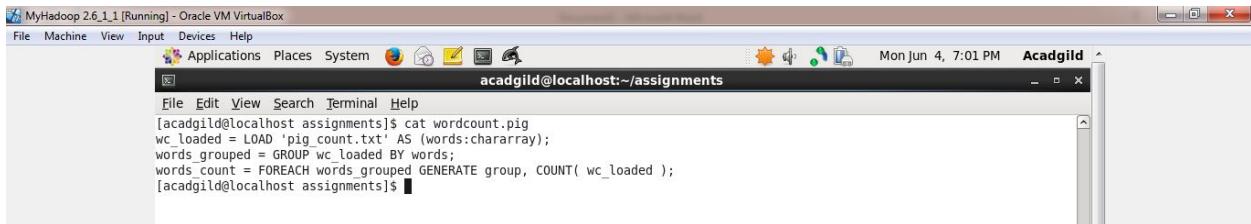


## Task 1

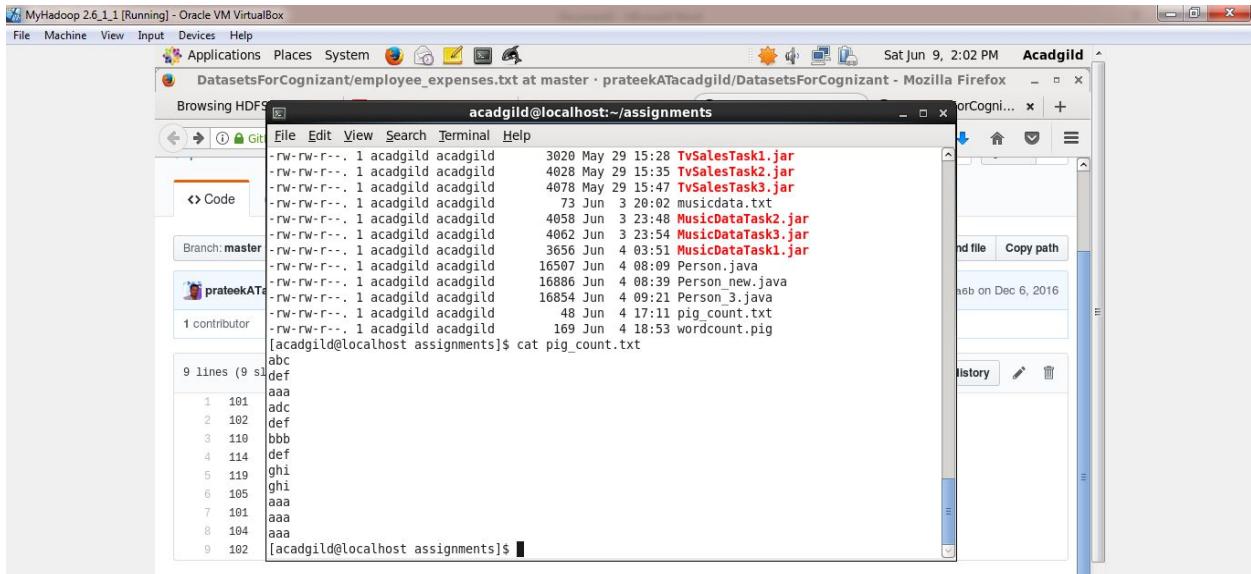
Write a program to implement wordcount using Pig.

Load input file using LOAD Command. Create a new file from loaded file by applying group by words. Get the word count by using FOREACH command.



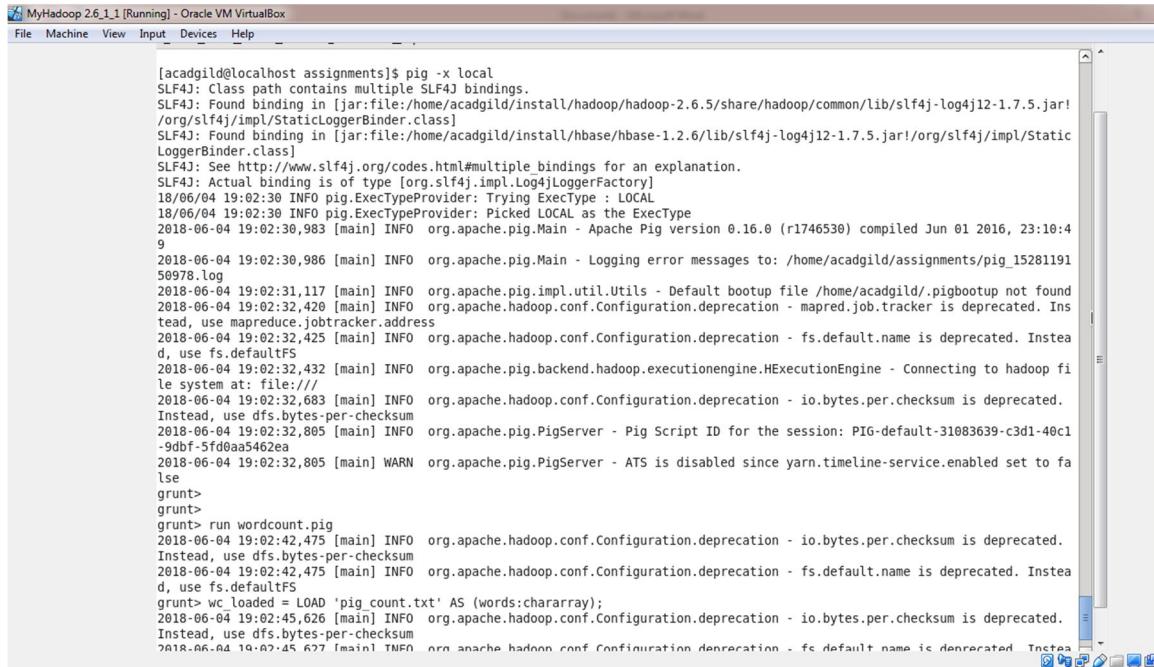
```
[MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox]
File Machine View Input Devices Help
Applications Places System Mozilla Firefox Mon Jun 4, 7:01 PM Acadgild
acadgild@localhost:~/assignments
File Edit View Search Terminal Help
[acadgild@localhost assignments]$ cat wordcount.pig
wc_loaded = LOAD 'pig_count.txt' AS (words:chararray);
words_grouped = GROUP wc_loaded BY words;
words_count = FOREACH words_grouped GENERATE group, COUNT( wc_loaded );
[acadgild@localhost assignments]$
```

Input file:



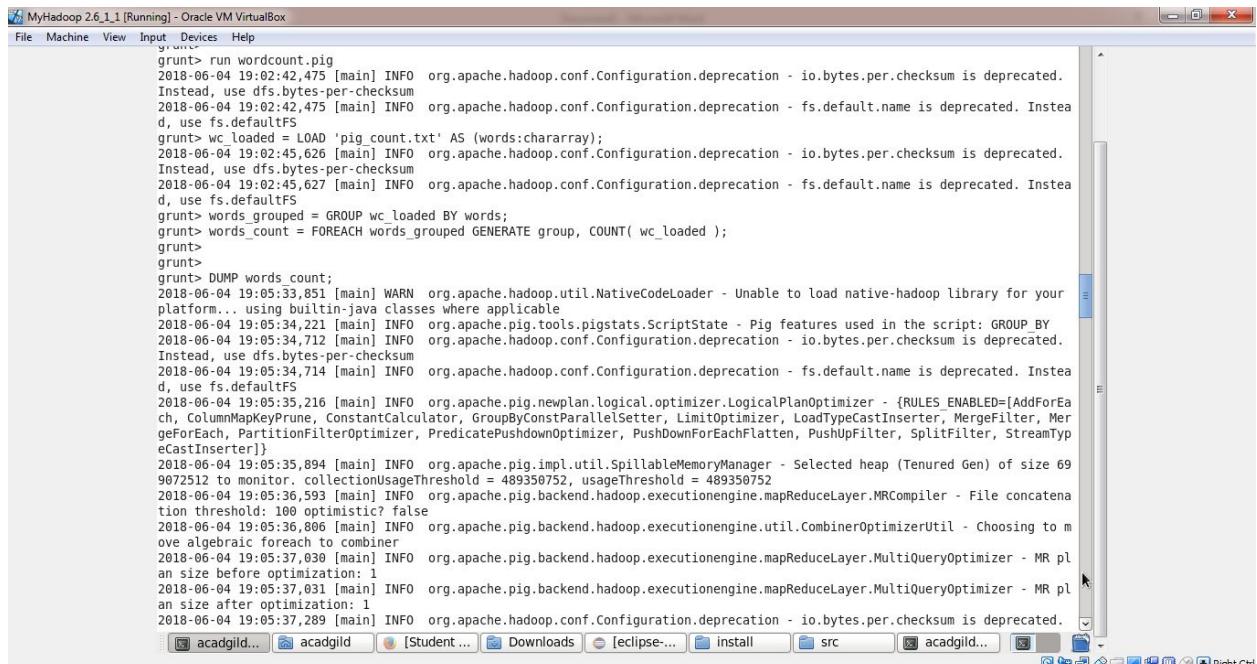
```
[MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox]
File Machine View Input Devices Help
Applications Places System Mozilla Firefox Sat Jun 9, 2:02 PM Acadgild
DatasetsForCognizant/employee_expenses.txt at master · prateekATAcadgild/DatasetsForCognizant - Mozilla Firefox
Browsing HDFS acadgild@localhost:~/assignments
File Edit View Search Terminal Help
Branch:master
prateekATAcadgild
1 contributor
abc
9 lines (9 s)
def
1 101 adc
2 102 def
3 110 bbb
4 114 def
5 119 ghi
6 105 aaa
7 101 aaa
8 104 aaa
9 102 [acadgild@localhost assignments]$ cat pig_count.txt
[acadgild@localhost assignments]$
```

Connect to pig grunt shell using command – **pig -x local** and run pig script using run **wordcount.pig**



```
[acadgild@localhost assignments]$ pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!
/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/06/04 19:02:30 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/06/04 19:02:30 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-06-04 19:02:30,983 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:4
9
2018-06-04 19:02:30,986 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/assignments/pig_15281191
56978.log
2018-06-04 19:02:31,117 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-06-04 19:02:32,420 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-06-04 19:02:32,425 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-04 19:02:32,432 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
le system at: file:///
2018-06-04 19:02:32,683 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-04 19:02:32,805 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PI0-default-31083639-c3d1-48c1
-9df5-5fd0aa5462ea
2018-06-04 19:02:32,805 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to fa
lse
grunt>
grunt>
grunt> run wordcount.pig
2018-06-04 19:02:42,475 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-04 19:02:42,475 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunts wc_loaded = LOAD 'pig_count.txt' AS (words:chararray);
2018-06-04 19:02:45,626 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-04 19:02:45,627 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunts words_grouped = GROUP wc_loaded BY words;
grunts words_count = FOREACH words_grouped GENERATE group, COUNT( wc_loaded );
grunt>
grunt>
grunts DUMP words_count;
2018-06-04 19:05:33,851 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using built-in java classes where applicable
2018-06-04 19:05:34,221 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2018-06-04 19:05:34,712 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-04 19:05:34,714 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-04 19:05:35,216 [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]}
2018-06-04 19:05:35,894 [main] INFO org.apache.pig.impl.util.SplittableMemoryManager - Selected heap (Tenured Gen) of size 69
9672512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 489350752
2018-06-04 19:05:36,593 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-06-04 19:05:36,886 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to m
ove algebraic foreach to combiner
2018-06-04 19:05:37,030 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size before optimization: 1
2018-06-04 19:05:37,031 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size after optimization: 1
2018-06-04 19:05:37,289 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
```

Check output using DUMP command



```
[acadgild@localhost assignments]$ pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!
/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/06/04 19:02:30 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/06/04 19:02:30 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-06-04 19:02:30,983 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:4
9
2018-06-04 19:02:30,986 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/assignments/pig_15281191
56978.log
2018-06-04 19:02:31,117 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-06-04 19:02:32,420 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-06-04 19:02:32,425 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-04 19:02:32,432 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
le system at: file:///
2018-06-04 19:02:32,683 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-04 19:02:32,805 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PI0-default-31083639-c3d1-48c1
-9df5-5fd0aa5462ea
2018-06-04 19:02:32,805 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to fa
lse
grunt>
grunt>
grunt> run wordcount.pig
2018-06-04 19:02:42,475 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-04 19:02:42,475 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunts wc_loaded = LOAD 'pig_count.txt' AS (words:chararray);
2018-06-04 19:02:45,626 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-04 19:02:45,627 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunts words_grouped = GROUP wc_loaded BY words;
grunts words_count = FOREACH words_grouped GENERATE group, COUNT( wc_loaded );
grunt>
grunt>
grunts DUMP words_count;
2018-06-04 19:05:33,851 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using built-in java classes where applicable
2018-06-04 19:05:34,221 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2018-06-04 19:05:34,712 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-04 19:05:34,714 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-04 19:05:35,216 [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]}
2018-06-04 19:05:35,894 [main] INFO org.apache.pig.impl.util.SplittableMemoryManager - Selected heap (Tenured Gen) of size 69
9672512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 489350752
2018-06-04 19:05:36,593 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-06-04 19:05:36,886 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to m
ove algebraic foreach to combiner
2018-06-04 19:05:37,030 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size before optimization: 1
2018-06-04 19:05:37,031 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size after optimization: 1
2018-06-04 19:05:37,289 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
```

```

MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Output(s):
Successfully stored 6 records in: "file:/tmp/temp-1737364731/tmp-805191814"

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

Job DAG:
job_local989426029_0001

2018-06-04 19:05:51,706 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-04 19:05:51,708 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-04 19:05:51,719 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-04 19:05:51,803 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLauncher - Success!
2018-06-04 19:05:51,824 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-06-04 19:05:51,826 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-04 19:05:51,826 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-04 19:05:51,916 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-04 19:05:51,916 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(aaa,4)
(abc,1)
(adc,1)
(bbb,1)
(def,3)
(ghi,2)
grunt> 

```

The screenshot shows a terminal window within a virtual machine. The terminal output displays Hadoop metrics, a successful map/reduce job completion, and a Pig script execution. The desktop environment includes icons for acadgild, eclipse, and various system files.

## Task 2

We have employee\_details and employee\_expenses files. Use local mode while running Pig and write Pig Latin script to get below results: employee\_details (EmpID,Name,Salary,DepartmentID)

[https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee\\_details.txt](https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee_details.txt)

employee\_expenses(EmpID,Expence)

[https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee\\_expenses.txt](https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee_expenses.txt)

```

MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Terminal Help
acadgild@localhost:~/assignments
File Edit View Search Terminal Help
[acadgild@localhost ~]$ cd assignments/
[acadgild@localhost assignments]$ ls
lgb.txt      employee_expenses.txt  MusicDataTask3.jar  Person_new.java  TvSalesTask2.jar  wordmean.txt
assignment1.txt  max-temp.txt    musicdata.txt    pig.temp.txt   TvSalesTask3.jar  wordmedian.txt
create_one_gb_file.sh  MusicDataTask1.jar  Person_3.java  television.txt  wordcount.pig  wordSD.txt
employee_details.txt  MusicDataTask2.jar  Person.java   TvSalesTask1.jar  word-count.txt
[acadgild@localhost assignments]$ cat employee_details.txt
101,Anitabh,20000,1
102,Shahrukh,10000,2
103,Akshay,11000,3
104,Anubhav,5000,4
105,Pawan,2500,5
106,Aamir,25000,1
107,Salman,17500,2
108,Rambir,14000,3
109,Katrina,1000,4
110,Priyanka,2000,5
111,Tushar,500,1
112,Ajay,5000,2
113,Jubeen,1000,1
114,Madhuri,2000,2
[acadgild@localhost assignments]$ cat employee_expenses.txt
101    200
102    100
110    400
114    200
119    200
105    100
101    100
104    300
102    400
[acadgild@localhost assignments]$ 

```

The screenshot shows a terminal window with a Linux desktop interface. It lists files in the assignments directory and displays the contents of employee\_details and employee\_expenses files. The desktop includes icons for Applications, Places, System, and a terminal window titled 'Acadgild'.

(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)

Load input file by using LOAD command. Order the input file by rating in descending order and name in ascending order. From the ordered list fetch first top five records.

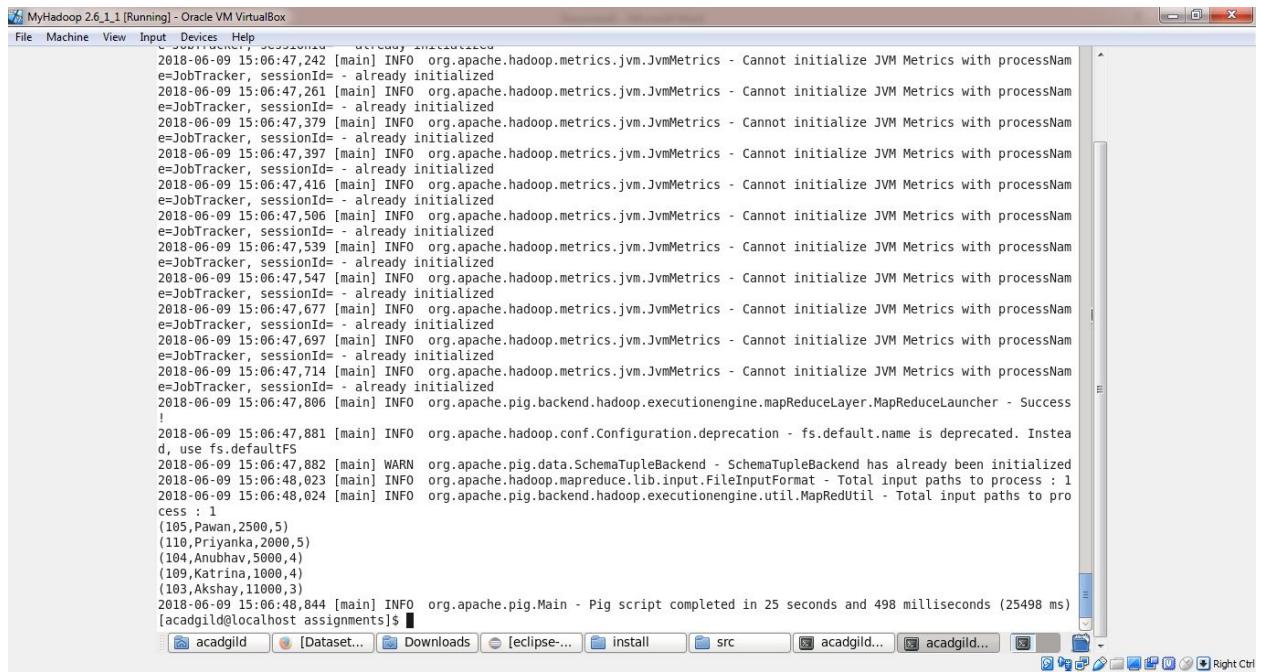
### Pig Script to fetch top 5 rating:

```
MyHadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Terminal acadgild@localhost:~/assignments
[acadgild@localhost assignments]$ ls
lgb.txt      employee_expenses.txt  MusicDataTask3.jar  Person_new.java  TVSalesTask1.jar  word-count.txt
assignment1.txt  max-temp.txt       musicdata.txt     pig_count.txt   TVSalesTask2.jar  wordmean.txt
create_one_gb_file.sh MusicDataTask1.jar  Person_3.java  television.txt  TVSalesTask3.jar  wordmedian.txt
employee_details.txt  MusicDataTask2.jar  Person.java    top5rating.pig  wordcount.pig   wordSD.txt
[acadgild@localhost assignments]$ vi top5rating.pig
[acadgild@localhost assignments]$ cat top5rating.pig
details = LOAD 'employee_details.txt' USING PigStorage(',') AS ( EmpId:int, Name:chararray, Salary:int, Rating: int );
details ordered = ORDER details BY Rating DESC, Name ASC;
Top5_Ratings = LIMIT details_ordered 5;
DUMP Top5_Ratings;

[acadgild@localhost assignments]$
```

```
MyHadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Terminal acadgild@localhost:~/assignments
[acadgild@localhost assignments]$ You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$ pig -x local top5rating.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/[org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/[org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/06/09 15:06:25 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/06/09 15:06:25 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-06-09 15:06:25,607 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-06-09 15:06:25,610 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/assignments/pig_1528536985600.log
2018-06-09 15:06:25,839 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2018-06-09 15:06:27,633 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-06-09 15:06:28,459 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-06-09 15:06:28,463 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-09 15:06:28,474 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2018-06-09 15:06:28,605 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PI6-top5rating.pig-c1048f83-d1cd-4dc5-92bf-fcc40b366c5c
2018-06-09 15:06:28,606 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-06-09 15:06:30,334 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```

## Output:



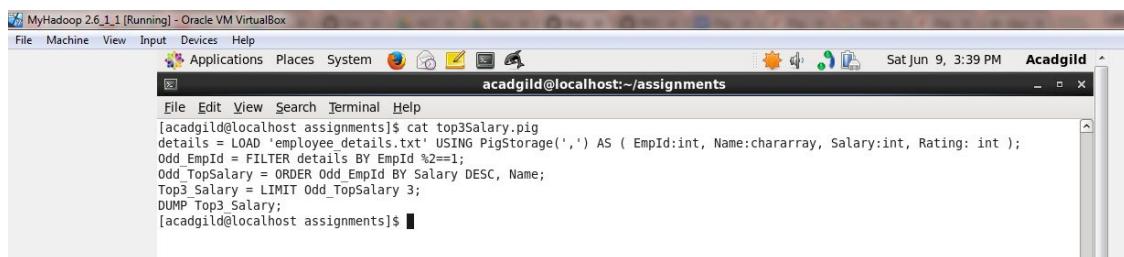
```

MyHadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
e=JobTracker, sessionId= - already initialized
2018-06-09 15:06:47,242 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 15:06:47,261 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 15:06:47,379 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 15:06:47,397 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 15:06:47,416 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 15:06:47,506 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 15:06:47,539 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 15:06:47,547 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 15:06:47,677 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 15:06:47,697 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 15:06:47,714 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 15:06:47,806 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLauncher - Success !
2018-06-09 15:06:47,881 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-09 15:06:47,882 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-09 15:06:48,023 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-09 15:06:48,024 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(105,Pawan,2500,5)
(110,Priyanka,2000,5)
(104,Anubhav,5000,4)
(109,Katrina,1000,4)
(103,Akshay,1100,3)
2018-06-09 15:06:48,844 [main] INFO org.apache.pig.Main - Pig script completed in 25 seconds and 498 milliseconds (25498 ms)
[acadgild@localhost assignments]$ 

```

(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)

Load input file using LOAD command. Filter odd records by using modulus function EmpId %2 == 1. Sort the list on descending order of salary and ascending order of name. Fetch top 3 records to get output.



```

MyHadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Terminal Help
acadgild@localhost:~/assignments$ cat top3Salary.pig
details = LOAD 'employee details.txt' USING PigStorage(',') AS ( EmpId:int, Name:chararray, Salary:int, Rating: int );
Odd_EmpId = FILTER details BY EmpId %2==1;
Odd_TopSalary = ORDER Odd_EmpId BY Salary DESC, Name;
Top3_Salary = LIMIT Odd_TopSalary 3;
DUMP Top3_Salary;
[acadgild@localhost assignments]$ 

```

```

[acadgild@localhost assignments]$ cat top3Salary.pig
details = LOAD 'employee details.txt' USING PigStorage(',') AS ( EmpId:int, Name:chararray, Salary:int, Rating: int );
Odd_EmpId = FILTER details BY Empid %2==1;
Odd_TopSalary = ORDER Odd_EmpId BY Salary DESC, Name;
Top3_Salary = LIMIT Odd_TopSalary 3;
DUMP Top3_Salary;
[acadgild@localhost assignments]$ pig -x local top3Salary.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/
/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/
/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/06/09 15:42:34 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/06/09 15:42:34 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-06-09 15:42:34,138 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:4
9
2018-06-09 15:42:34,139 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/assignments/pig_15285391
54134.log
2018-06-09 15:42:34,259 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use
mapreduce.job.user.name
2018-06-09 15:42:35,354 [main] INFO org.apache.pig.impl.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-06-09 15:42:36,015 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Inste
ad, use mapreduce.jobtracker.address
2018-06-09 15:42:36,015 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-09 15:42:36,030 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
le system at: file:///
2018-06-09 15:42:36,149 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PI6-top3Salary.pig-12d4bc9f-f1
01-4a42-8c95-c887ff9d8cb5
2018-06-09 15:42:36,149 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to fa
lse

```

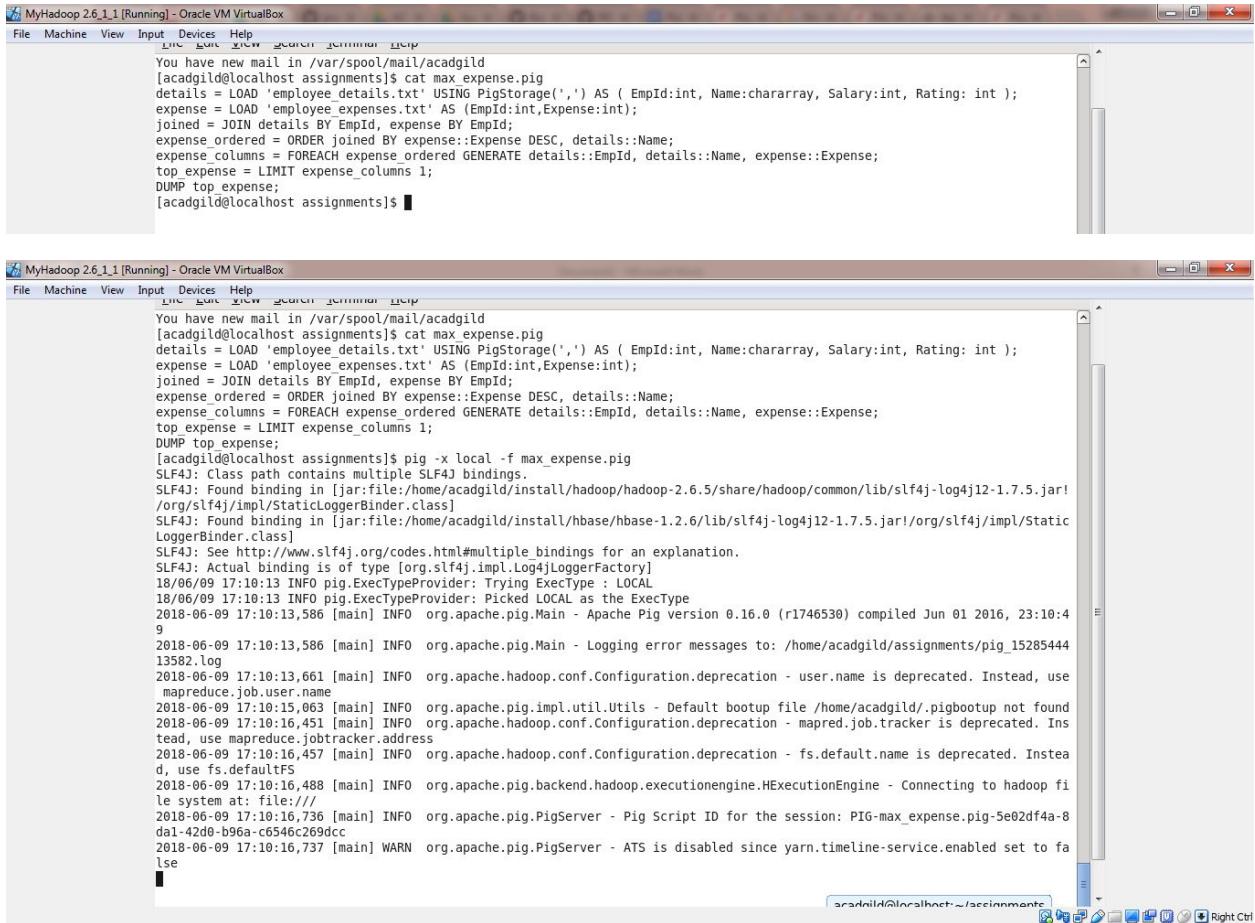
```

[acadgild@localhost assignments]$ 
2018-06-09 15:42:47,190 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 15:42:47,193 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 15:42:47,203 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 15:42:47,221 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 15:42:47,225 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 15:42:47,232 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 15:42:47,249 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 15:42:47,253 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 15:42:47,255 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 15:42:47,285 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 15:42:47,287 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 15:42:47,297 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 15:42:47,309 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLauncher - Success !
2018-06-09 15:42:47,324 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-09 15:42:47,324 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-09 15:42:47,358 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-09 15:42:47,358 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(101,Amitabh,20000,1)
(107,Salman,17500,2)
(103,Akshay,11000,3)
2018-06-09 15:42:47,479 [main] INFO org.apache.pig.Main - Pig script completed in 18 seconds and 716 milliseconds (18716 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$ 

```

(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)

Load input files using LOAD command. Join both files using common field employee Id. Sort the list on descending order of expenses and ascending order of name. Filter columns employee id , name and expense from the above result. TO get the employee with maximum expense fetch the first row from result.



The image shows two terminal windows side-by-side, both titled "MyHadoop 2.6.1\_1 [Running] - Oracle VM VirtualBox".

**Terminal Window 1:**

```
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$ cat max_expense.pig
details = LOAD 'employee details.txt' USING PigStorage(',') AS ( EmpId:int, Name:chararray, Salary:int, Rating: int );
expense = LOAD 'employee expenses.txt' AS ( EmpId:int,Expense:int );
joined = JOIN details BY EmpId, expense BY EmpId;
expense_ordered = ORDER joined BY expense::Expense DESC, details::Name;
expense_columns = FOREACH expense_ordered GENERATE details::EmpId, details::Name, expense::Expense;
top_expense = LIMIT expense_columns 1;
DUMP top_expense;
[acadgild@localhost assignments]$
```

**Terminal Window 2:**

```
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$ cat max_expense.pig
details = LOAD 'employee details.txt' USING PigStorage(',') AS ( EmpId:int, Name:chararray, Salary:int, Rating: int );
expense = LOAD 'employee expenses.txt' AS ( EmpId:int,Expense:int );
joined = JOIN details BY EmpId, expense BY EmpId;
expense_ordered = ORDER joined BY expense::Expense DESC, details::Name;
expense_columns = FOREACH expense_ordered GENERATE details::EmpId, details::Name, expense::Expense;
top_expense = LIMIT expense_columns 1;
DUMP top_expense;
[acadgild@localhost assignments]$ pig -x local -f max_expense.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/
org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/
org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/06/09 17:10:13 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/06/09 17:10:13 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-06-09 17:10:13,586 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:4
9
2018-06-09 17:10:13,586 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/assignments/pig_15285444
13582.log
2018-06-09 17:10:13,661 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use
mapreduce.job.user.name
2018-06-09 17:10:15,063 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-06-09 17:10:16,451 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2018-06-09 17:10:16,457 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-09 17:10:16,488 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
le system at: file:///
2018-06-09 17:10:16,736 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIQ-max_expense.pig-5e02df4a-8
da1-42d0-b96a-c6546c269dcc
2018-06-09 17:10:16,737 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to fa
lse
```

```

2018-06-09 17:10:46,252 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:10:46,253 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:10:46,265 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:10:46,310 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:10:46,317 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:10:46,317 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:10:46,345 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:10:46,350 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:10:46,367 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:10:46,409 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:10:46,421 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:10:46,462 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-09 17:10:46,500 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-09 17:10:46,500 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(110,Priyanka,400)
2018-06-09 17:10:46,654 [main] INFO org.apache.pig.Main - Pig script completed in 37 seconds and 329 milliseconds (37329 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$ 

```

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

Load input files using LOAD command. Join both files using common field employee Id. Filter columns employee id , name and expense from the above result.

```

[acadgild@localhost assignments]$ ls
1gb.txt          max.expense.pig    musicdata.txt   taskd.pig      TvSalesTask2.jar  wordmedian.txt
assignment1.txt  max-temp.txt       Person.java     television.txt TvSalesTask3.jar  wordSD.txt
create_one_gb_file.sh MusicDataTask1.jar Person.java     top3Salary.pig  wordcount.pig
employee.details.txt MusicDataTask2.jar Person.new.java top5Rating.pig word-count.txt
employee.expenses.txt MusicDataTask3.jar pig_count.txt TvSalesTask1.jar  wordmean.txt
[acadgild@localhost assignments]$ cat taskd.pig
details = LOAD 'employee.details.txt' USING PigStorage(',') AS ( EmpId:int, Name:chararray, Salary:int, Rating: int );
expense = LOAD 'employee.expenses.txt' AS (EmpId:int,Expense:int);
joined = JOIN details BY EmpId, expense BY EmpId;
expense_columns = FOREACH joined GENERATE details::EmpId, details::Name;
Result = DISTINCT expense_columns;
DUMP Result;
[acadgild@localhost assignments]$ 

```

```

MyHadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Sat Jun 9, 5:21 PM Acadgild
acadgild@localhost:~/assignments
File Edit View Search Terminal Help
assignment1.txt max-temp.txt Person_3.java television.txt TvSalesTask3.jar wordSD.txt
create_one_gb_file.sh MusicDataTask1.jar Person.java top3Salary.pig wordcount.pig
employee_details.txt MusicDataTask2.jar Person_new.java topRating.pig word-count.txt
employee_expenses.txt MusicDataTask3.jar pig_count.txt TvSalesTask1.jar wordmean.txt
[acadgild@localhost assignments]$ cat taskd.pig
details = LOAD 'employee_details.txt' USING PigStorage(',') AS ( EmpId:int, Name:chararray, Salary:int, Rating: int );
expense = LOAD 'employee_expenses.txt' AS ( EmpId:int,Expense:int );
joined = JOIN details BY EmpId, expense BY EmpId;
expense_columns = FOREACH joined GENERATE details::EmpId, details::Name;
Result = DISTINCT expense_columns;
DUMP Result;
[acadgild@localhost assignments]$ pig -x local -f taskd.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar! /org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar! /org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/06/09 17:21:22 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/06/09 17:21:22 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-06-09 17:21:22,726 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-06-09 17:21:22,726 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/assignments/pig_1528545082723.log
2018-06-09 17:21:22,767 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2018-06-09 17:21:23,253 [main] INFO org.apache.pig.impl.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-06-09 17:21:23,631 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-06-09 17:21:23,631 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-09 17:21:23,634 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///2018-06-09 17:21:23,712 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: DTG-taskd.nin-hAf0729-7a12-4A
[acadgild@localhost assignments]$ Right Ctrl

MyHadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
SplitDecide Memory Manager Spill Counter : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local367527205_0001 -> job_local961446326_0002,
job_local961446326_0002

2018-06-09 17:21:32,238 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:21:32,244 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:21:32,247 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:21:32,290 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:21:32,299 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:21:32,305 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 17:21:32,324 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success !
2018-06-09 17:21:32,338 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-09 17:21:32,342 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-09 17:21:32,384 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-09 17:21:32,384 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Amitabh)
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
2018-06-09 17:21:32,513 [main] INFO org.apache.pig.Main - Pig script completed in 11 seconds and 403 milliseconds (11403 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$ Right Ctrl

```

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

Load input files using LOAD command. Right outer Join both files using common field employee Id. From the joined results remove rows which has expense filed null. Filter distinct columns employee id , name .

MyHadoop 2.6.1\_1 [Running] - Oracle VM VirtualBox

```
File Machine View Input Devices Help
Applications Places System Sat Jun 9, 6:09 PM Acadgild
File Edit View Search Terminal Help
acadgild@localhost:~/assignments
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$ cat taske.pig
details = LOAD 'employee_details.txt' USING PigStorage(',') AS ( EmpId:int, Name:chararray, Salary:int, Rating: int );
expense = LOAD 'employee_expenses.txt' AS ( EmpId:int,Expense:int );
joined = JOIN expense BY EmpId RIGHT OUTER, details BY EmpId;
filtered = FILTER joined BY expense::Expense is null;
no_expense = FOREACH filtered GENERATE details::EmpId, details::Name;
result = DISTINCT no_expense;
DUMP result;
[acadgild@localhost assignments]$
```

MyHadoop 2.6.1\_1 [Running] - Oracle VM VirtualBox

```
File Machine View Input Devices Help
Applications Places System Sat Jun 9, 6:10 PM Acadgild
File Edit View Search Terminal Help
acadgild@localhost:~/assignments
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$ cat taske.pig
details = LOAD 'employee_details.txt' USING PigStorage(',') AS ( EmpId:int, Name:chararray, Salary:int, Rating: int );
expense = LOAD 'employee_expenses.txt' AS ( EmpId:int,Expense:int );
joined = JOIN expense BY EmpId RIGHT OUTER, details BY EmpId;
filtered = FILTER joined BY expense::Expense is null;
no_expense = FOREACH filtered GENERATE details::EmpId, details::Name;
result = DISTINCT no_expense;
DUMP result;
[acadgild@localhost assignments]$ pig -x local taske.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [/jar:/file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/_org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [/jar:/file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/_org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/06/09 18:10:28 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/06/09 18:10:28 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-06-09 18:10:28,741 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-06-09 18:10:28,743 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/assignments/pig_1528548028738.log
2018-06-09 18:10:28,809 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2018-06-09 18:10:30,076 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
```

MyHadoop 2.6.1\_1 [Running] - Oracle VM VirtualBox

```
File Machine View Input Devices Help
Total records productively splitted.
Job DAG:
job local715085489 0001 -> job_local632535421_0002,
job_local632535421_0002

2018-06-09 18:10:54,731 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 18:10:54,740 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 18:10:54,756 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 18:10:54,847 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 18:10:54,852 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 18:10:54,857 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 18:10:54,895 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success !
2018-06-09 18:10:54,931 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-09 18:10:54,939 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-09 18:10:55,002 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-09 18:10:55,003 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(103,Akshay)
(106,Aamir)
(107,Salman)
(108,Ranbir)
(109,Katrina)
(111,Tushar)
(112,Ajay)
(113,Jubeen)
2018-06-09 18:10:55,223 [main] INFO org.apache.pig.Main - Pig script completed in 28 seconds and 635 milliseconds (28635 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$
```

### Task 3

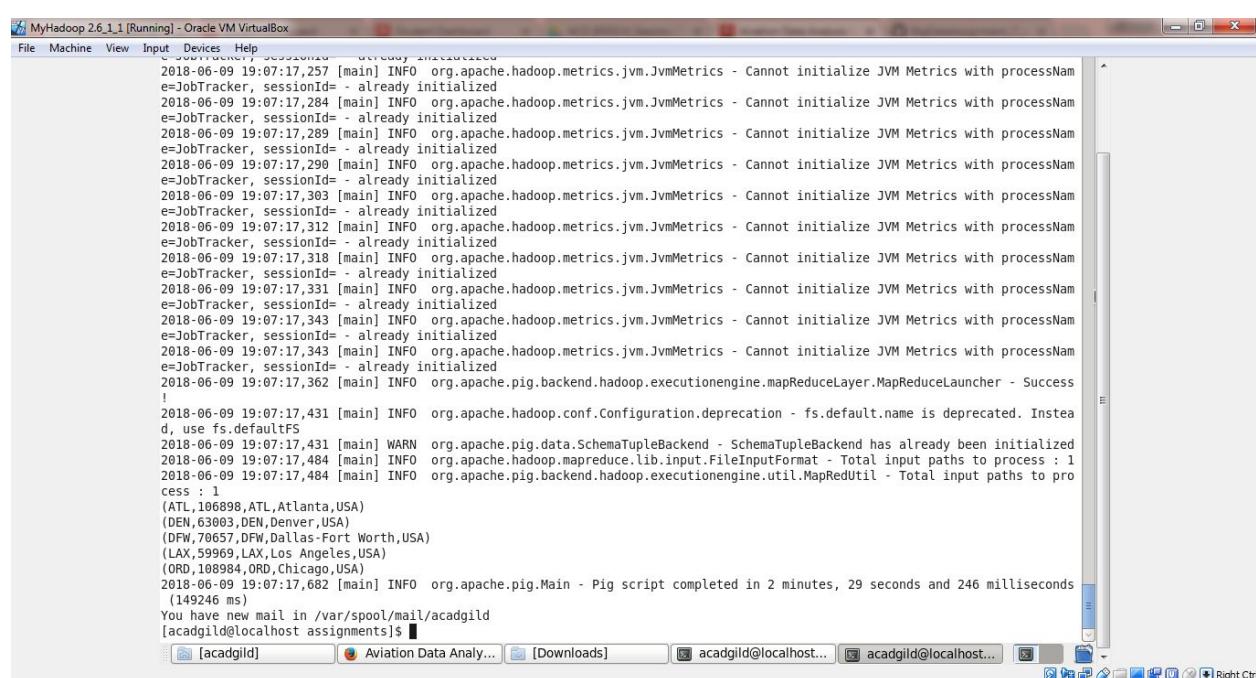
Implement the use case present in below blog link and share the complete steps along with screenshot(s) from your end.

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



The terminal window shows the execution of a Pig Latin script named task3.pig. The script reads two CSV files: 'DelayedFlights.csv' and 'airports.csv', joins them, and then groups by destination to find the top 5 most visited places. The output is a list of destination names and their counts.

```
[acadgild@localhost assignments]$ cat task3.pig
A = load '/home/acadgild/assignments/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18 as dest;
C = filter B by dest is not null;
D = group C by dest;
E = foreach D generate group, COUNT(C.dest);
F = order E by $1 DESC;
Result = LIMIT F 5;
A1 = load '/home/acadgild/assignments/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
joined_table = join Result by $0, A2 by dest;
dump joined_table;
[acadgild@localhost assignments]$
```

The terminal window shows the logs of a Hadoop job. It displays multiple INFO messages from the org.apache.hadoop.metrics.jvm.JvmMetrics class, indicating that JVM Metrics cannot be initialized because processName is already initialized. It also shows INFO messages from org.apache.hadoop.mapreduce.lib.input.FileInputFormat and org.apache.hadoop.mapreduce.util.MapRedUtil, and a success message from org.apache.pig.backend.hadoop.executionengine.mapReduceLauncher.

```
2018-06-09 19:07:17,257 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 19:07:17,284 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 19:07:17,289 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 19:07:17,290 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 19:07:17,303 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 19:07:17,312 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 19:07:17,318 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 19:07:17,331 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 19:07:17,343 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 19:07:17,343 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-06-09 19:07:17,362 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLauncher - Success !
2018-06-09 19:07:17,431 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-09 19:07:17,431 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-09 19:07:17,484 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-09 19:07:17,484 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
( ATL,106898,ATL,Atlanta,USA)
(DEN,63003,DEN,Denver,USA)
(DFW,70657,DFW,Dallas-Fort Worth,USA)
(LAX,59969,LAX,Los Angeles,USA)
(ORD,108984,ORD,Chicago,USA)
2018-06-09 19:07:17,682 [main] INFO  org.apache.pig.Main - Pig script completed in 2 minutes, 29 seconds and 246 milliseconds (14924 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$
```

### Problem Statement 2

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

```

[acadgild@localhost assignments]$ cat task3.2.pig
A = load '/home/acadgild/assignments/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTI
LINE','UNIX','SKIP_INPUT_HEADER');
B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
C = filter B by cancelled == 1 AND cancel_code == 'B';
D = group C by month;
E = foreach D generate group, COUNT(C.cancelled);
F= order E by $1 DESC;
Result = limit F 1;
dump Result;
[acadgild@localhost assignments]$ 

```

```

2018-06-09 19:14:33,701 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,707 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,714 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,720 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,726 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,733 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,739 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,745 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,751 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,757 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,763 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,769 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,775 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,781 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,821 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,824 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:14:33,842 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLauncher - Success !
2018-06-09 19:14:33,853 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-09 19:14:33,853 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-09 19:14:33,896 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-09 19:14:33,897 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(12,250)
2018-06-09 19:14:34,047 [main] INFO org.apache.pig.Main - Pig script completed in 1 minute, 32 seconds and 651 milliseconds

```

### Problem Statement 3

Top ten origins with the highest AVG departure delay

```

[acadgild@localhost assignments]$ cat task3.3.pig
A = load '/home/acadgild/assignments/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTI
LINE','UNIX','SKIP_INPUT_HEADER');
B1 = foreach A generate (int)$16 as dep_delay, (chararray)$17 as origin;
C1 = filter B1 by (dep_delay is not null) AND (origin is not null);
D1 = group C1 by origin;
E1 = foreach D1 generate group, AVG(C1.dep_delay);
Result = order E1 by $1 DESC;
Top_ten = limit Result 10;
Lookup = load '/home/acadgild/assignments/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTI
LINE','UNIX','SKIP_INPUT_HEADER');
Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;
Joined = join Lookup1 by origin, Top_ten by $0;
Final = foreach Joined generate $0,$1,$2,$4;
Final Result = ORDER Final by $3 DESC;
dump Final.Result;
[acadgild@localhost assignments]$ 

```

```

MyHadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
2018-06-09 19:28:28,077 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
eJobTracker, sessionId= - already initialized
2018-06-09 19:28:28,881 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
eJobTracker, sessionId= - already initialized
2018-06-09 19:28:28,891 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
eJobTracker, sessionId= - already initialized
2018-06-09 19:28:28,892 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
eJobTracker, sessionId= - already initialized
2018-06-09 19:28:28,903 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
eJobTracker, sessionId= - already initialized
2018-06-09 19:28:28,910 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
eJobTracker, sessionId= - already initialized
2018-06-09 19:28:28,910 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
eJobTracker, sessionId= - already initialized
2018-06-09 19:28:28,916 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
eJobTracker, sessionId= - already initialized
2018-06-09 19:28:28,920 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLauncher - Success !
2018-06-09 19:28:28,939 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-09 19:28:28,940 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-09 19:28:28,973 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-09 19:28:28,973 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(CMX, Hancock,USA, 116.1470588235294)
(PLN,Pellston,USA, 93.76190476190476)
(SPI,Springfield,USA, 83.84873949579831)
(ALO,Waterloo,USA, 82.2258064516129)
(MOT,NA,USA, 79.55665824630542)
(ACY,Atlantic City,USA, 79.3103448275862)
(MOT,Minot,USA, 78.66165413533835)
(HHH,NA,USA, 76.53005464480874)
(EGE,Eagle,USA, 74.12891986062718)
(BGN,Binghamton,USA, 73.15533980582525)
2018-06-09 19:28:29,103 [main] INFO org.apache.pig.Main - Pig script completed in 1 minute, 13 seconds and 805 milliseconds
(73805 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$ 

```

## Problem Statement 4

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

```

MyHadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Terminal Help
acadgild@localhost:~/assignments$ cat task3.pig
A = load '/home/acadgild/assignments/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER');
B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;
C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);
D = GROUP C by (origin,dest);
E = FOREACH D generate group, COUNT(C.diversion);
F = ORDER E BY $1 DESC;
Result = limit F 10;
dump Result;
[acadgild@localhost assignments]$ 

```

MyHadoop 2.6.1\_1 [Running] - Oracle VM VirtualBox

```
File Machine View Input Devices Help
2018-06-09 19:33:06,229 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-06-09 19:33:06,238 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-06-09 19:33:06,261 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-06-09 19:33:06,269 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-06-09 19:33:06,272 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-06-09 19:33:06,301 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-06-09 19:33:06,308 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-06-09 19:33:06,310 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-06-09 19:33:06,328 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLauncher - Success !
2018-06-09 19:33:06,344 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-06-09 19:33:06,344 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-09 19:33:06,399 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-09 19:33:06,399 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
((ORD,LGA),39)
((DAL,HOU),35)
((DFW,LGA),33)
((ATL,LGA),32)
((ORD,SNA),31)
((SLC,SUN),31)
((MIA,LGA),31)
((BUR,JFK),29)
((HRL,HOU),28)
((BUR,DFW),25)
2018-06-09 19:33:06,557 [main] INFO org.apache.pig.Main - Pig script completed in 1 minute, 10 seconds and 398 milliseconds (70398 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$
```