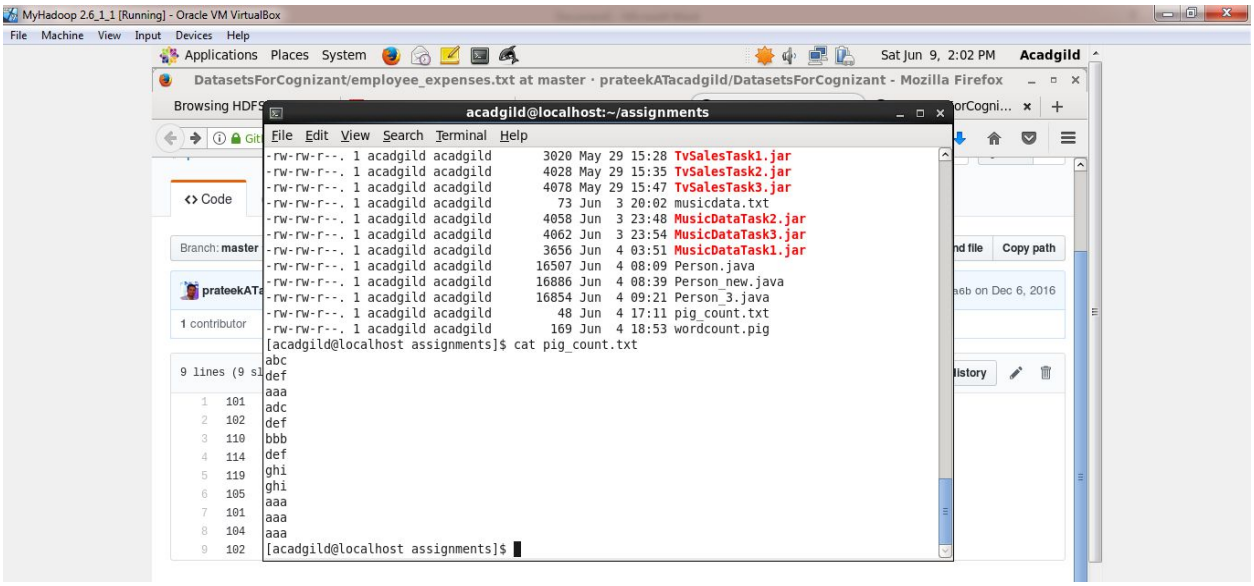
Task 1

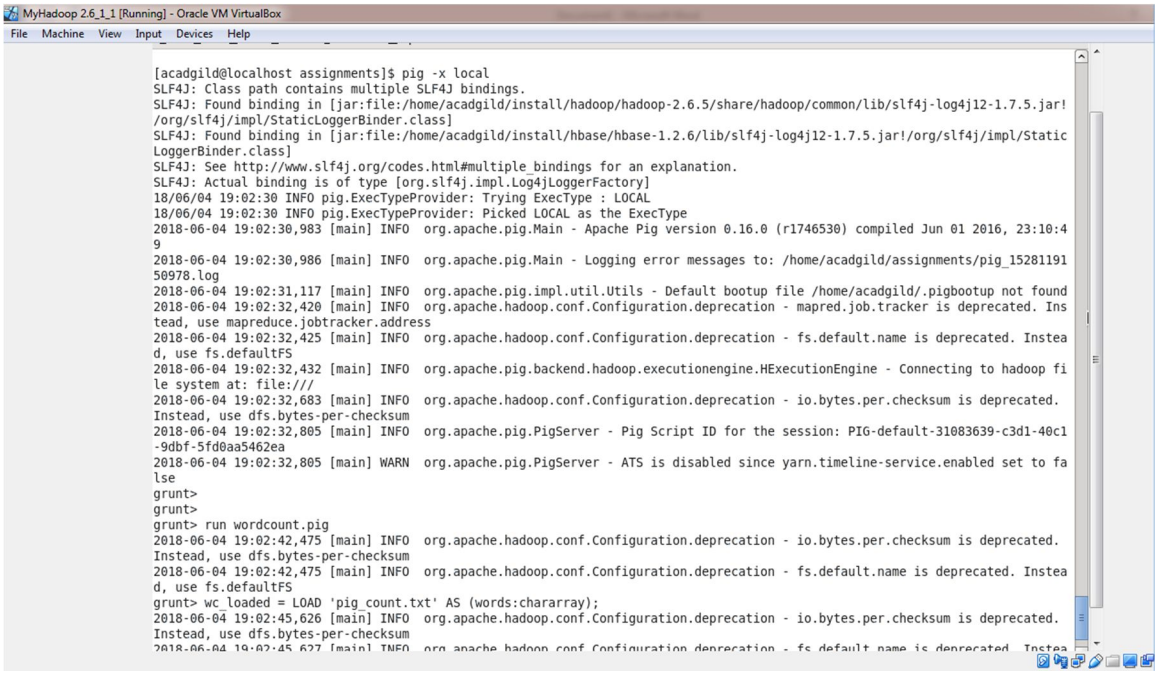
Write a program to implement wordcount using Pig.

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.

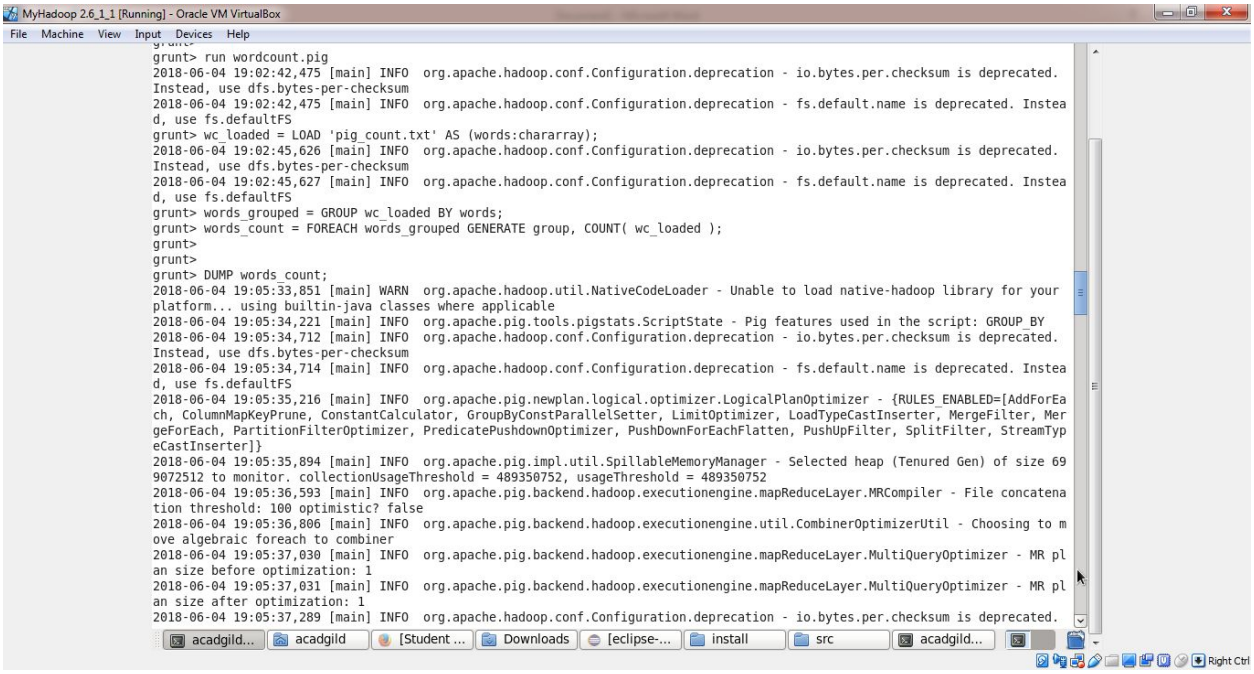
Input file:

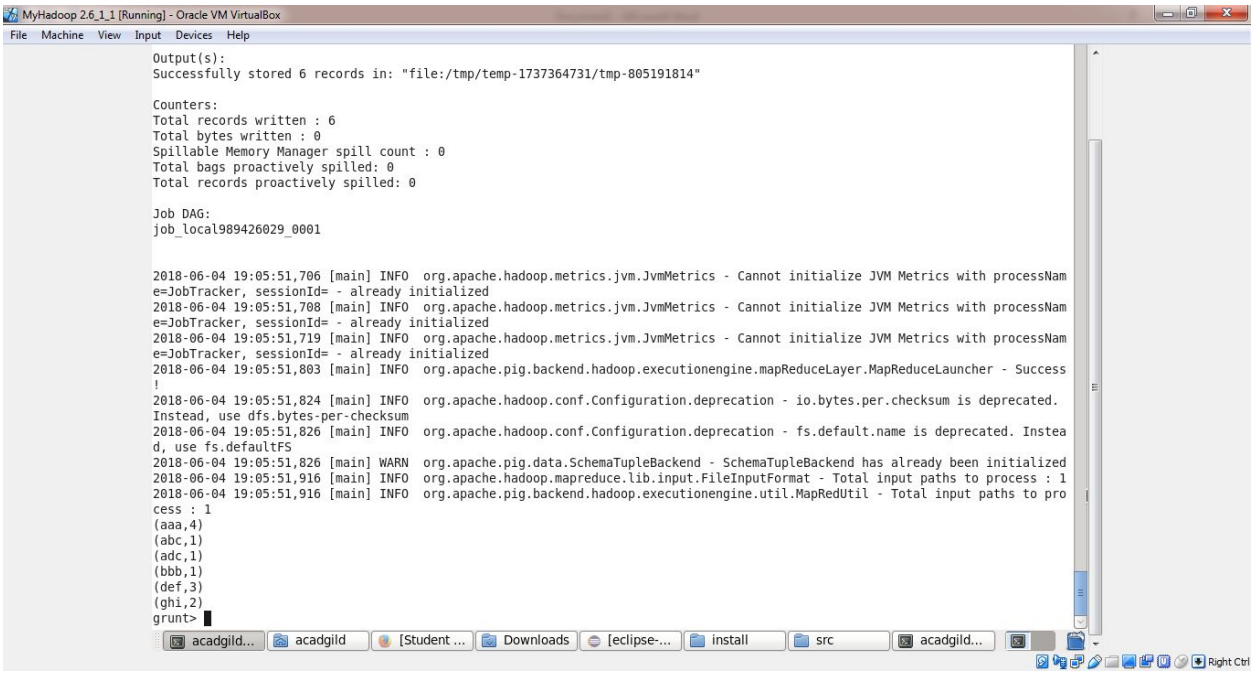


Connect to pig grunt shell using command – **pig –x local** and run pig script using run **wordcount.pig**



Check output using DUMP command





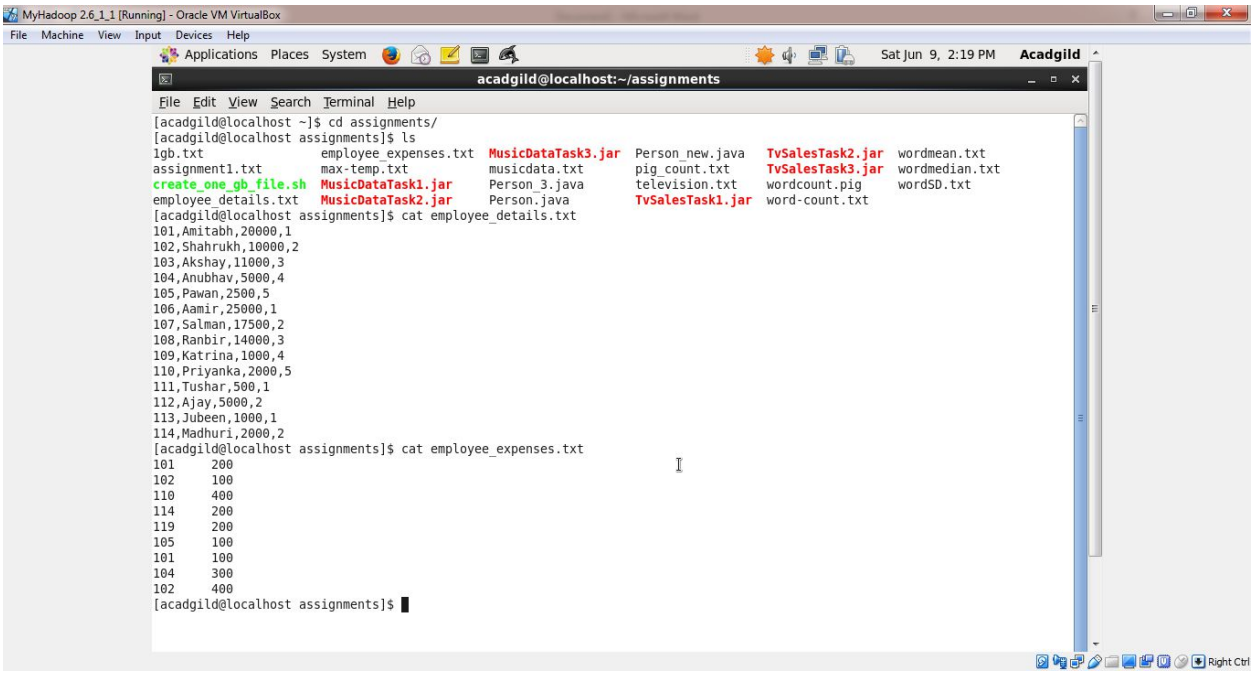
Task 2

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

https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee_details.txt

employee_expenses(EmpID,Expence)

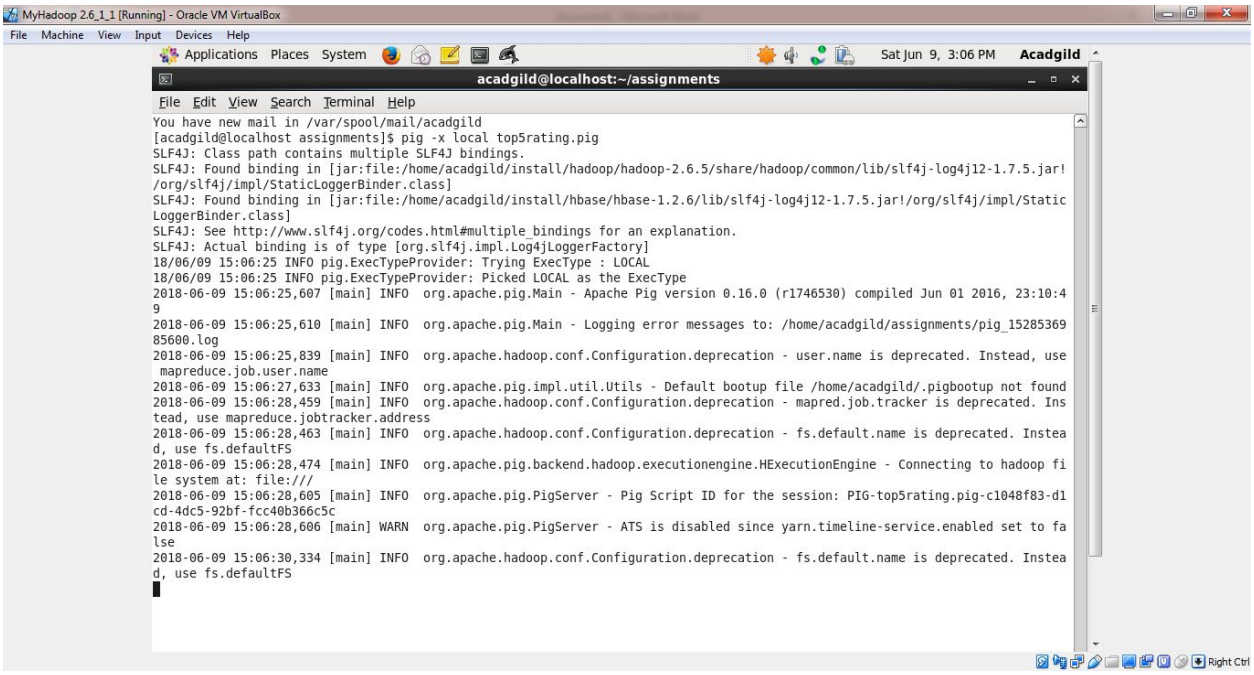
https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee_expenses.txt



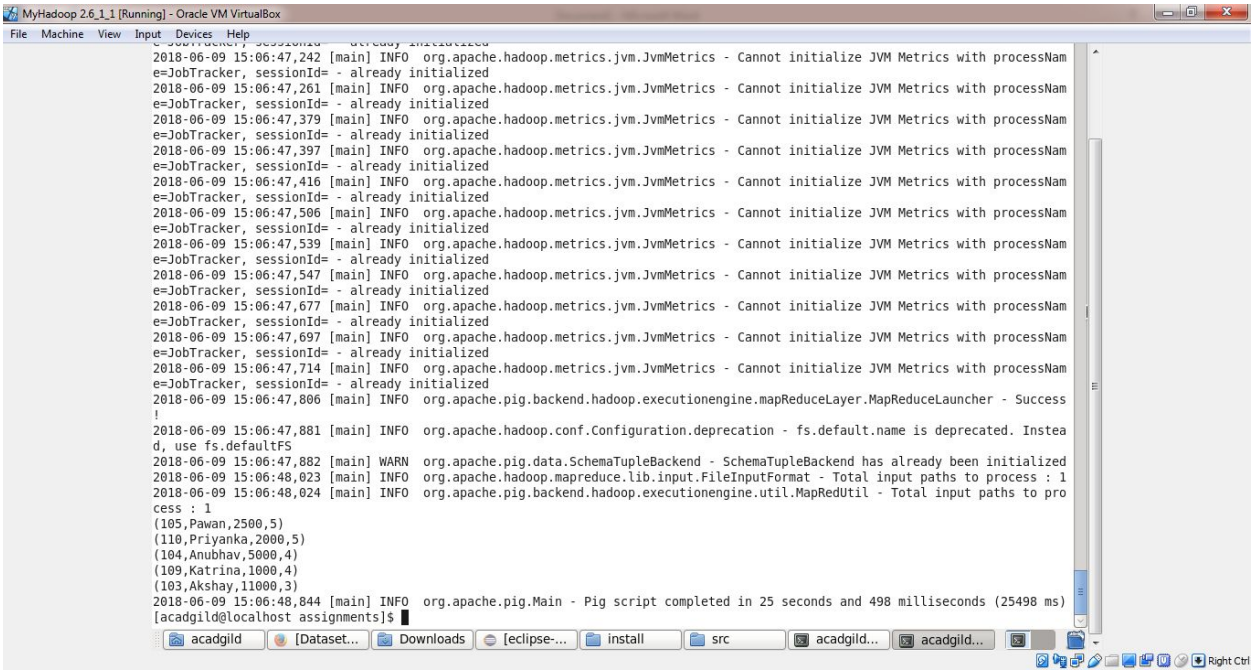
(a) Top 5 employees (employee id and employee name) with highest rating. (In case two employees have same rating, employee with name coming first in dictionary should get preference)

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:

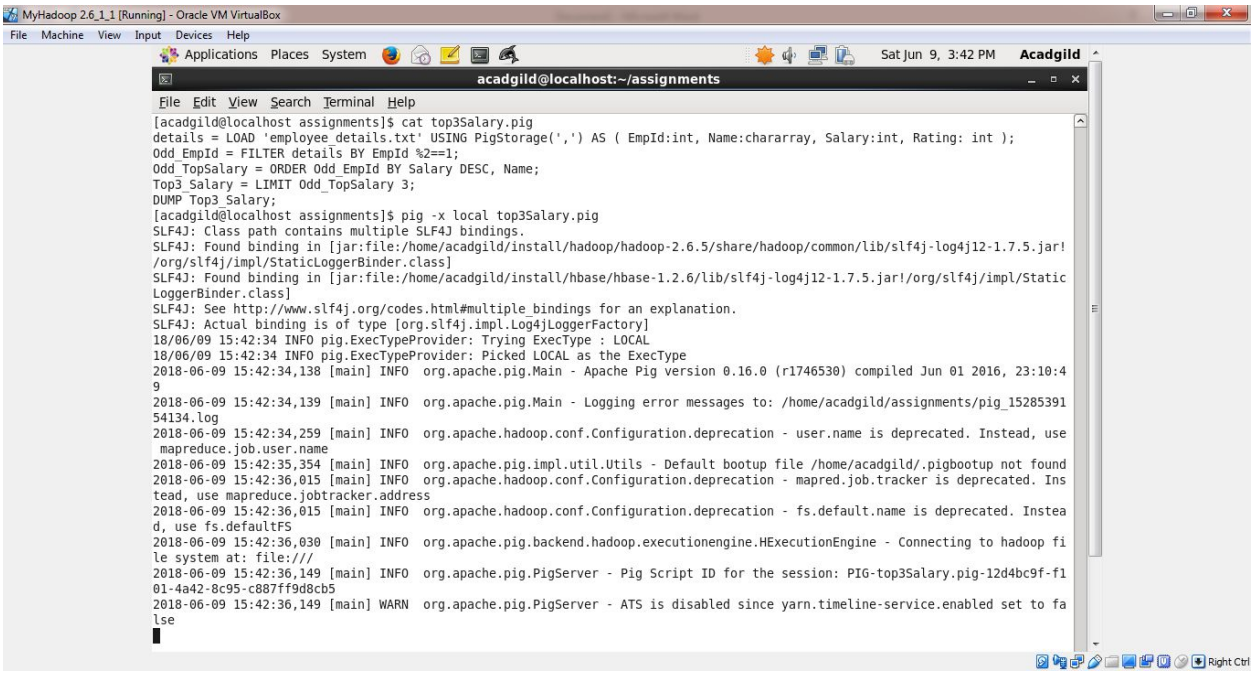


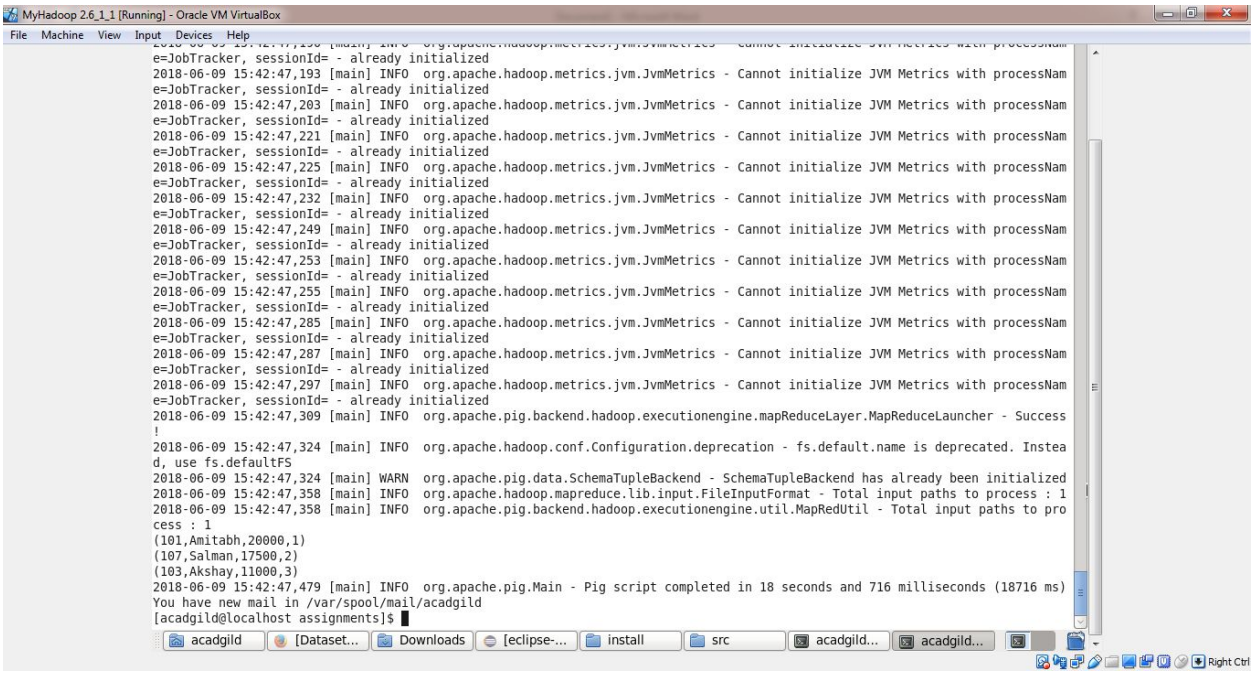
Output:



(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)

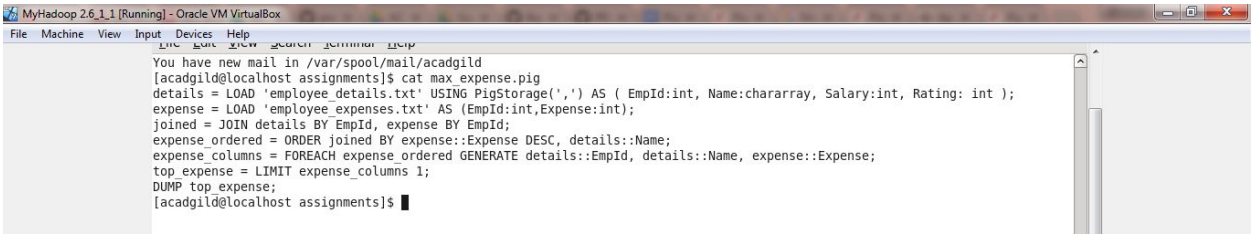
Load input file using LOAD command. Filter odd records by using modulus function EmpId %2 == 1. Sort the list on descending order of salary and ascending order of name. Fetch top 3 records to get output.

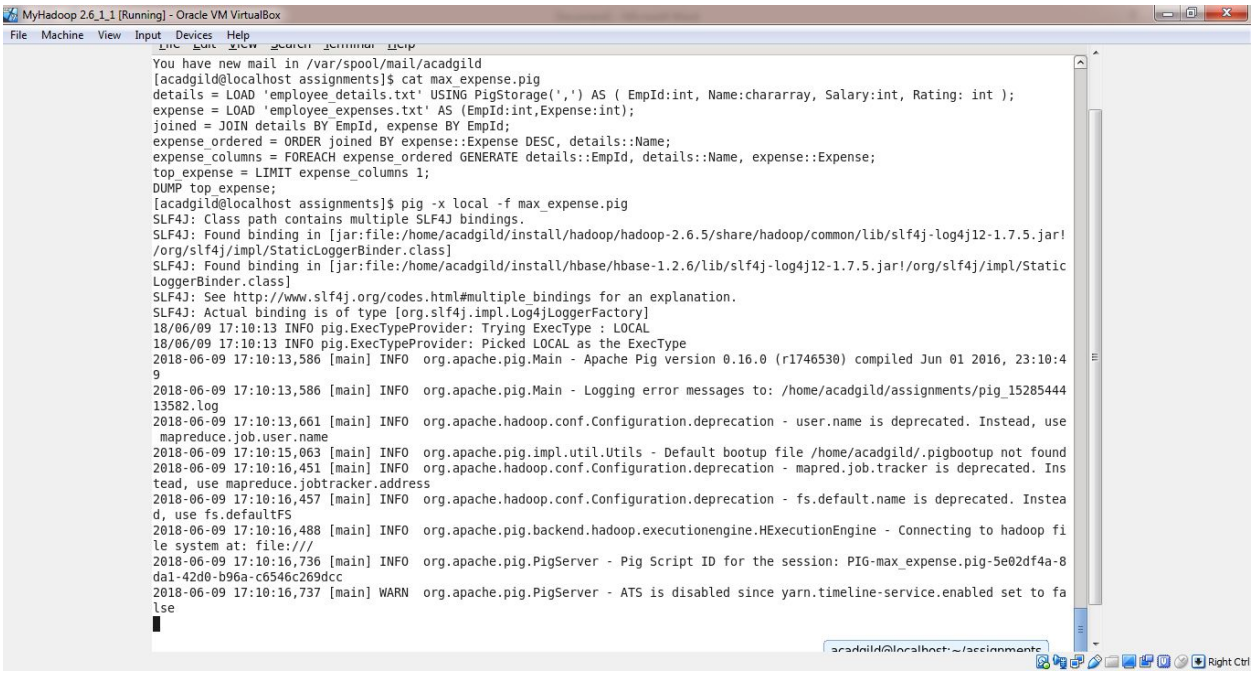


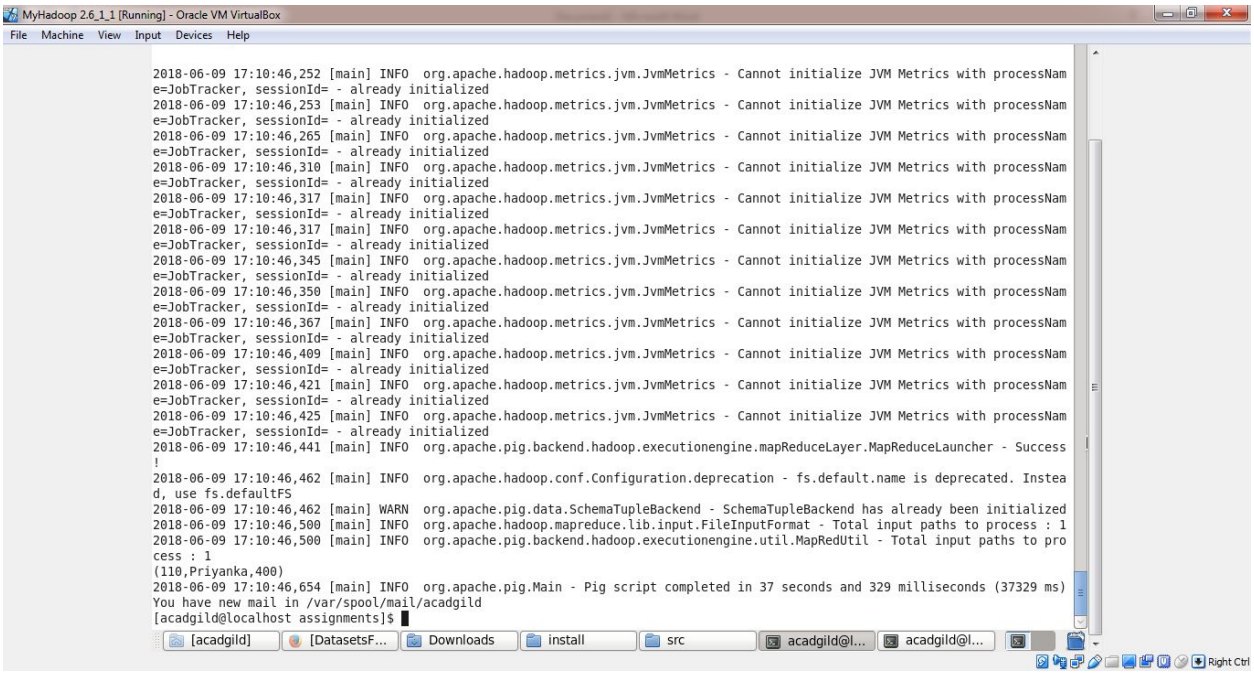


(c) Employee (employee id and employee name) with maximum expense (In case two employees have same expense, employee with name coming first in dictionary should get preference)

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.

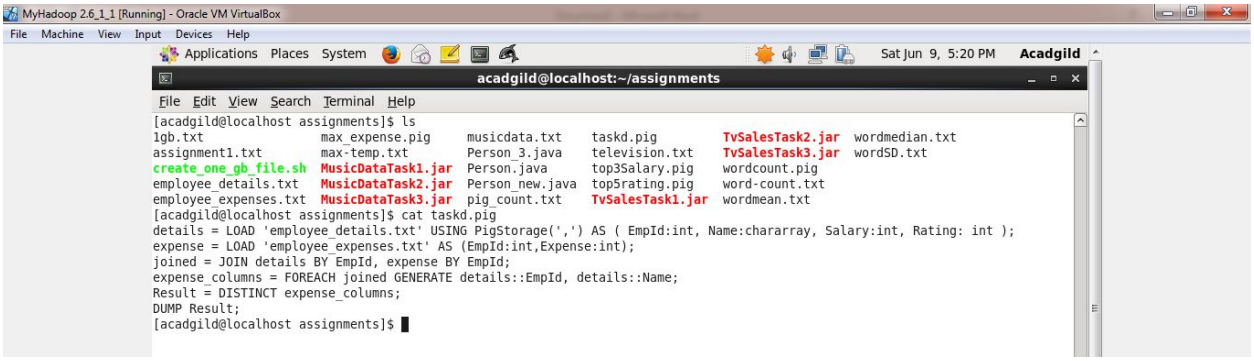


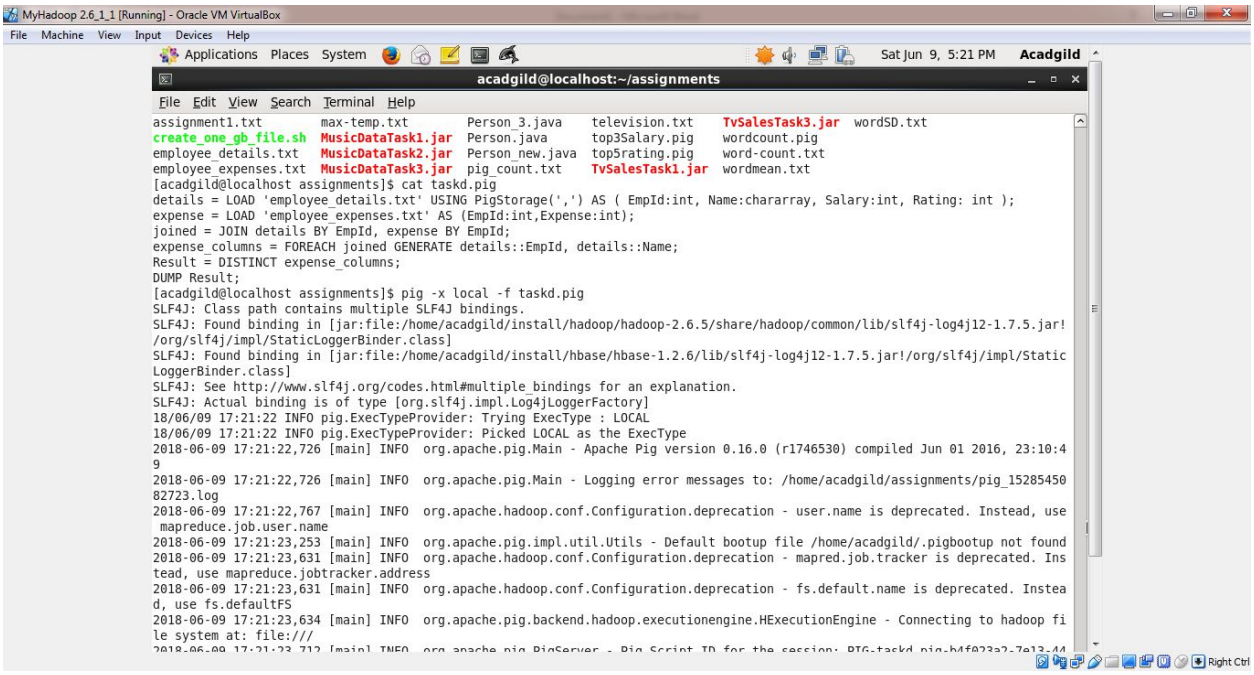


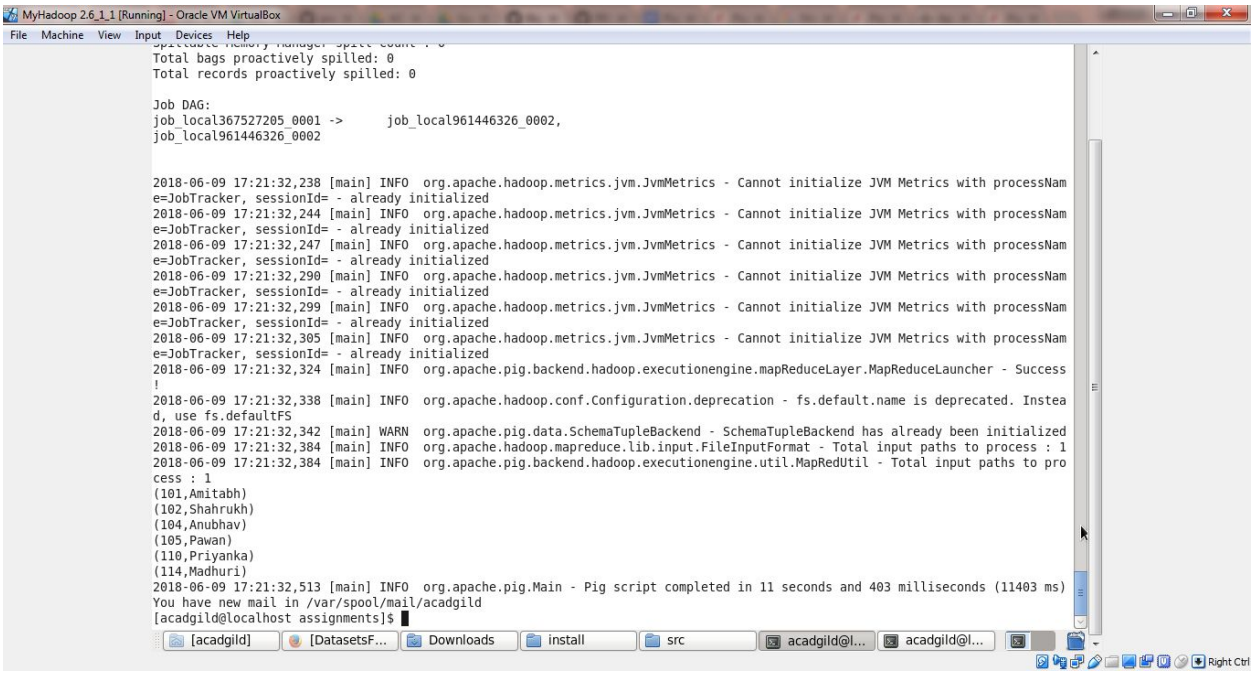


(d) List of employees (employee id and employee name) having entries in employee_expenses file.

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.

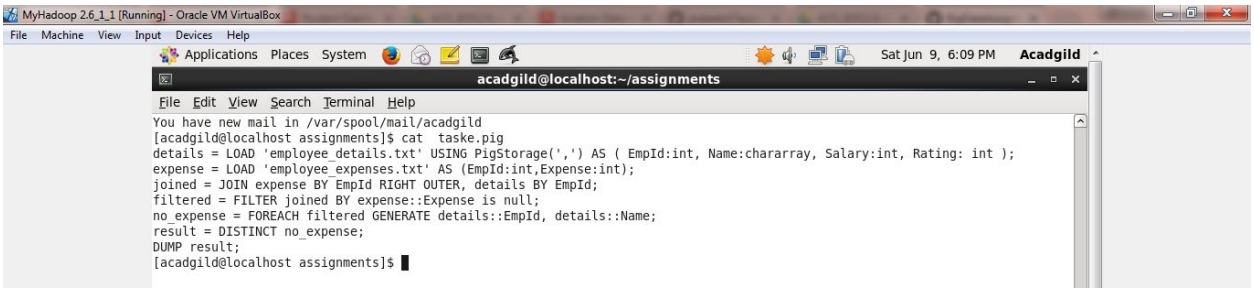


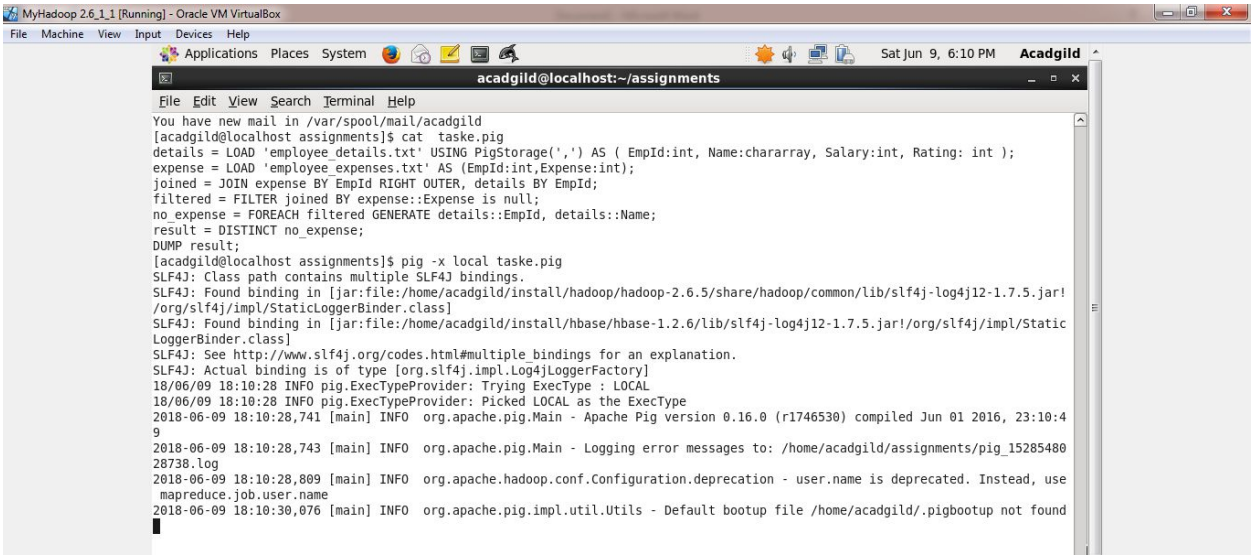


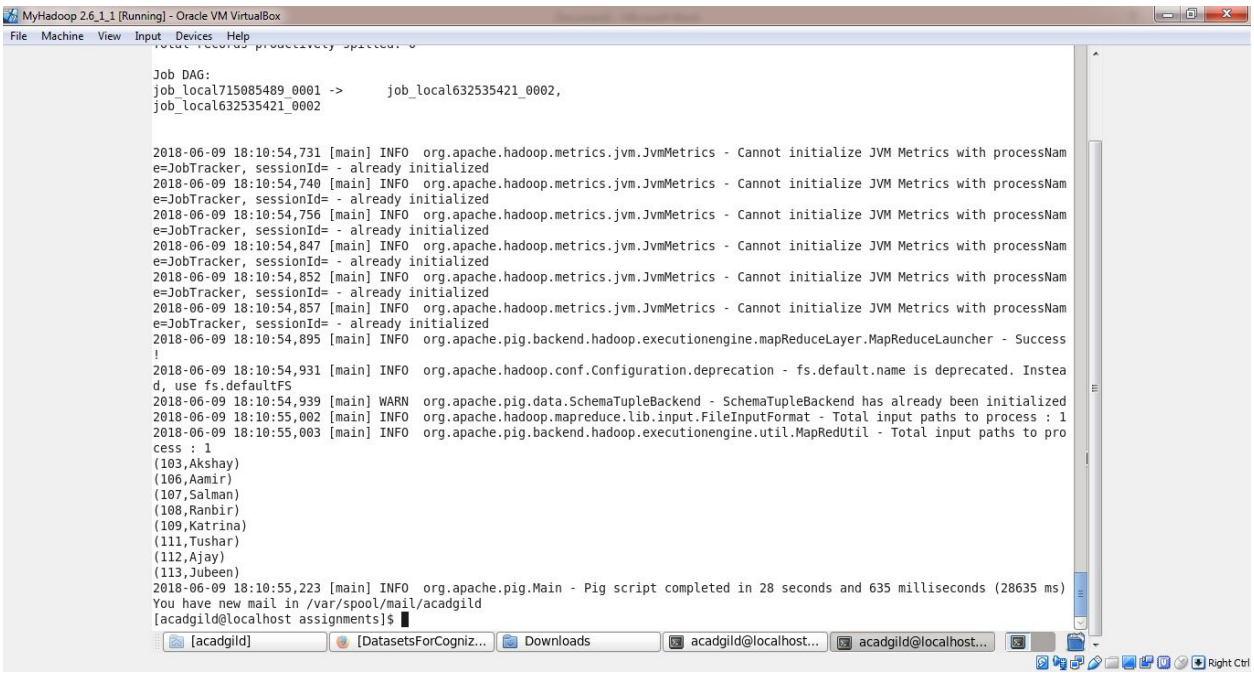


(e) List of employees (employee id and employee name) having no entry in employee_expenses file.

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
acadgild@localhost:~/assignments
File Edit View Search Terminal Help
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]$

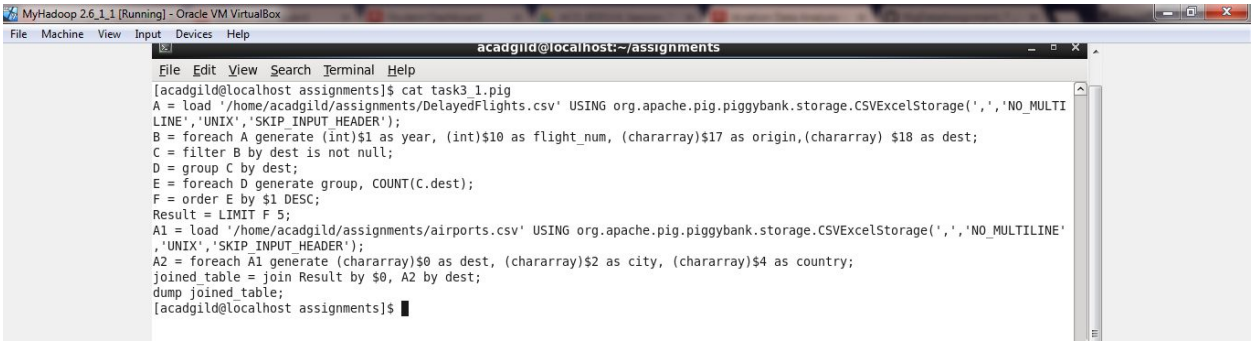

MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jun 9, 6:10 PM
Acadgild
acadgild@localhost:~/assignments
File Edit View Search Terminal Help
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost assignments]$ cat taske.pig
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

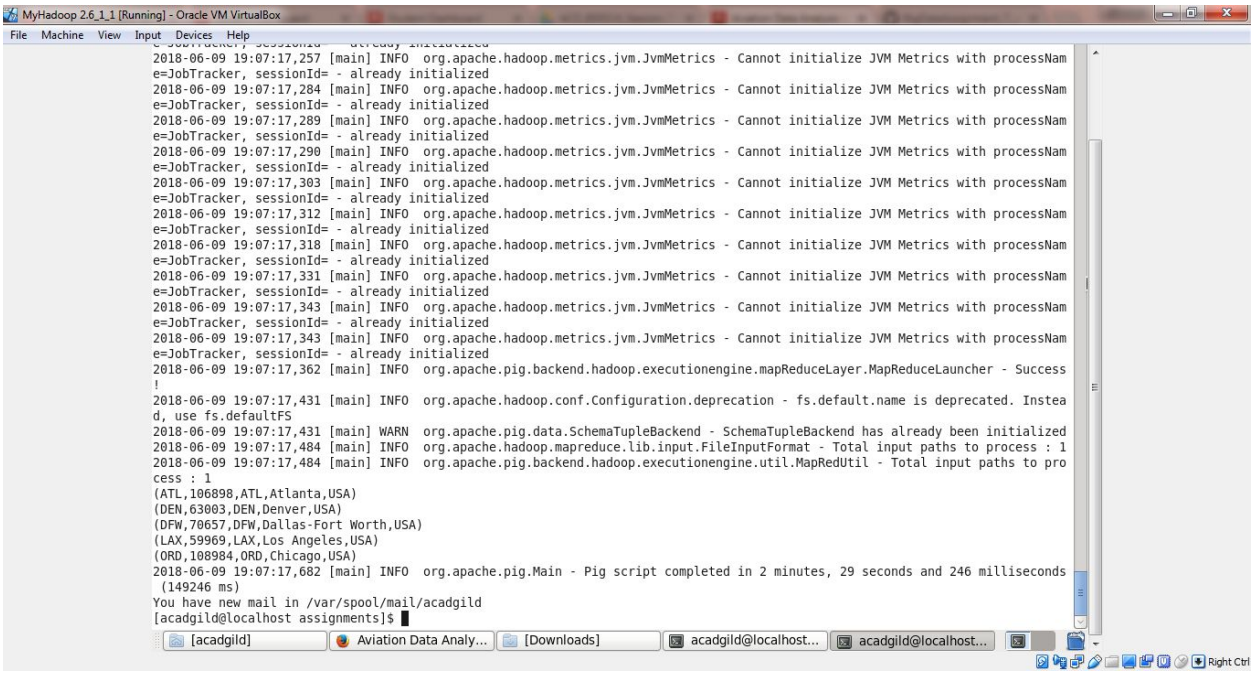

MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Job DAG:
job_local715085489_0001 -> job_local632535421_0002,
job_local632535421_0002
2018-06-09 18:10:54,731 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, 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]$
[acadgild]
[DatasetsForCogniz...
Downloads
acadgild@localhost...
acadgild@localhost...
Right Ctrl

Task 3

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

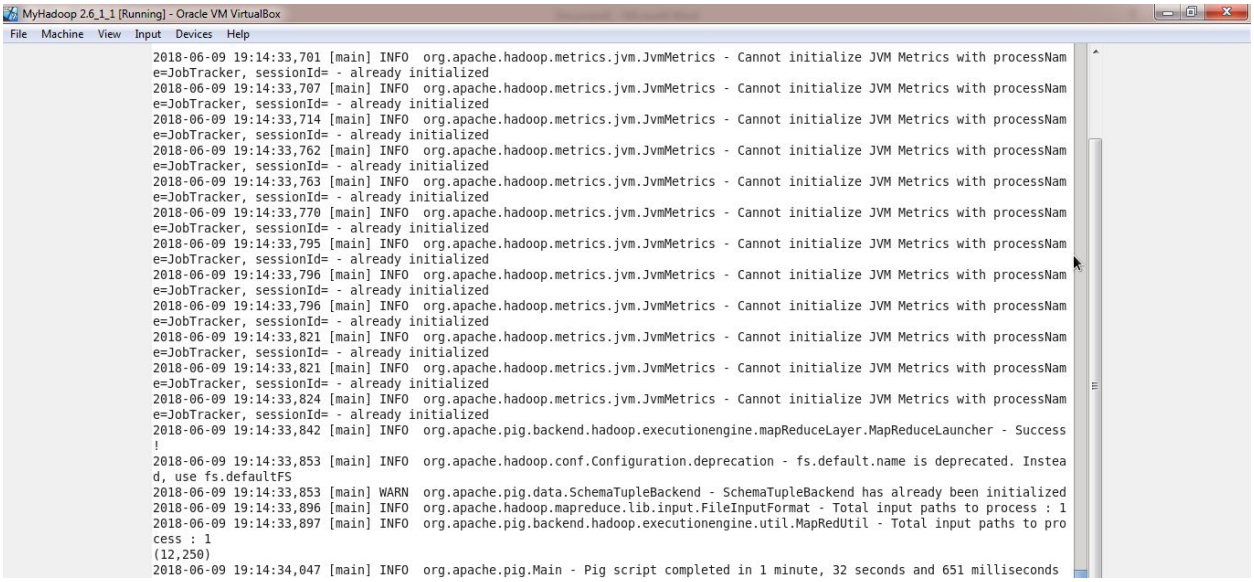
Problem Statement 1 : Find out the top 5 most visited destinations.





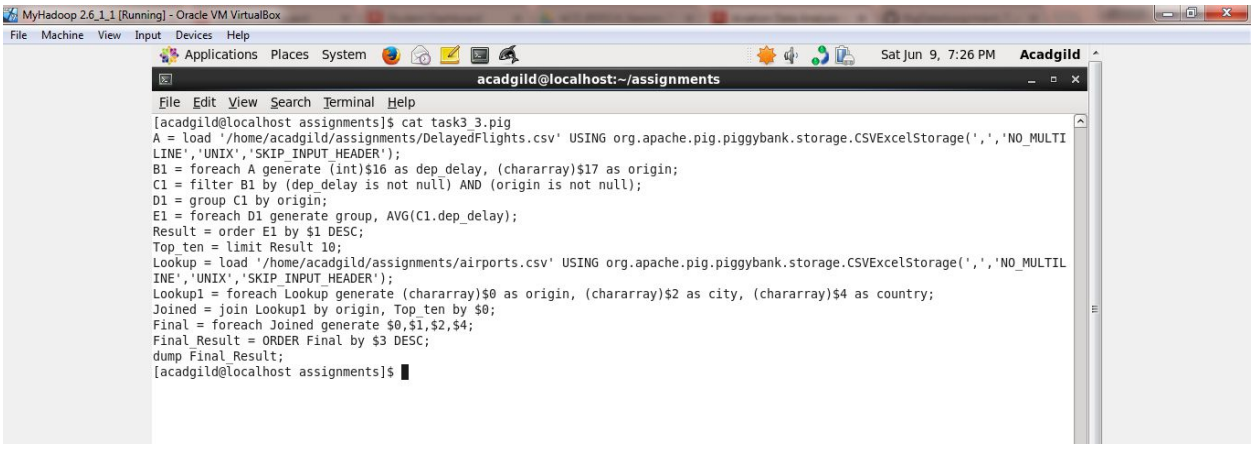
Problem Statement 2

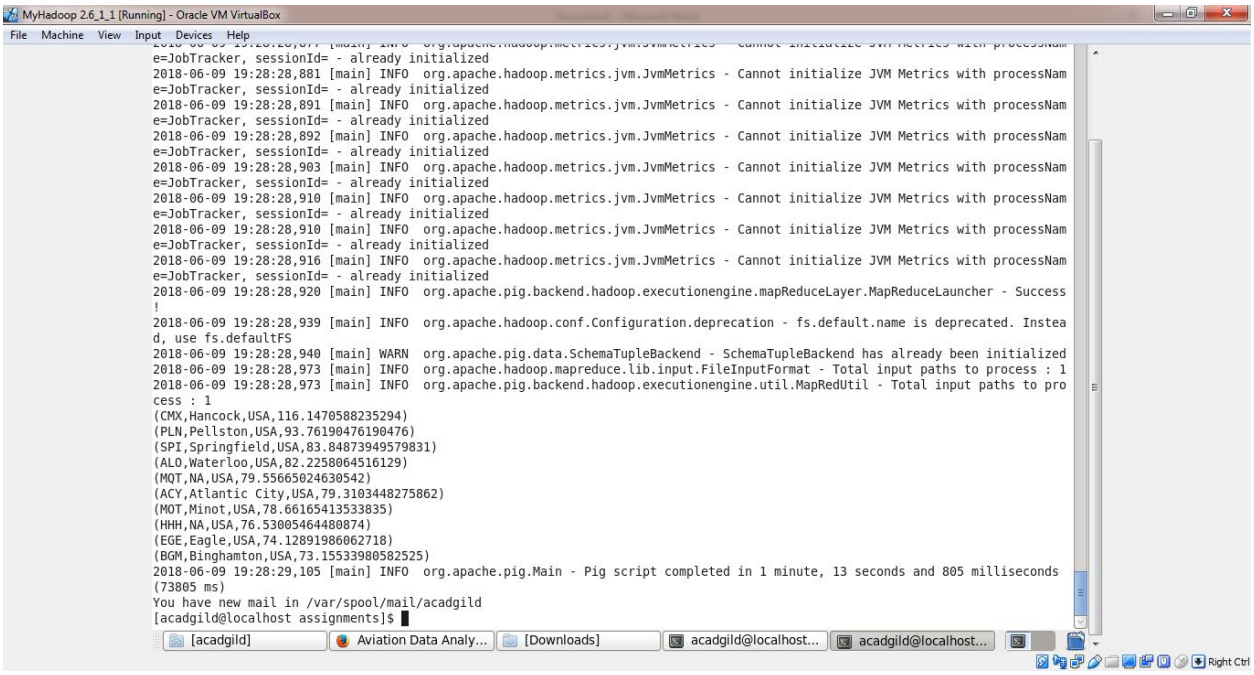
Which month has seen the most number of cancellations due to bad weather?



Problem Statement 3

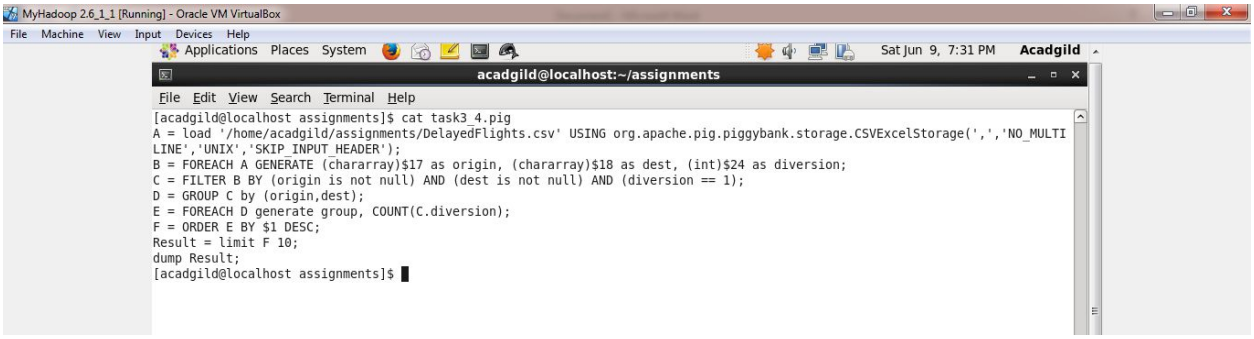
Top ten origins with the highest AVG departure delay

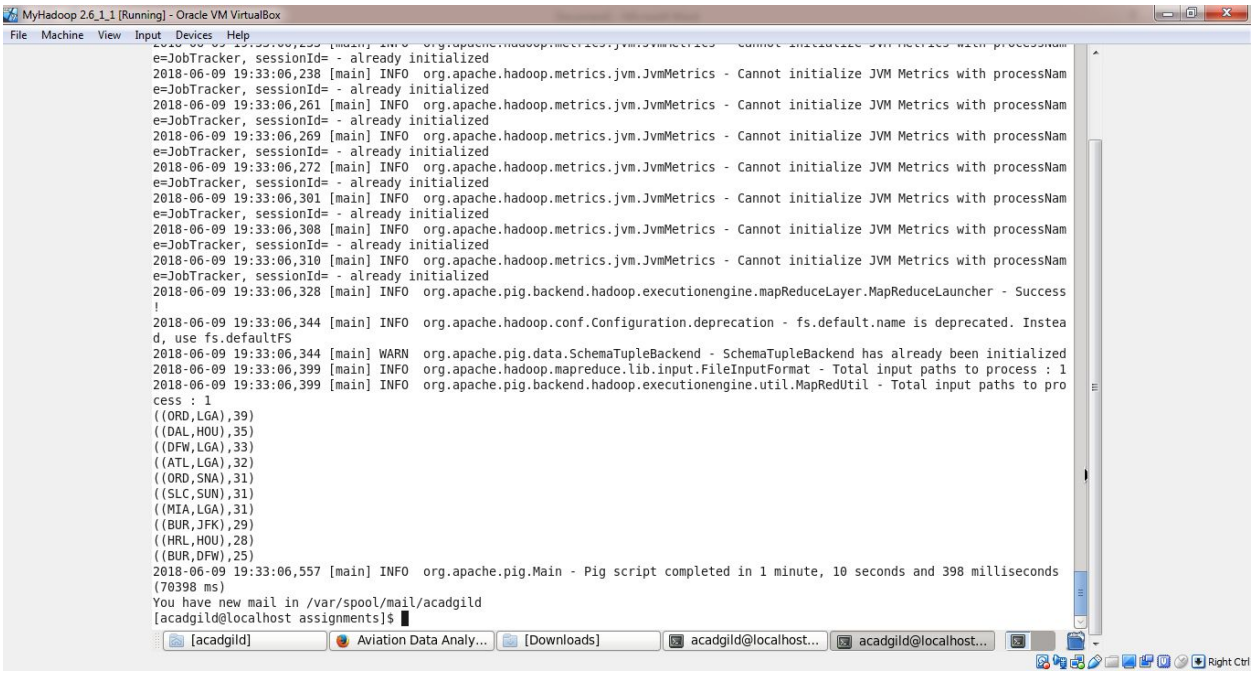




Problem Statement 4

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




MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
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 processNam
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 processNam
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 processNam
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 processNam
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 processNam
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 processNam
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 processNam
e=JobTracker, sessionId= - already initialized
2018-06-09 19:33:06,328 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-06-09 19:33:06,344 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, 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 pro
cess : 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]$
[acadgild]
Aviation Data Analy...
[Downloads]
acadgild@localhost...
acadgild@localhost...
Right Ctrl