254029359,

Input(s):

Successfully read 14 records (714 bytes) from: "/hadoopdata/pig/employee_details1.txt"

Output(s):

Successfully stored 3 records (69 bytes) in: "hdfs://localhost:8020/tmp/temp-574938762/tmp-1254029359"

Counters:

Total records written : 3
Total bytes written : 69
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:

```

job_1522148259963_0063 -> job_1522148259963_0064,
job_1522148259963_0064 -> job_1522148259963_0065,

```

```

2018-03-27 19:22:45,242 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 19:22:45,248 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 19:22:45,305 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 19:22:45,310 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 19:22:45,341 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 19:22:45,344 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 19:22:45,381 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 19:22:45,384 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 19:22:45,418 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 19:22:45,424 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 19:22:45,466 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 19:22:45,473 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 19:22:45,547 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 19:22:45,555 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 19:22:45,618 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 19:22:45,622 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 19:22:45,651 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-27 19:22:45,654 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-27 19:22:45,688 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-03-27 19:22:45,689 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-27 19:22:45,690 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
2018-03-27 19:22:45,694 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-03-27 19:22:45,694 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

```

```

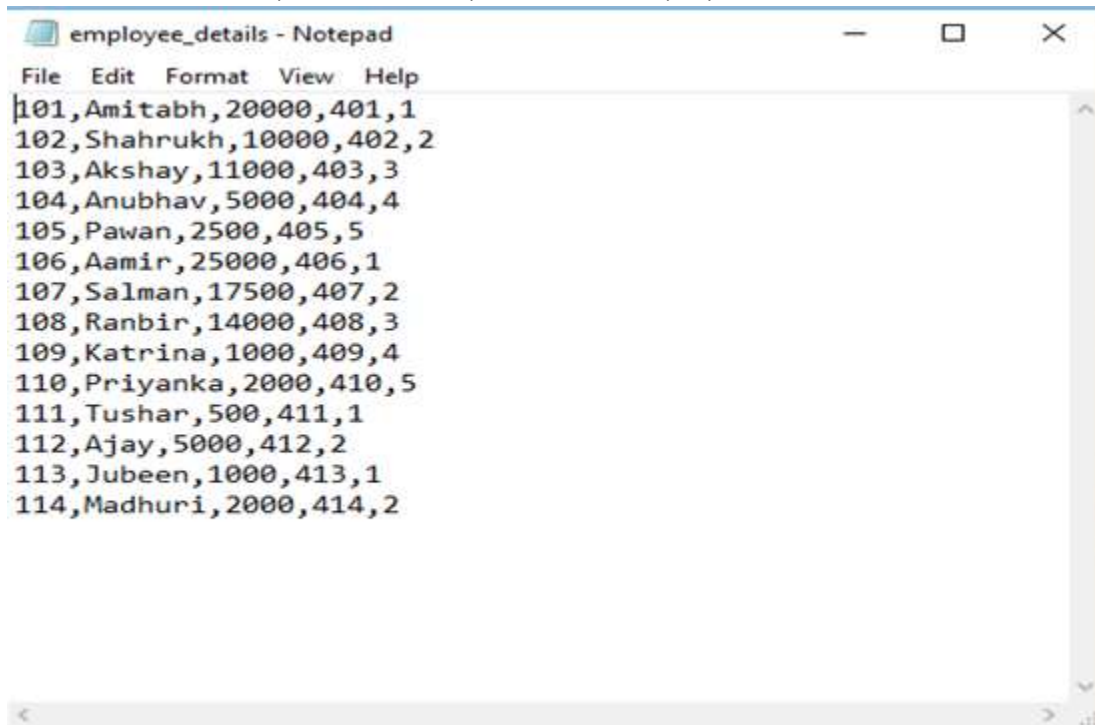
(101,Amitabh,20000,401,1)
(107,Salman,17500,407,2)
(103,Akshay,11000,403,3)
grunt>

```

Maximum salary for odd
number of employees
salary

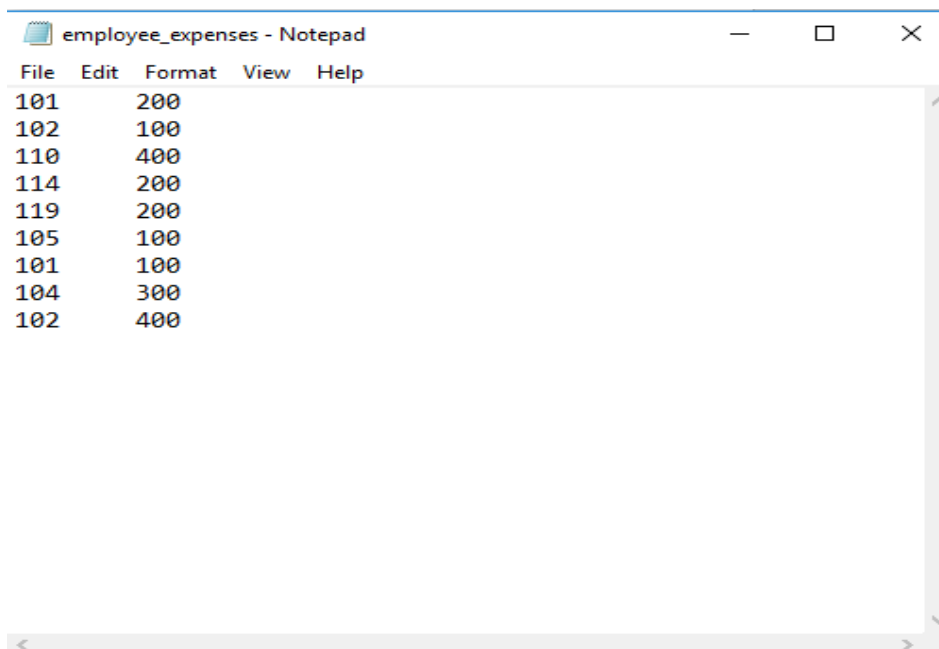
(c) Employee (employee id and employee name) with maximum expense (In case two employees have same expense, employee with name coming first in dictionary should get preference)

Below screen shot represents the input file for “employee_details.txt”:



```
File Edit Format View Help
101,Amitabh,20000,401,1
102,Shahrukh,10000,402,2
103,Akshay,11000,403,3
104,Anubhav,5000,404,4
105,Pawan,2500,405,5
106,Aamir,25000,406,1
107,Salman,17500,407,2
108,Ranbir,14000,408,3
109,Katrina,1000,409,4
110,Priyanka,2000,410,5
111,Tushar,500,411,1
112,Ajay,5000,412,2
113,Jubeen,1000,413,1
114,Madhuri,2000,414,2
```

Below screen shot represents the input file for “employeeexpense.txt”:



```
File Edit Format View Help
101 200
102 100
110 400
114 200
119 200
105 100
101 100
104 300
102 400
```


Below is the pig script run in local mode:

```
empl = LOAD '/hadoopdata/pig/employee_details1.txt' using PigStorage(',') as  
(eid:int,name:chararray,salary:int,dno:int,rating:int);  
  
emplexp = LOAD '/hadoopdata/pig/employee_expenses1.txt' using PigStorage('\t') as  
(eid:int,expen:int);
```

The above commands will load **employee_details1.txt** & **employee_expenses1.txt** to empl and emplexp relations respectively.

```
A = join empl by (eid), emplexp by (eid);
```

This command will join two tables employee_details & employee_expenses using employee id as joining factor and save it in the relation “A”.

```
sorted = ORDER A by expence DESC,name ASC;
```

This command will sort all the elements of relation “B” as expense in descending order and employee name is ascending order, so that if two employees have same expense, employee with name coming first in dictionary should get preference.

```
B = GROUP sorted by name;
```

This command will group all sorted records by name and save it in the relation “B”.

```
C = foreach B generate sorted.name, SUM(sorted.expence);
```

This command will group each element in relation “B” as name from employee_details table and expense from employee_expense table and find the sum of expenses for all employees and save it in the relation “C”.

```
dump C;
```

This command executes the mapreduce program to find maximum expense from the input file.

```

grunt> empl = LOAD '/hadoopdata/pig/employee_details1.txt' using PigStorage(',') as (eid:int,name:chararray,salary:int,dno:int,rating:int);
2018-03-28 18:57:43,702 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
grunt> emlexp = LOAD '/hadoopdata/pig/employee_expenses1.txt' using PigStorage(',') as (eid:int,expenditure:int);
2018-03-28 18:57:47,725 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
grunt> A = join empl by (eid), emlexp by (eid);
grunt> sorted = ORDER A by expenditure DESC,name ASC;
grunt> B = GROUP sorted by name;
grunt> C = foreach B generate sorted.name, SUM(sorted.expenditure);
grunt> dump C;
2018-03-28 18:59:08,742 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN, GROUP_BY, ORDE
R_BY
2018-03-28 18:59:08,828 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-28 18:59:08,829 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2018-03-28 18:59:08,833 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMa
pKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilt
erOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-28 18:59:08,855 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thresho
ld: 100 optimistic? false
2018-03-28 18:59:08,886 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Seconda
ry Key Optimization for MapReduce node scope-164
2018-03-28 18:59:08,888 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer
- Rewrite: POForEach->POForEach to POForEach(JoinPacker)
2018-03-28 18:59:08,892 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size befo
re optimization: 4
2018-03-28 18:59:08,892 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size afte
r optimization: 4
2018-03-28 18:59:09,043 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS

```

```

tus=SUCCESSFUL. Redirecting to job history server
2018-03-28 19:06:30,655 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-28 19:06:30,676 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

```

HadoopVersion	PigVersion	UserId	StartedAt	FinishedAt	Features
2.6.5	0.16.0	acadgild	2018-03-28 19:03:41	2018-03-28 19:06:30	HASH_JOIN, GROUP_BY, ORDER_BY

Success!

Job Stats (time in seconds):

JobId	Maps	Reducers	MaxMapTime	MinMapTime	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime	M
job_1522227177035_0022	2	1	17	17	17	8	8	8	8	A,empl,emlexp HASH_JOIN
job_1522227177035_0023	1	1	7	7	7	8	8	8	8	sorted SAMPLER
job_1522227177035_0024	1	1	7	7	7	7	7	7	7	sorted ORDER_BY
job_1522227177035_0025	1	1	6	6	6	8	8	8	8	B,C GROUP_BY hdfs://lo

calhost:8020/tmp/Temp-1964416219/tmp884242353,

Input(s):

Successfully read 14 records from: "/hadoopdata/pig/employee_details1.txt"
 Successfully read 9 records from: "/hadoopdata/pig/employee_expenses1.txt"

Output(s):

Successfully stored 6 records (142 bytes) in: "hdfs://localhost:8020/tmp/temp-1964416219/tmp884242353"

Counters:

Total records written : 6
 Total bytes written : 142
 Spillable Memory Manager spill count : 0
 Total bags proactively spilled: 0
 Total records proactively spilled: 0

Job DAG:

```

job_1522227177035_0022 -> job_1522227177035_0023,
job_1522227177035_0023 -> job_1522227177035_0024,
job_1522227177035_0024 -> job_1522227177035_0025,
job_1522227177035_0025

```

```

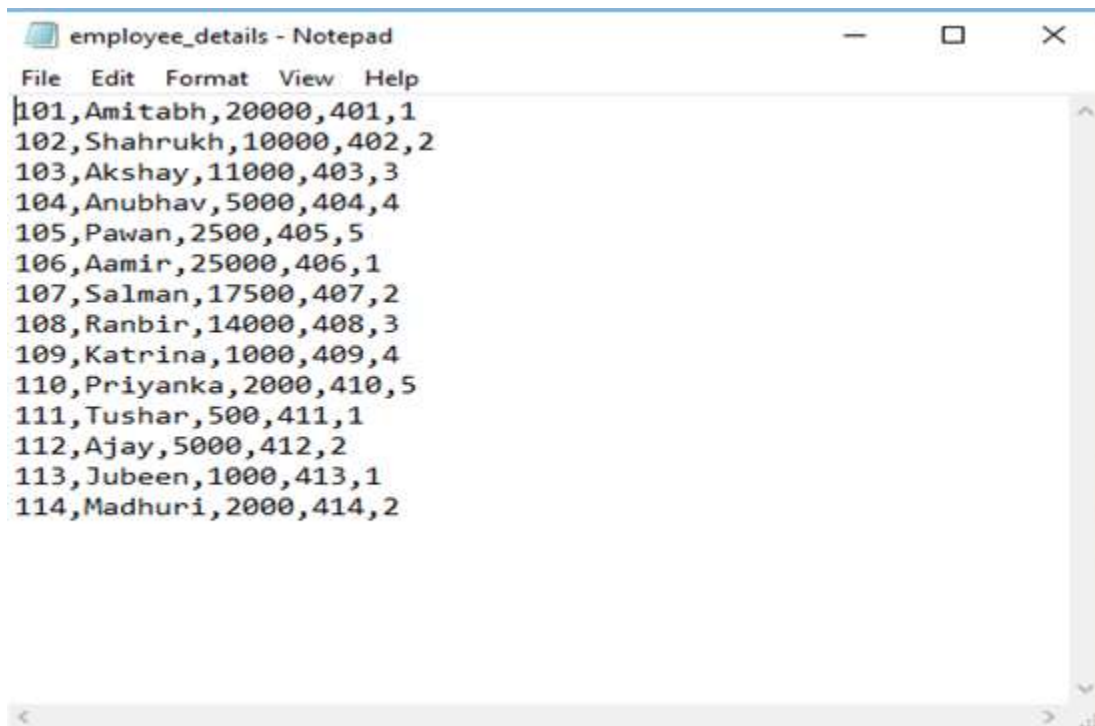
2018-03-28 19:06:31,034 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:06:31,042 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:06:31,110 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:06:31,115 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:06:31,178 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:06:31,193 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:06:31,264 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:06:31,269 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:06:31,318 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:06:31,323 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:06:31,376 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:06:31,393 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:06:31,459 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:06:31,464 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:06:31,509 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:06:31,515 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:06:31,555 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-03-28 19:06:31,555 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-28 19:06:31,556 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2018-03-28 19:06:31,565 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-03-28 19:06:31,565 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
{(Pawan)},100)
{(Amitabh),(Amitabh)},300)
{(Anubhav)},300)
{(Madhuri)},200)
{(Priyanka)},400)
{(Shahrukh),(Shahrukh)},500)
grunt>

```

Employees with
maximum expense

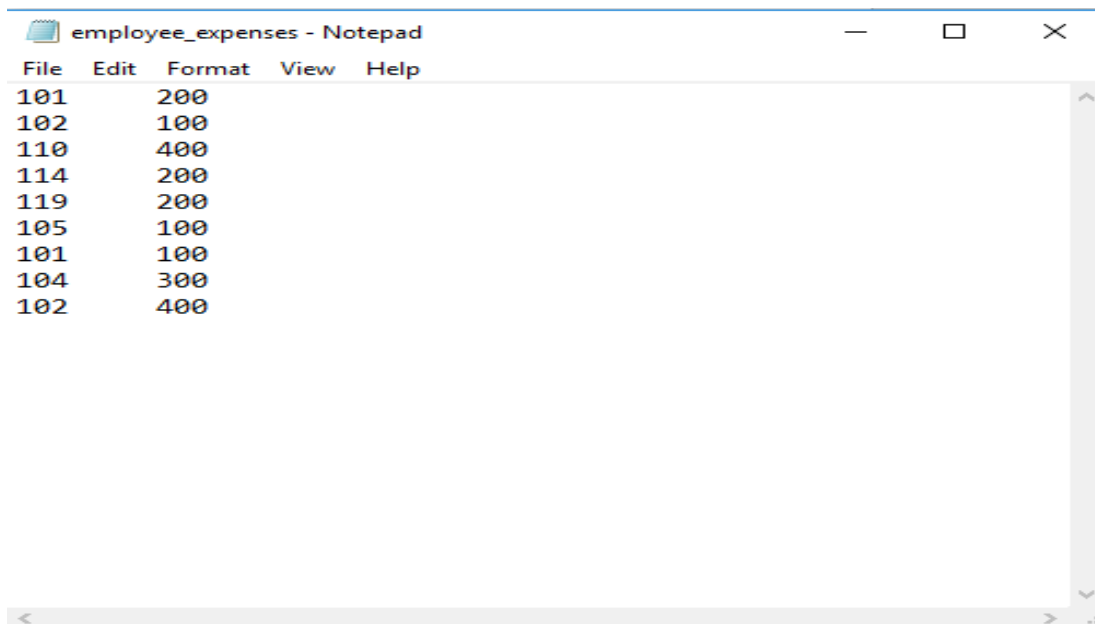
(d) List of employees (employee id and employee name) having entries in employee_expenses file.

Below screen shot represents the input file for “employee_details.txt”:



```
File Edit Format View Help
101,Amitabh,20000,401,1
102,Shahrukh,10000,402,2
103,Akshay,11000,403,3
104,Anubhav,5000,404,4
105,Pawan,2500,405,5
106,Aamir,25000,406,1
107,Salman,17500,407,2
108,Ranbir,14000,408,3
109,Katrina,1000,409,4
110,Priyanka,2000,410,5
111,Tushar,500,411,1
112,Ajay,5000,412,2
113,Jubeen,1000,413,1
114,Madhuri,2000,414,2
```

Below screen shot represents the input file for “employeeexpense.txt”:



```
File Edit Format View Help
101      200
102      100
110      400
114      200
119      200
105      100
101      100
104      300
102      400
```

Below is the pig script run in local mode:


```
empl = LOAD '/hadoopdata/pig/employee_details1.txt' using PigStorage(',') as  
(eid:int,name:chararray,salary:int,dno:int,rating:int);
```

```
emplexp = LOAD '/hadoopdata/pig/employee_expenses1.txt' using PigStorage('\t') as  
(eid:int,expenditure:int);
```

The above commands will load **employee_details1.txt** & **employee_expenses1.txt** to **empl** and **emplexp** relations respectively.

```
A = join empl by (eid) LEFT OUTER, emplexp by (eid);
```

This command will join using left outer join(The **left outer Join** operation returns all rows from the left table, even if there are no matches in the right relation) two tables **employee_details** & **employee_expenses** using employee id as joining factor and save it in the relation “A”.

```
B = FILTER A by expenditure is not null;
```

This command will filter each element of relations “B” where values of expenditure is not null, which indicates the list of all employees who have entries in the expense file.

```
C = GROUP B by name;
```

This command will group all elements of relation “B” by employee name.

```
D = foreach C generate B.name, SUM(B.expenditure);
```

This command will group each element in relation “B” as name from **employee_details** table and expenditure from **employeeexpense** table and find the sum of expenses for all employees and save it in the relation “C”.

```
dump C;
```

This command executes the mapreduce program to find entries of all employees who have entries in **employeeexpense** file from the input file.

```

grunt>
grunt> empl = LOAD '/hadoopdata/pig/employee_details1.txt' using PigStorage(',') as (eid:int,name:chararray,salary:int,dno:int,rating:int);
2018-03-28 19:28:31,793 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> emplexp = LOAD '/hadoopdata/pig/employee_expenses1.txt' using PigStorage(',') as (eid:int,expenditure:int);
2018-03-28 19:28:39,372 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> A = join empl by (eid) LEFT OUTER, emplexp by (eid);
grunt> B = FILTER A by expenditure is not null;
grunt> C = GROUP B by name;
grunt> D = foreach C generate B.name, SUM(B.expenditure);
grunt> dump D;
2018-03-28 19:29:43,027 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN, GROUP_BY, FILTER
2018-03-28 19:29:43,060 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-28 19:29:43,062 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
2018-03-28 19:29:43,063 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMajorKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-28 19:29:43,068 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-03-28 19:29:43,077 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 2
2018-03-28 19:29:43,077 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 2
2018-03-28 19:29:43,089 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-28 19:29:43,092 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:29:43,095 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2018-03-28 19:29:43,096 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce

```

```

2018-03-28 19:31:05,090 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

```

HadoopVersion	PigVersion	UserId	StartedAt	FinishedAt	Features
2.6.5	0.16.0	acadgild	2018-03-28 19:29:43	2018-03-28 19:31:05	HASH_JOIN, GROUP_BY, FILTER

Success!

Job Stats (time in seconds):

JobId	Maps	Reduces	MaxMapTime	MinMapTime	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime	Features
job_1522227177035_0034	2	1	16	15	16	7	7	7	7	A,B,empl,emplexp
job_1522227177035_0035	1	1	6	6	6	8	8	8	8	C,D GROUP_BY

Input(s):

Successfully read 14 records from: "/hadoopdata/pig/employee_details1.txt"
 Successfully read 9 records from: "/hadoopdata/pig/employee_expenses1.txt"

Output(s):

Successfully stored 6 records (142 bytes) in: "hdfs://localhost:8020/tmp/temp-1964416219/tmp1124132297"

Counters:

Total records written : 6
 Total bytes written : 142
 Spillable Memory Manager spill count : 0
 Total bags proactively spilled: 0
 Total records proactively spilled: 0

Job DAG:

```

job_1522227177035_0034 -> job_1522227177035_0035,
job_1522227177035_0035

```

```

2018-03-28 19:31:05,092 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:31:05,096 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:31:05,136 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:31:05,141 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

```

```

Job DAG:
job_1522227177035_0034 -> job_1522227177035_0035,
job_1522227177035_0035

2018-03-28 19:31:05,092 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:31:05,096 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:31:05,136 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:31:05,141 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:31:05,184 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:31:05,191 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:31:05,246 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:31:05,250 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:31:05,280 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:31:05,284 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:31:05,327 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 19:31:05,332 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 19:31:05,375 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-03-28 19:31:05,375 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-28 19:31:05,381 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set..., will not generate code
.
2018-03-28 19:31:05,389 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-03-28 19:31:05,389 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
({(Pawan)},100)
({(Amitabh)},(Amitabh)},300)
({(Anubhav)},300)
({(Madhuri)},200)
({(Priyanka)},400)
({(Shahrukh)},(Shahrukh)},500)
grt>

```

← List of employees having entries in employee expense file

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

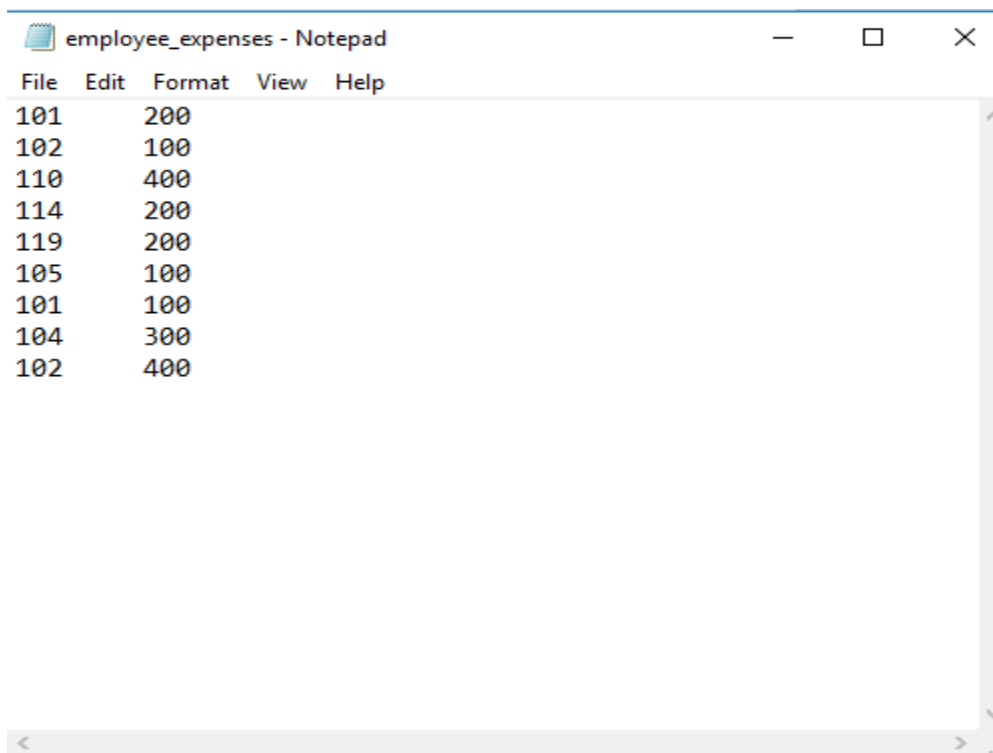
Below screen shot represents the input file for “employee_details.txt”:

```

File Edit Format View Help
101,Amitabh,20000,401,1
102,Shahrukh,10000,402,2
103,Akshay,11000,403,3
104,Anubhav,5000,404,4
105,Pawan,2500,405,5
106,Aamir,25000,406,1
107,Salman,17500,407,2
108,Ranbir,14000,408,3
109,Katrina,1000,409,4
110,Priyanka,2000,410,5
111,Tushar,500,411,1
112,Ajay,5000,412,2
113,Jubeen,1000,413,1
114,Madhuri,2000,414,2

```

Below screen shot represents the input file for “employeeexpense.txt” :



Below is the pig script run in local mode:

```
empl = LOAD '/hadoopdata/pig/employee_details1.txt' using PigStorage(',') as  
(eid:int,name:chararray,salary:int,dno:int,rating:int);  
  
emplexp = LOAD '/hadoopdata/pig/employee_expenses1.txt' using PigStorage('\t') as  
(eid:int,expen:int);
```

The above commands will load **employee_details1.txt** & **employee_expenses1.txt** to empl and emplexp relations respectively.

A = join empl by (eid) LEFT OUTER, emplexp by (eid);

This command will join using left outer join (The **left outer Join** operation returns all rows from the left table, even if there are no matches in the right relation) two tables employee_details & employee_expenses using employee id as joining factor and save it in the relation “A”.

B = foreach A generate empl::eid,empl::name,emplexp::expen;

This command will group each element in relation “A” as employee id, name from employee_details file and expence from employeeexpense file and save it in the relation “B”.

C = FILTER B by expense is null;

This command will filter each element of relations “B” where values of expense is null, which indicates the list of all employees who not have entries in the expense file.

dump C;

This command executes the mapreduce program to find entries of all employees who do not have entries in employeeexpense file from the input file.

```
grunt>
grunt> empl = LOAD '/hadoopdata/pig/employee_details1.txt' using PigStorage(',') as (eid:int,name:chararray,salary:int,dno:int,rating:int);
2018-03-28 15:07:31,958 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> emplexp = LOAD '/hadoopdata/pig/employee_expenses1.txt' using PigStorage(',') as (eid:int,expense:int);
2018-03-28 15:07:44,509 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> A = join empl by (eid) LEFT OUTER, emplexp by (eid);
grunt> B = foreach A generate empl::eid,empl::name,empl::expense;
grunt> C = FILTER B by expense is null;
grunt> dump C;
2018-03-28 15:08:16,121 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH JOIN,FILTER
2018-03-28 15:08:16,182 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-28 15:08:16,183 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-03-28 15:08:16,188 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-28 15:08:16,193 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for empl: $2, $3, $4
2018-03-28 15:08:16,202 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-03-28 15:08:16,210 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2018-03-28 15:08:16,210 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2018-03-28 15:08:16,252 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-28 15:08:16,258 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 15:08:16,266 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2018-03-28 15:08:16,267 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-03-28 15:08:16,267 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2018-03-28 15:08:16,267 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2018-03-28 15:08:16,273 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=408
```

```

2018-03-28 15:09:02,545 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 15:09:02,673 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-28 15:09:02,674 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

```

```

HadoopVersion PigVersion      UserId StartedAt      FinishedAt      Features
2.6.5 0.16.0 acadgild 2018-03-28 15:08:16 2018-03-28 15:09:02 HASH_JOIN,FILTER

```

Success!

Job Stats (time in seconds):

JobId	Maps	Reduces	MaxMapTime	MinMapTime	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime	MedianReduceTime	Alias	Feature	Outputs
job_1522227177035_0006	2	1	16	15	16	16	8	8	8	8	A,B,C,empl,emplexp	HASH_JOIN	

hdfs://localhost:8020/tmp/temp-886396143/tmp1902265131,

Input(s):

Successfully read 14 records from: "/hadoopdata/pig/employee_details1.txt"

Successfully read 9 records from: "/hadoopdata/pig/employee_expenses1.txt"

Output(s):

Successfully stored 8 records (126 bytes) in: "hdfs://localhost:8020/tmp/temp-886396143/tmp1902265131"

Counters:

Total records written : 8

Total bytes written : 126

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

Job DAG:

job_1522227177035_0006

```

2018-03-28 15:09:02,678 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 15:09:02,689 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 15:09:02,803 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032

```

Successfully stored 8 records (126 bytes) in: "hdfs://localhost:8020/tmp/temp-886396143/tmp1902265131"

Counters:

Total records written : 8

Total bytes written : 126

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

Job DAG:

job_1522227177035_0006

```

2018-03-28 15:09:02,678 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 15:09:02,689 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 15:09:02,803 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 15:09:02,816 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 15:09:02,934 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-28 15:09:02,951 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-28 15:09:03,055 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-03-28 15:09:03,055 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-28 15:09:03,064 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
2018-03-28 15:09:03,091 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-03-28 15:09:03,091 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

```

```

(102,Akshay,)
(106,Aamir,)
(107,Salman,)
(108,Ranbir,)
(109,Katrina,)
(111,Tushar,)
(112,Ajay,)
(113,Jubeen,)
grunt>

```

List of employee who do
not have entry in
employee expenses

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/>

You can download the datasets from the following links:

[Delayed_Flights.csv](#)

[Airports.csv](#)

Delayed_Flights.csv Datasets

There are 29 columns in this dataset. Some of them have been mentioned below:

- Year: 1987 – 2008
- Month: 1 – 12
- FlightNum: Flight number
- Canceled: Was the flight canceled?
- CancellationCode: The reason for cancellation.

For complete details, refer to [this link](#).

Airports.csv Datasets

- iata: the international airport abbreviation code
- name of the airport
- city and country in which airport is located.
- lat and long: the latitude and longitude of the airport

Problem Statement 1: Find out the top 5 most visited destinations.

Below is the pig script run in local mode:

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

We are registering the “piggybank.jar” file which is saved in the above mentioned path.

The *Piggy Bank* is a place for Pig users to share their functions.

Piggybank.jar file is an inbuilt jar file from pig which is used to parse nested data for the dataset used in the program.

For Example: Below is the table of employee details

EmployeeID	City	Employee Name,Company	designation	phoneno
101	Bangalore	{Aaradya, IBM}	Developer	942545654
102	Bangalore	{Kayal,IBM}	Tester	983452343
103	Bangalore	{Arun,IBM}	Developer	953342534

If this data is getting loaded to Pig *without the Piggybank.jar file*, it will treat all the columns as separate columns even though the employee name and company name data is nested. As shown below:

Column1	Column2	Column3	Column4	Column5	Column6
EmployeeID	City	Employee Name	Company	designation	phoneno
101	Bangalore	{Aaradya	IBM}	Developer	942545654
102	Bangalore	{Kayal	IBM}	Tester	983452343
103	Bangalore	{Arun	IBM}	Developer	953342534

To avoid this issue, we will use *piggybank.jar file* to make explicitly consider if a column consists of multiple values which are under bracket then consider them as single column value and not split in the next subsequent column

If this data is getting loaded to Pig *with the Piggybank.jar file*

Column1	Column2	Column3	Column4	Column5
EmployeeID	City	Employee Name,Company	designation	phoneno
101	Bangalore	{Aaradya, IBM}	Developer	942545654
102	Bangalore	{Kayal,IBM}	Tester	983452343
103	Bangalore	{Arun,IBM}	Developer	953342534

```
A = load '/hadoopdata/pig/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_H
EADER');
```

This command will load the DelayedFlights.csv file from HDFS to the relation A and skip the header of the input file.

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

This command will generate columns namely: “year,flight number,origin and destination” for each elements of relation A and save it in relation B.

C = filter B by dest is not null;

This command will filter the elements of relation B whose values are not null and discard the values with null.

D = group C by dest;

This command will group all the elements of relation C by destination.

E = foreach D generate group, COUNT(C.dest);

This command will count the number of visits to all destinations to find the total number of visits for all destinations respectively.

F = order E by \$1 DESC;

This command will order all elements of relation E by descending order.

Result = LIMIT F 5;

This is command will restrict the number of rows to be displayed in the result to 5.

**A1 = load '/hadoopdata/pig/airports.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_H
EADER');**

This command will load “airports.csv” file from HDFS to relation “A1” and skip the header of the input file.

**A2 = foreach A1 generate (chararray)\$0 as dest, (chararray)\$2 as city, (chararray)\$4 as
country;**

This command will generate the columns namely: destination, city and name of the country for each elements of relation A2. We are loading this table from which we will look-up and find the city as well as the country.

joined_table = join Result by \$0, A2 by dest;

This command will join two tables from relation A2 and result, to obtain the top5 most visited destinations.

dump joined_table;

This command will run the Map Reduce program to get the top5 number of visits to the destination

```

grunt> A = load '/hadoopdata/pig/DelavedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(' ','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-03-29 10:02:29,932 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin, (chararray)$18 as dest;
grunt> C = filter B by dest is not null;
grunt> D = group C by dest;
grunt> E = foreach D generate group, COUNT(C.dest);
grunt> F = order E by $1 DESC;
grunt> Result = LIMIT F 5;
grunt> A1 = load '/hadoopdata/pig/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(' ','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-03-29 10:03:40,016 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
grunt> joined_table = join Result by $0, A2 by dest;
grunt> dump joined_table;
2018-03-29 10:04:08,060 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN, GROUP_BY, ORDER_BY, FILTER, LIMIT
2018-03-29 10:04:08,139 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-29 10:04:08,142 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
2018-03-29 10:04:08,145 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-29 10:04:08,157 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-03-29 10:04:08,259 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2018-03-29 10:04:08,266 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-177
2018-03-29 10:04:08,266 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POForEach->POForEach to POPackage(JoinPackager)
2018-03-29 10:04:08,268 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 5
2018-03-29 10:04:08,268 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 5

2018-03-29 10:20:17,081 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:20:17,103 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:20:17,298 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-29 10:20:17,443 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.5 0.16.0 acadgild 2018-03-29 10:15:47 2018-03-29 10:20:17 HASH_JOIN, GROUP_BY, ORDER_BY, FILTER, LIMIT

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime M
MedianReduceTime Alias Feature Outputs
job_1522298651170_0003 2 1 52 50 51 51 8 8 8 8 A,B,C,D,E GROUP_BY,COMBINER
job_1522298651170_0004 1 1 7 7 7 7 8 8 8 8 F SAMPLER
job_1522298651170_0005 1 1 7 7 7 7 7 7 7 7 F ORDER_BY,COMBINER
job_1522298651170_0006 1 1 7 7 7 7 7 7 7 7 F
job_1522298651170_0007 2 1 18 17 18 18 9 9 9 9 A1,A2,joined_table HASH_JOIN
hdfs://localhost:8020/tmp/temp-864112345/tmp-1344442008,

Input(s):
Successfully read 1936758 records (247968072 bytes) from: "/hadoopdata/pig/DelavedFlights.csv"
Successfully read 3376 records from: "/hadoopdata/pig/airports.csv"

Output(s):
Successfully stored 5 records (198 bytes) in: "hdfs://localhost:8020/tmp/temp-864112345/tmp-1344442008"

Counters:
Total records written : 5
Total bytes written : 198
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1522298651170_0003 -> job_1522298651170_0004,
job_1522298651170_0004 -> job_1522298651170_0005,

```

```

tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:20:18,594 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:20:18,609 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:20:18,743 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:20:18,760 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:20:18,871 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:20:18,886 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:20:19,020 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:20:19,037 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:20:19,151 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:20:19,160 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:20:19,277 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:20:19,286 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:20:19,393 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:20:19,415 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:20:19,520 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:20:19,530 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:20:19,641 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-03-29 10:20:19,655 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-29 10:20:19,660 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2018-03-29 10:20:19,677 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-03-29 10:20:19,677 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(ATL,106898,ATL,Atlanta,USA)
(DEN,63803,DEN,Denver,USA)
(DFW,70657,DFW,Dallas-Fort Worth,USA)
(LAX,59969,LAX,Los Angeles,USA)
(ORD,108984,ORD,Chicago,USA)
grunt>

```

← Top 5 visited destinations

Problem Statement 2: Which month has seen the most number of cancellations due to bad weather?

Below is the pig script run in local mode:

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

We are registering the “piggybank.jar” file which is saved in the above mentioned path.

```
A = load '/hadoopdata/pig/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_H
EADER');
```

This command will load the DelayedFlights.csv file from HDFS to the relation A and skip the header of the input file.

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

This command will generate columns namely: “month, flightnumber, cancelled, cancelledcode” for each elements in relation “A”.

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

This command will filter all the cancelled flights due to ***“Bad weather”***.

We are filtering the data based on cancellation and cancellation code, i.e., cancelled = 1 means flight have been cancelled and cancel_code = 'B' means the reason for cancellation is “weather.” So relation C will point to the data which consists of canceled flights due to bad weather.

D = group C by month;

This command will group all elements of relation “C” by month and save it in the relation “D”.

E = foreach D generate group, COUNT(C.cancelled);

This command will generate the count of all cancelled flights for every month and group them together to save it in the relation “E”.

F= order E by \$1 DESC;

This command will sort the elements of the relation “E” in descending order.

Result = limit F 1;

This command will limit the number of outputs to 1st row of relation “F”, i.e. finding the top month based on cancellation.

dump Result;

This command will run the Map Reduce program to get the most number of cancellations due to bad weather.


```

grunt>
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
2018-03-29 10:27:02,987 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
grunt> A = load '/hadoopdata/pig/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SK
IP_INPUT_HEADER');
2018-03-29 10:27:14,434 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
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;
2018-03-29 10:28:15,133 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY, ORDER_BY, FILTE
R, LIMIT
2018-03-29 10:28:15,206 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-29 10:28:15,208 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2018-03-29 10:28:15,208 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMa
pKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilt
erOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-29 10:28:15,222 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thresho
ld: 100 optimistic? false
2018-03-29 10:28:15,236 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebrai
c foreach to combiner
2018-03-29 10:28:15,242 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Seconda
ry Key Optimization for MapReduce node scope-159
2018-03-29 10:28:15,247 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size befo
re optimization: 4
2018-03-29 10:28:15,247 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size afte
r optimization: 4
2018-03-29 10:28:15,277 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-29 10:28:15,291 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:28:15,298 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job

```

```

tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:31:23,035 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:31:23,043 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:31:23,139 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-29 10:31:23,168 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SamplePigStats - Script Statistics:

```

HadoopVersion	PigVersion	UserId	StartedAt	FinishedAt	Features
2.6.5	0.16.0	acadgild	2018-03-29 10:28:15	2018-03-29 10:31:23	GROUP_BY, ORDER_BY, FILTER, LIMIT

Success!

Job Stats (time in seconds):

JobId	Maps	Reduces	MaxMapTime	MinMapTime	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime	M
job_1522298651170_0008	2	1	37	36	37	37	8	8	8	A,B,C,D,E GROUP_BY,COMBINER
job_1522298651170_0009	1	1	8	8	8	8	8	8	8	F SAMPLER
job_1522298651170_0010	1	1	7	7	7	7	7	7	7	F ORDER_BY,COMBINER
job_1522298651170_0011	1	1	7	7	7	7	8	8	8	F hdfs://localhost:8020/tmp/temp-864112345/tmp654158051,

Input(s):

Successfully read 1936758 records (247968072 bytes) from: "/hadoopdata/pig/DelayedFlights.csv"

Output(s):

Successfully stored 1 records (9 bytes) in: "hdfs://localhost:8020/tmp/temp-864112345/tmp654158051"

Counters:

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

Job DAG:

```

job_1522298651170_0008 -> job_1522298651170_0009,
job_1522298651170_0009 -> job_1522298651170_0010,
job_1522298651170_0010 -> job_1522298651170_0011,

```

```

2018-03-29 10:31:23,363 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:31:23,432 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:31:23,438 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:31:23,483 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:31:23,491 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:31:23,537 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:31:23,560 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:31:23,643 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:31:23,650 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:31:23,725 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:31:23,743 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:31:23,821 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:31:23,831 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:31:23,888 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:31:23,900 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:31:23,977 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:31:23,991 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:31:24,061 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:31:24,072 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:31:24,153 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-03-29 10:31:24,157 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-29 10:31:24,158 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2018-03-29 10:31:24,166 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-03-29 10:31:24,166 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
grunt>

```

← This indicates that december month had most number of cancellations due to bad weather

Problem Statement 3: Top ten origins with the highest AVG departure delay

Below is the pig script run in local mode:

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

We are registering the “piggybank.jar” file which is saved in the above mentioned path.

```
A = load '/hadoopdata/pig/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_H
EADER');
```

This command will load the DelayedFlights.csv file from HDFS to the relation A and skip the header of the input file

```
B1 = foreach A generate (int)$16 as dep_delay, (chararray)$17 as origin;
```

This command will generate the following columns for each element of relation “A”:
departure_delay, origin.

C1 = filter B1 by (dep_delay is not null) AND (origin is not null);

This command will filter all the elements in relation B1 by, departure_delay and origin values should not be null. i.e. we are removing all null values in dep_delay and origin columns.

D1 = group C1 by origin;

This command will group all the non-null values in origin and save it in the relation "D1".

E1 = foreach D1 generate group, AVG(C1.dep_delay);

This command will generate average of departure_delay values for each unique origin, group the average and save it in the relation "E1"

Result = order E1 by \$1 DESC;

This command will sort all the elements of relation E1 in descending order.

Top_ten = limit Result 10;

This command will limit to output to top 10 values.

**Lookup = load '/hadoopdata/pig/airports.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_H
EADER');**

This command will load the airports.csv file from HDFS to the relation "Lookup" and skip the header of the input file. we are loading another table so that we can look up and find the city as well as the country.

**Lookup1 = foreach Lookup generate (chararray)\$0 as origin, (chararray)\$2 as city,
(chararray)\$4 as country;**

This command will generate following columns for each element of relation "Lookup": origin, city and country.

Joined = join Lookup1 by origin, Top_ten by \$0;

This command will join values of relation "top_ten" and "Lookup1" by using origin as the joining factor, because we have to find top ten origins with the highest AVG departure delay

Final = foreach Joined generate \$0, \$1, \$2, \$4;

This command will generate required column values of relation "joined" and save it in the relation "Final".

Final_Result = ORDER Final by \$3 DESC;

sort the elements of "Final" in descending order.

dump Final_Result;

This command will run the Map Reduce program to get top ten origins with the highest AVG departure delay.

```
grunt>
grunt> REGISTER '/home/acadqild/install/pig/pig-0.16.0/lib/piggybank.jar';
grunt> A = load '/hadoopdata/pig/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-03-29 10:38:33,464 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B1 = foreach A generate (int)$16 as dep_delay, (chararray)$17 as origin;
grunt> C1 = filter B1 by (dep_delay is not null) AND (origin is not null);
grunt> D1 = group C1 by origin;
grunt> E1 = foreach D1 generate group, AVG(C1.dep_delay);
grunt> Result = order E1 by $1 DESC;
grunt> Top_ten = limit Result 10;
grunt> Lookup = load '/hadoopdata/pig/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-03-29 10:39:33,361 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;
grunt> Joined = join Lookup1 by origin, Top_ten by $0;
grunt> Final = foreach Joined generate $0,$1,$2,$4;
grunt> Final_Result = ORDER Final by $3 DESC;
grunt> dump Final_Result;
2018-03-29 10:40:24,557 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN, GROUP_BY, ORDER_BY, FILTER, LIMIT
2018-03-29 10:40:24,675 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-29 10:40:24,678 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
2018-03-29 10:40:24,679 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-29 10:40:24,694 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-03-29 10:40:24,709 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2018-03-29 10:40:24,715 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-282
2018-03-29 10:40:24,715 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-315
```



```

2.6.5 0.16.0 acadgild 2018-03-29 10:40:24 2018-03-29 10:46:07 HASH_JOIN, GROUP_BY, ORDER_BY, FILTER, LIMIT

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime M
MedianReduceTime Alias Feature Outputs
job_1522298651170_0012 2 1 54 51 53 53 9 9 9 9 A,B1,C1,D1,E1 GROUP_BY,COMBINER
job_1522298651170_0013 1 1 7 7 7 7 8 8 8 8 Result SAMPLER
job_1522298651170_0014 1 1 7 7 7 7 8 8 8 8 Result ORDER_BY,COMBINER
job_1522298651170_0015 1 1 7 7 7 7 7 7 7 7 Result
job_1522298651170_0016 2 1 16 15 15 15 8 8 8 8 Final,Joined,Lookup,Lookup1 H
HASH_JOIN
job_1522298651170_0017 1 1 8 8 8 8 8 8 8 8 Final_Result SAMPLER
job_1522298651170_0018 1 1 8 8 8 8 8 8 8 8 Final_Result ORDER_BY h
hdfs://localhost:8020/tmp/temp-864112345/tmp1745689741,

Input(s):
Successfully read 1936758 records (247968072 bytes) from: "/hadoopdata/pig/DelayedFlights.csv"
Successfully read 3376 records from: "/hadoopdata/pig/airports.csv"

Output(s):
Successfully stored 10 records (351 bytes) in: "hdfs://localhost:8020/tmp/temp-864112345/tmp1745689741"

Counters:
Total records written : 10
Total bytes written : 351
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1522298651170_0012 -> job_1522298651170_0013,
job_1522298651170_0013 -> job_1522298651170_0014,
job_1522298651170_0014 -> job_1522298651170_0015,
job_1522298651170_0015 -> job_1522298651170_0016,
job_1522298651170_0016 -> job_1522298651170_0017,
job_1522298651170_0017 -> job_1522298651170_0018,

```

```

2018-03-29 10:46:08,758 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:46:08,797 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:46:08,801 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:46:08,886 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:46:08,892 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:46:08,962 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:46:08,973 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:46:09,077 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:46:09,089 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:46:09,218 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:46:09,234 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:46:09,364 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 10:46:09,376 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 10:46:09,489 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-03-29 10:46:09,498 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-29 10:46:09,499 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2018-03-29 10:46:09,516 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-03-29 10:46:09,516 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(CMX,Hancock,USA,116.1478588235294)
(PLN,Pellston,USA,93.76198476198476)
(SPI, Springfield, USA, 83.84873949579831)
(ALO, Waterloo, USA, 82.2258864516129)
(MQT, NA, USA, 79.55665824630542)
(ACY, Atlantic City, USA, 79.3103448275862)
(MOT, Minot, USA, 78.66165413533835)
(HHH, NA, USA, 76.53005464488874)
(EGE, Eagle, USA, 74.12891986062718)
(BGM, Binghamton, USA, 73.15533980582525)
grant>

```

← List of top10 origins with the highest AVG departure delay

Problem Statement 4: Which route (origin & destination) has seen the maximum diversion?

Below is the pig script run in local mode:

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

We are registering the “piggybank.jar” file which is saved in the above mentioned path.

```
A = load '/hadoopdata/pig/DelayedFlights.csv' USING  
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_H  
EADER');
```

This command will load the DelayedFlights.csv file from HDFS to the relation A and skip the header of the input file.

```
B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as  
diversion;
```

This command will generate following columns for each elements of the relation “A”: origin, destination and diversion.

```
C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);
```

This command will filter out all null values in origin and destination and diversion =1, which means it will remove all null records if any in columns origin and destination and save all the records for which the diversion is “True”, in the relation “C”.

```
D = GROUP C by (origin,dest);
```

This will group all values by origin and destination and save them in the relation “D”.

```
E = FOREACH D generate group, COUNT(C.diversion);
```

This will generate count of number of diversion taken per unique origin and destination.

```
F = ORDER E BY $1 DESC;
```

This will sort the elements of relation “E” in descending order.

```
Result = limit F 10;
```

This will limit the output to top 10 values, because you require route (origin & destination) with maximum diversion

```
dump Result;
```

This command will run the Map Reduce program to get route (origin & destination) with maximum diversion

```

2018-03-29 12:03:26,634 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
2018-03-29 12:06:32,966 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> A = load '/hadoopdata/pig/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER');
2018-03-29 12:06:42,967 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;
grunt> C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);
grunt>
grunt> D = GROUP C by (origin,dest);
grunt> E = FOREACH D generate group, COUNT(C.diversion);
grunt> F = ORDER E BY $1 DESC;
grunt> Result = limit F 10;
grunt> dump Result;
2018-03-29 12:07:41,547 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY,FILTER,LIMIT
2018-03-29 12:07:41,635 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-29 12:07:41,644 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
2018-03-29 12:07:41,753 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMajorKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-29 12:07:41,929 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 489350752
2018-03-29 12:07:42,047 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-03-29 12:07:42,117 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2018-03-29 12:07:42,186 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope=41
2018-03-29 12:07:42,243 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4
2018-03-29 12:07:42,243 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 4

```

```

tus=SUCCEEDED. Redirecting to job history server
2018-03-29 12:11:31,652 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-29 12:11:31,724 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt      FinishedAt      Features
2.6.5          0.16.0      acadgild  2018-03-29 12:07:43  2018-03-29 12:11:31  GROUP_BY,ORDER_BY,FILTER,LIMIT

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  M
MedianReductetime Alias  Feature Outputs
job_1522305150621_0001  2  1  45  45  45  45  10  10  10  10  A,B,C,D,E  GROUP_BY,COMBINER
job_1522305150621_0002  1  1  9  9  9  8  8  8  8  8  F  SAMPLER
job_1522305150621_0003  1  1  7  7  7  8  8  8  8  8  F  ORDER_BY,COMBINER
job_1522305150621_0004  1  1  7  7  7  10  10  10  10  10  F
8020/tmp/temp-1425484776/tmp-271330620,

Input(s):
Successfully read 1936758 records (247968072 bytes) from: "/hadoopdata/pig/DelayedFlights.csv"

Output(s):
Successfully stored 10 records (190 bytes) in: "hdfs://localhost:8020/tmp/temp-1425484776/tmp-271330620"

Counters:
Total records written : 10
Total bytes written : 190
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1522305150621_0001 -> job_1522305150621_0002,
job_1522305150621_0002 -> job_1522305150621_0003,
job_1522305150621_0003 -> job_1522305150621_0004,
job_1522305150621_0004

```



```

2018-03-29 12:11:32,411 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 12:11:32,569 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 12:11:32,585 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 12:11:32,739 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 12:11:32,753 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 12:11:32,950 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 12:11:32,969 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 12:11:33,122 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 12:11:33,132 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 12:11:33,227 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 12:11:33,239 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 12:11:33,332 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-29 12:11:33,349 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-03-29 12:11:33,491 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-03-29 12:11:33,498 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-03-29 12:11:33,503 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2018-03-29 12:11:33,517 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-03-29 12:11:33,517 [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)
grunt>

```



List of route (origin &
destination) has seen the
maximum diversion