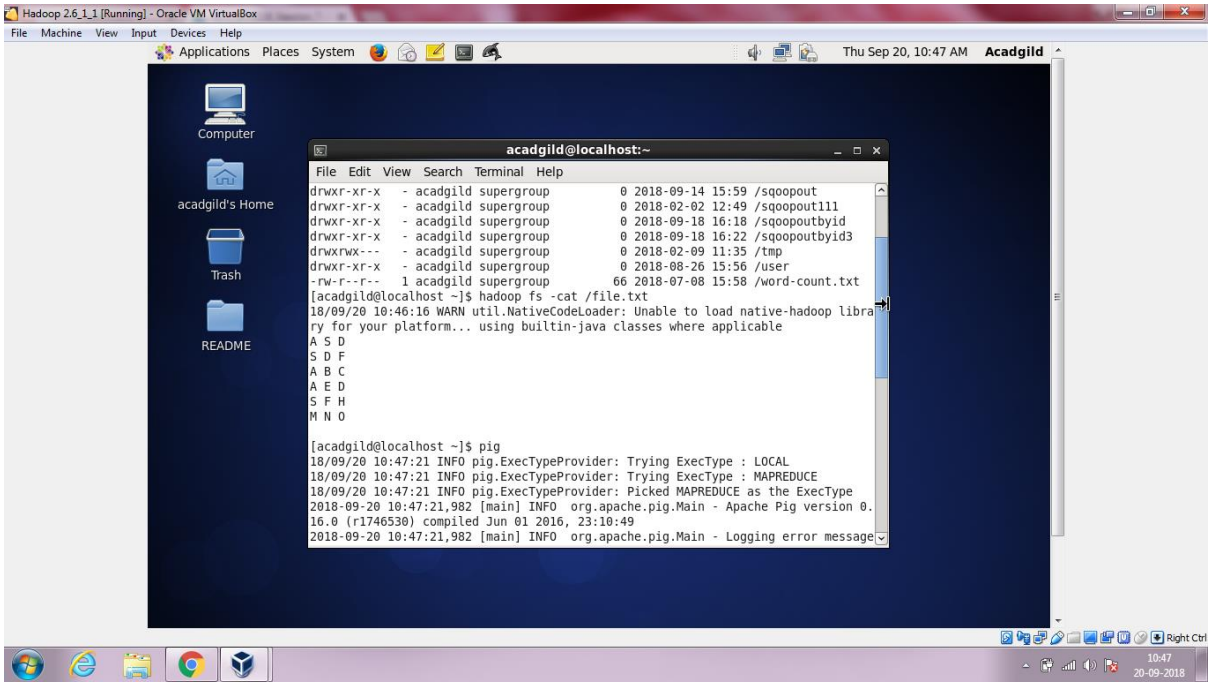
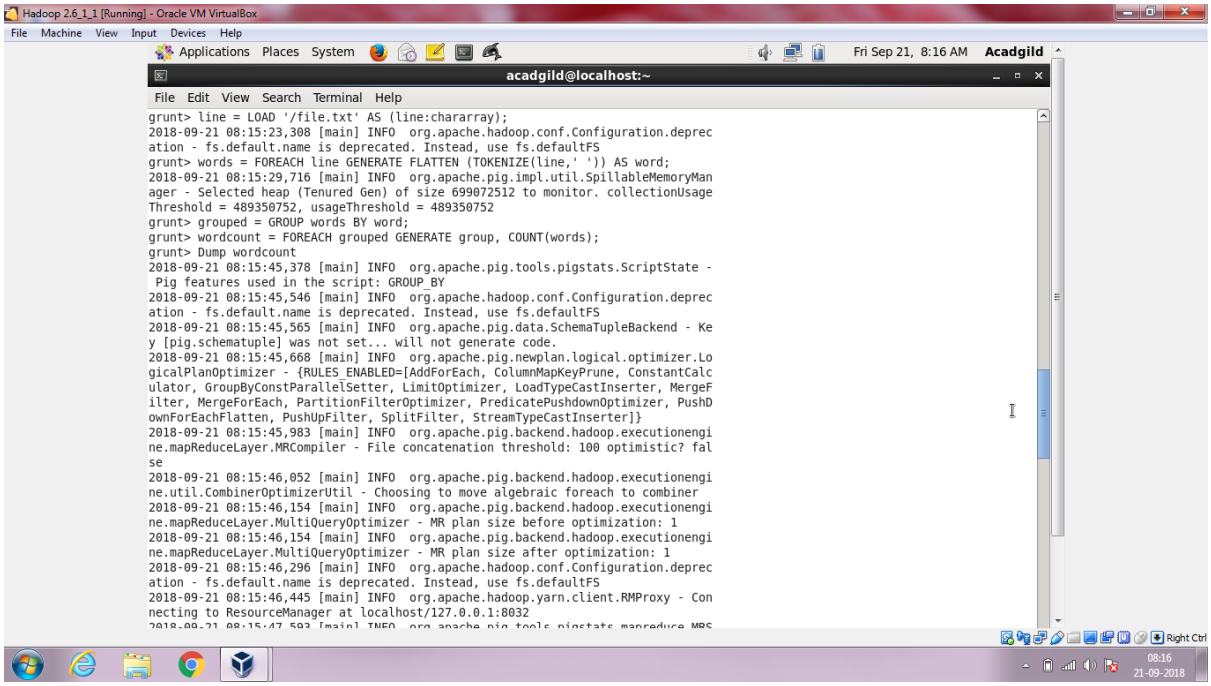
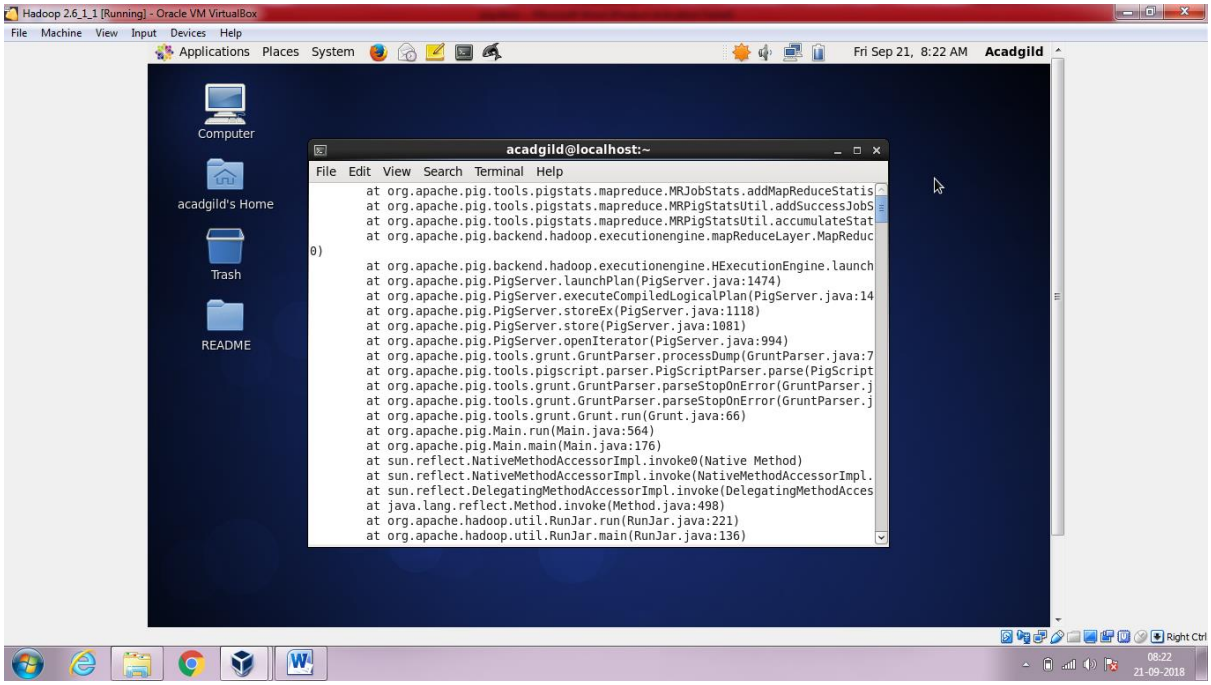
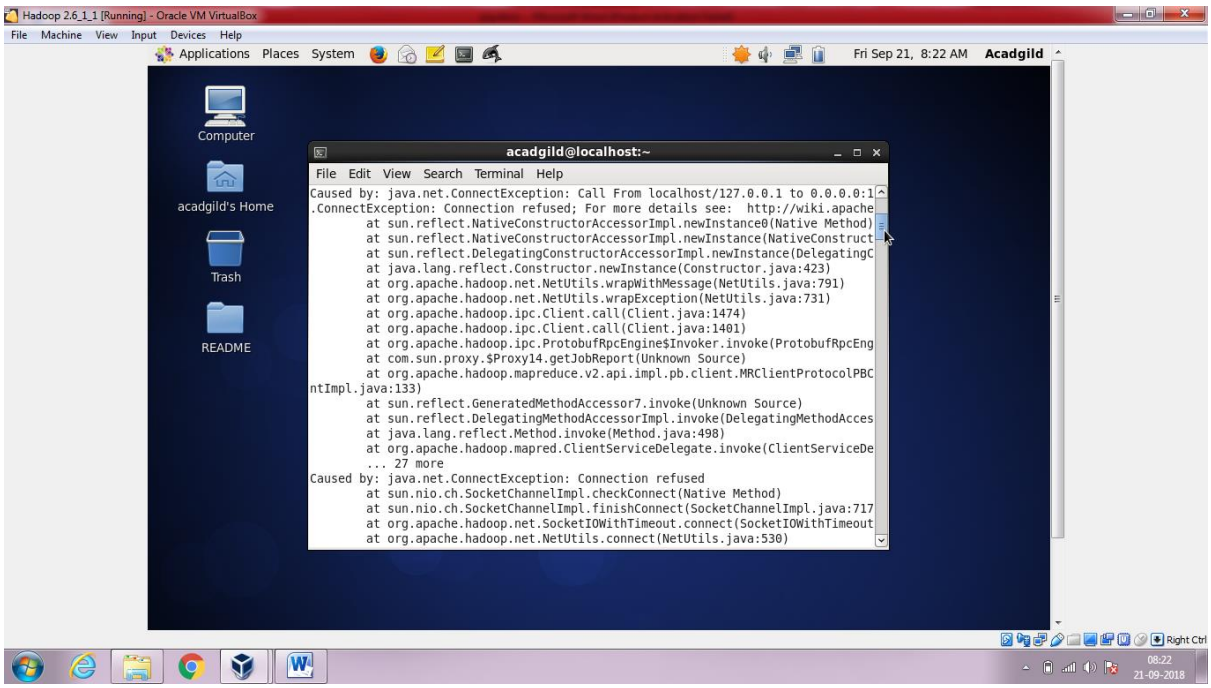
Task 1

Write program to implement word count using pig.

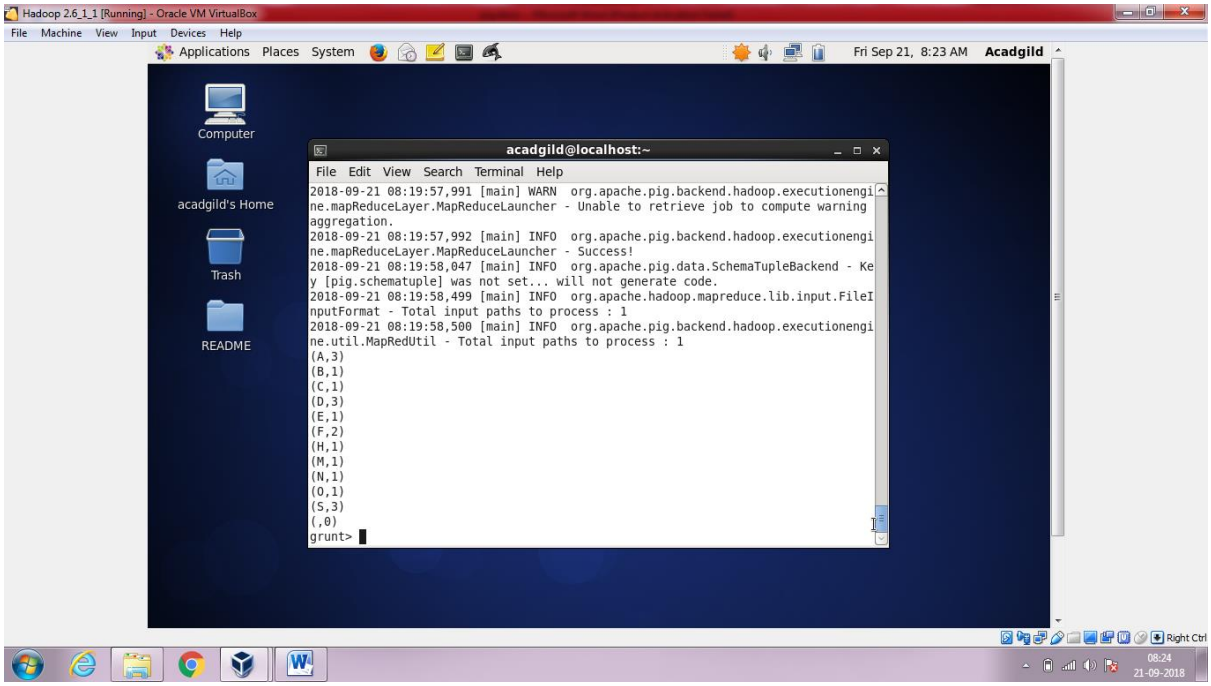


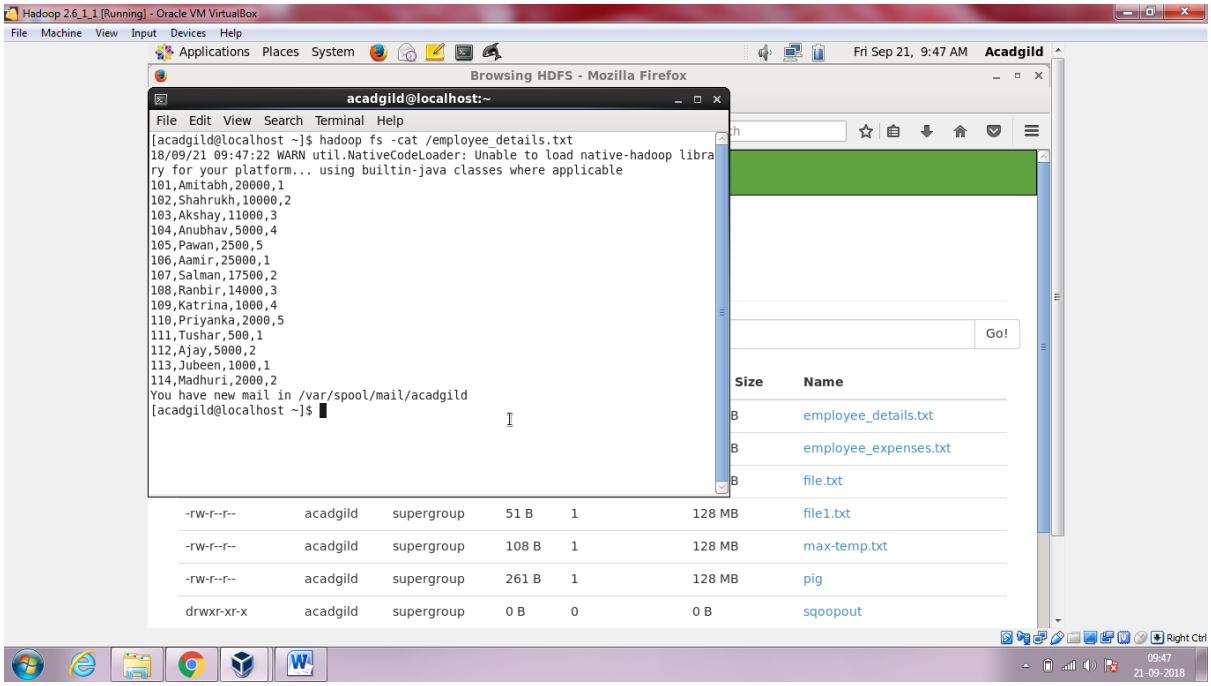

Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Fri Sep 21, 8:22 AM Acadgild
Computer
acadgild's Home
Trash
README
acadgild@localhost:~
File Edit View Search Terminal Help
at org.apache.pig.tools.pigstats.mapreduce.MRJobStats.addMapReduceStatis
at org.apache.pig.tools.pigstats.mapreduce.MRPigStatsUtil.addSuccessJobS
at org.apache.pig.tools.pigstats.mapreduce.MRPigStatsUtil.accumulateStat
at org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduc
0)
at org.apache.pig.backend.hadoop.executionengine.HExecutionEngine.launch
at org.apache.pig.PigServer.launchPlan(PigServer.java:1474)
at org.apache.pig.PigServer.executeCompiledLogicalPlan(PigServer.java:14
at org.apache.pig.PigServer.storeEx(PigServer.java:1118)
at org.apache.pig.PigServer.store(PigServer.java:1081)
at org.apache.pig.PigServer.openIterator(PigServer.java:994)
at org.apache.pig.tools.grunt.GruntParser.processDump(GruntParser.java:7
at org.apache.pig.tools.pigscript.parser.PigScriptParser.parse(PigScript
at org.apache.pig.tools.grunt.GruntParser.parseStopOnError(GruntParser.j
at org.apache.pig.tools.grunt.GruntParser.parseStopOnError(GruntParser.j
at org.apache.pig.tools.grunt.Grunt.run(Grunt.java:66)
at org.apache.pig.Main.run(Main.java:564)
at org.apache.pig.Main.main(Main.java:176)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAcces
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
Right Ctrl
08:22
21-09-2018

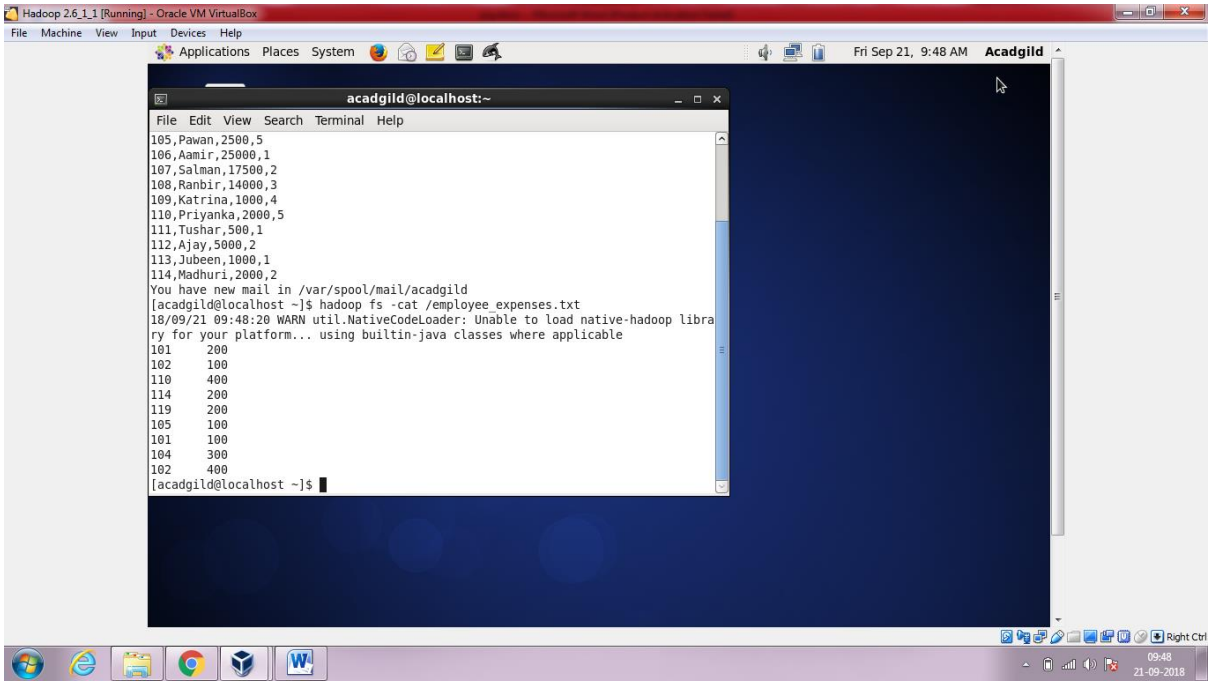

Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Fri Sep 21, 8:22 AM Acadgild
Computer
acadgild's Home
Trash
README
acadgild@localhost:~
File Edit View Search Terminal Help
Caused by: java.net.ConnectException: Call From localhost/127.0.0.1 to 0.0.0.0:1
.ConnectException: Connection refused; For more details see: http://wiki.apache
at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstruct
at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingC
at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
at org.apache.hadoop.net.NetUtils.wrapWithMessage(NetUtils.java:791)
at org.apache.hadoop.net.NetUtils.wrapException(NetUtils.java:731)
at org.apache.hadoop.ipc.Client.call(Client.java:1474)
at org.apache.hadoop.ipc.Client.call(Client.java:1401)
at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEng
at com.sun.proxy.$Proxy14.getJobReport(Unknown Source)
at org.apache.hadoop.mapreduce.v2.api.impl.pb.client.MRClientProtocolPBC
ntImpl.java:133)
at sun.reflect.GeneratedMethodAccessor7.invoke(Unknown Source)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAcces
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.mapred.ClientServiceDelegate.invoke(ClientServiceDe
... 27 more
Caused by: java.net.ConnectException: Connection refused
at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
at sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:717
at org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout
at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:530)
Right Ctrl
08:22
21-09-2018

Output:

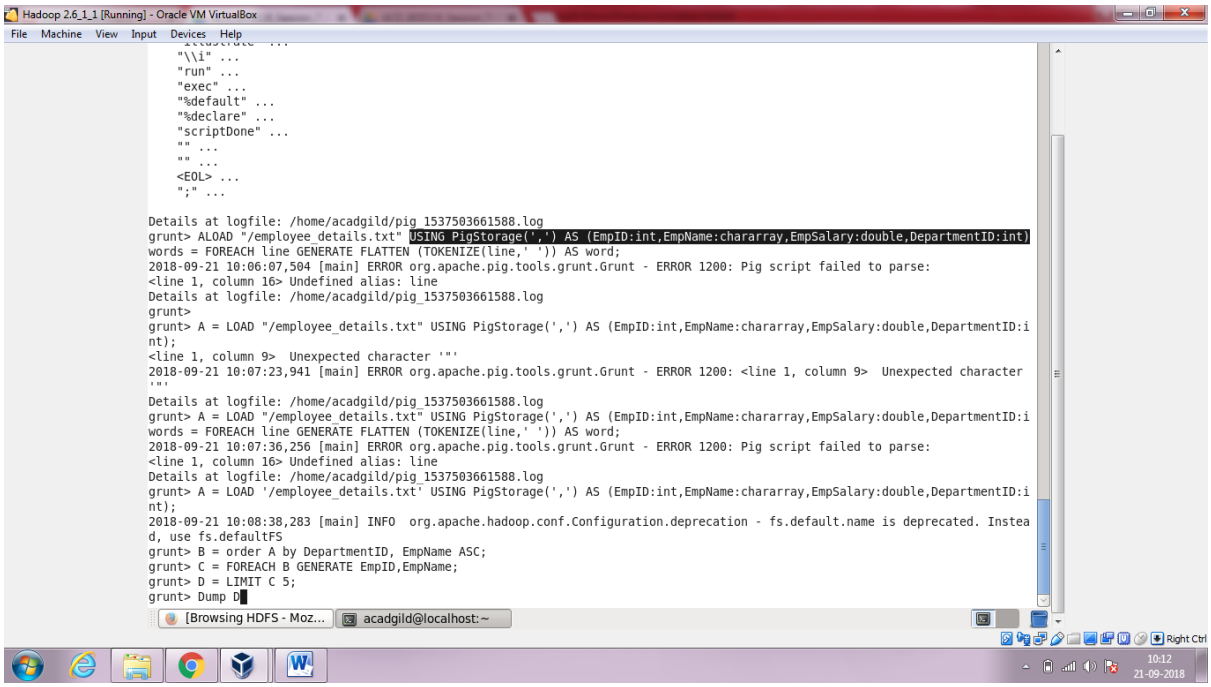


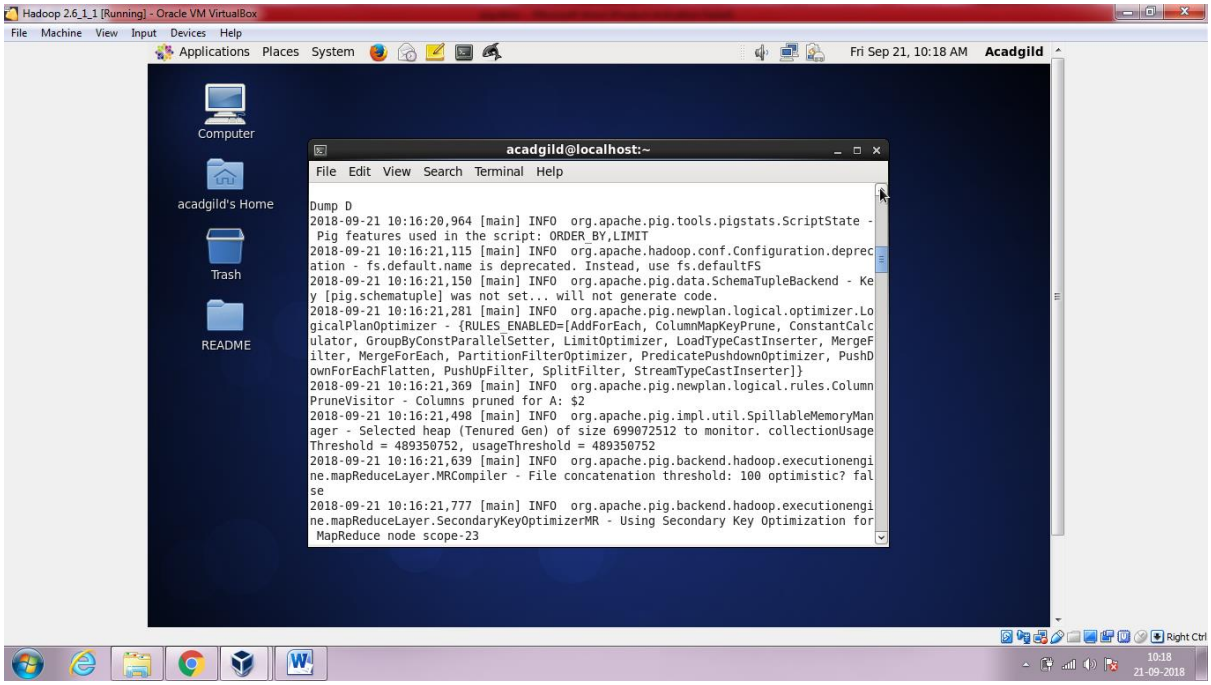

Task2:input file



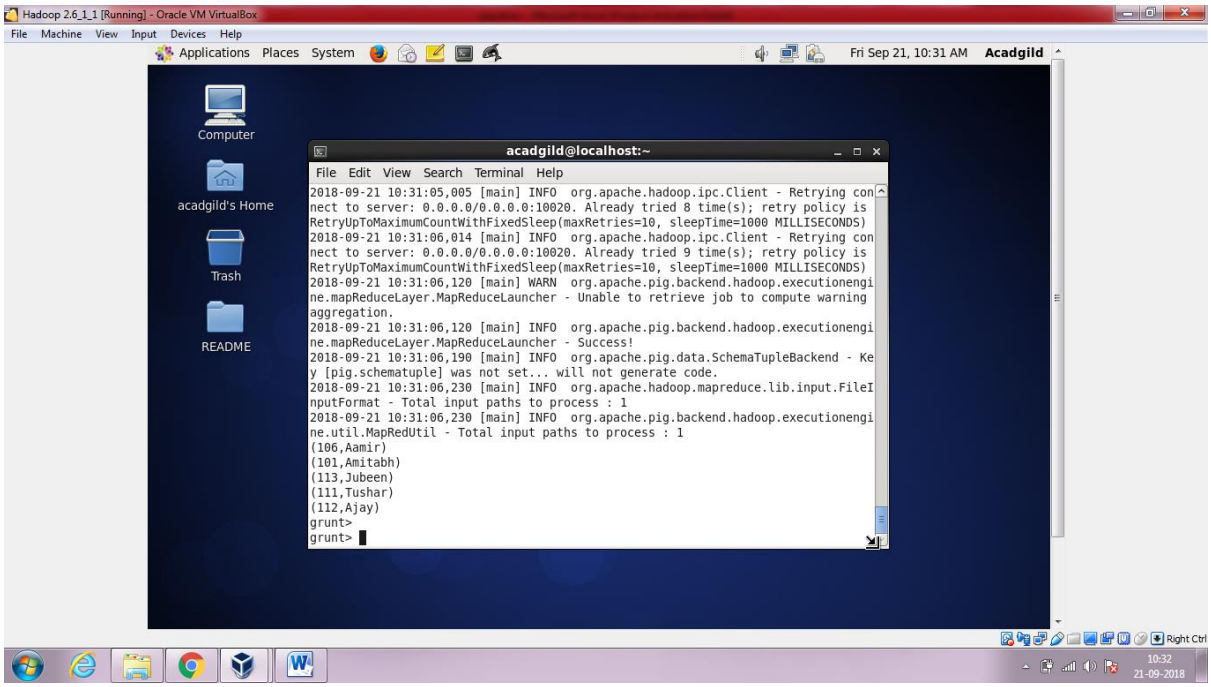



a)Top 5 employees (employee id and employee name) with highest rating.

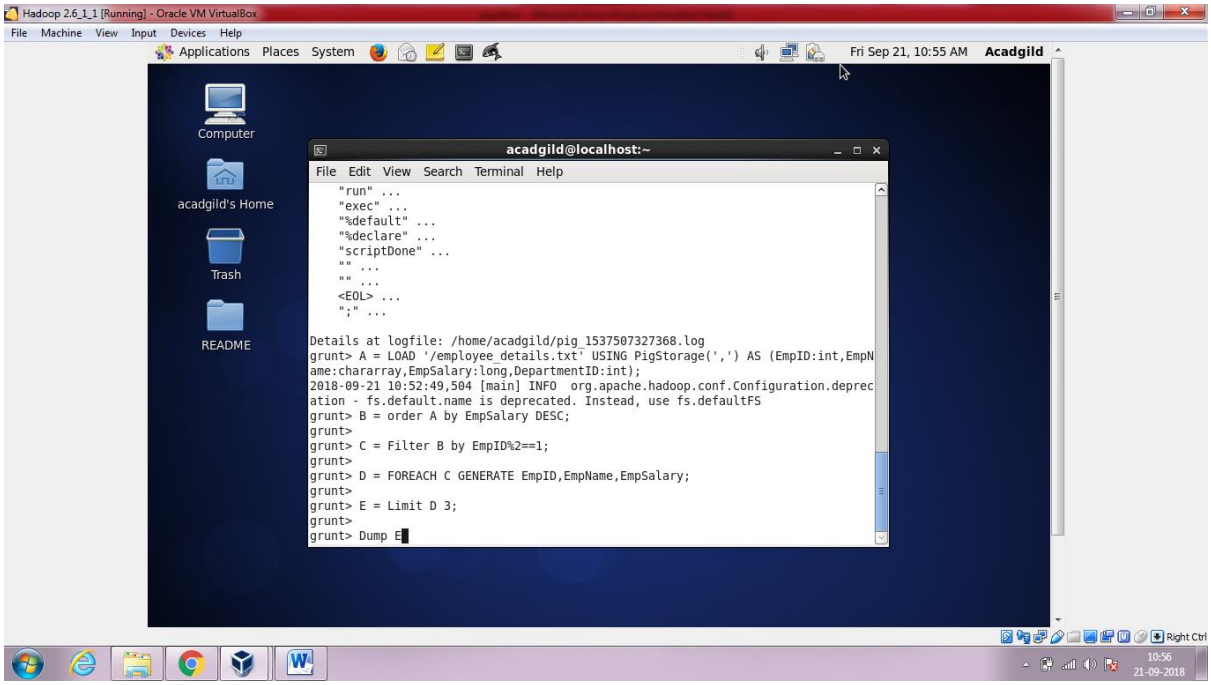


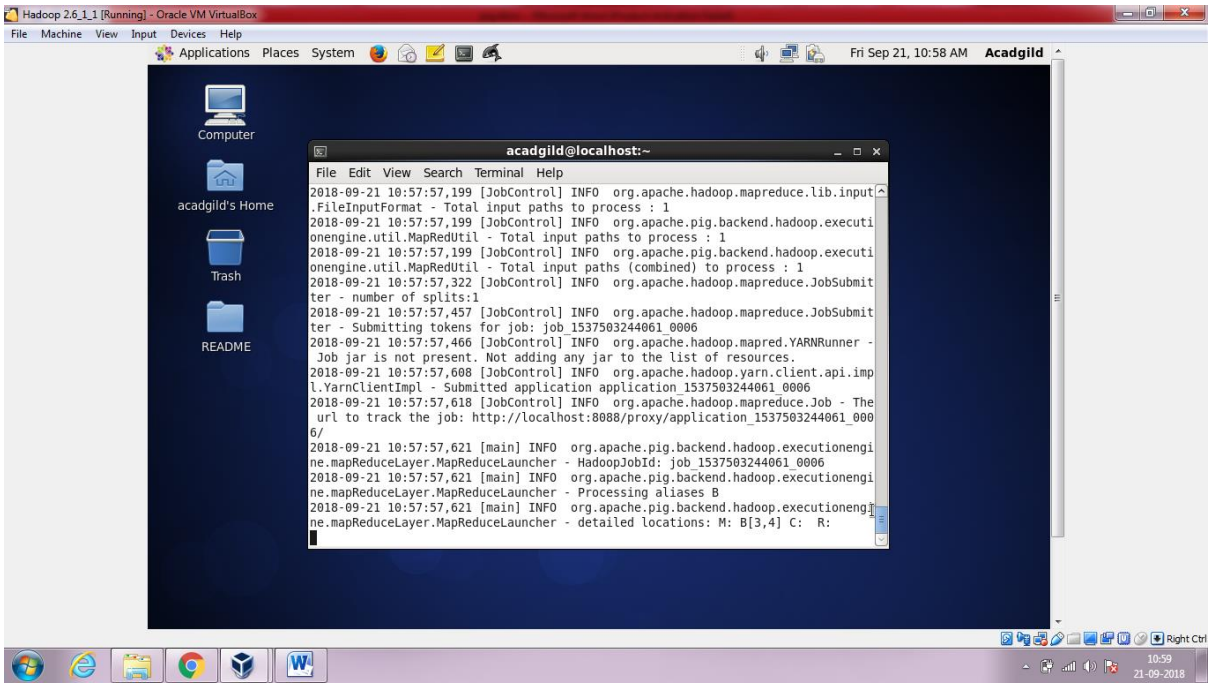



Output:

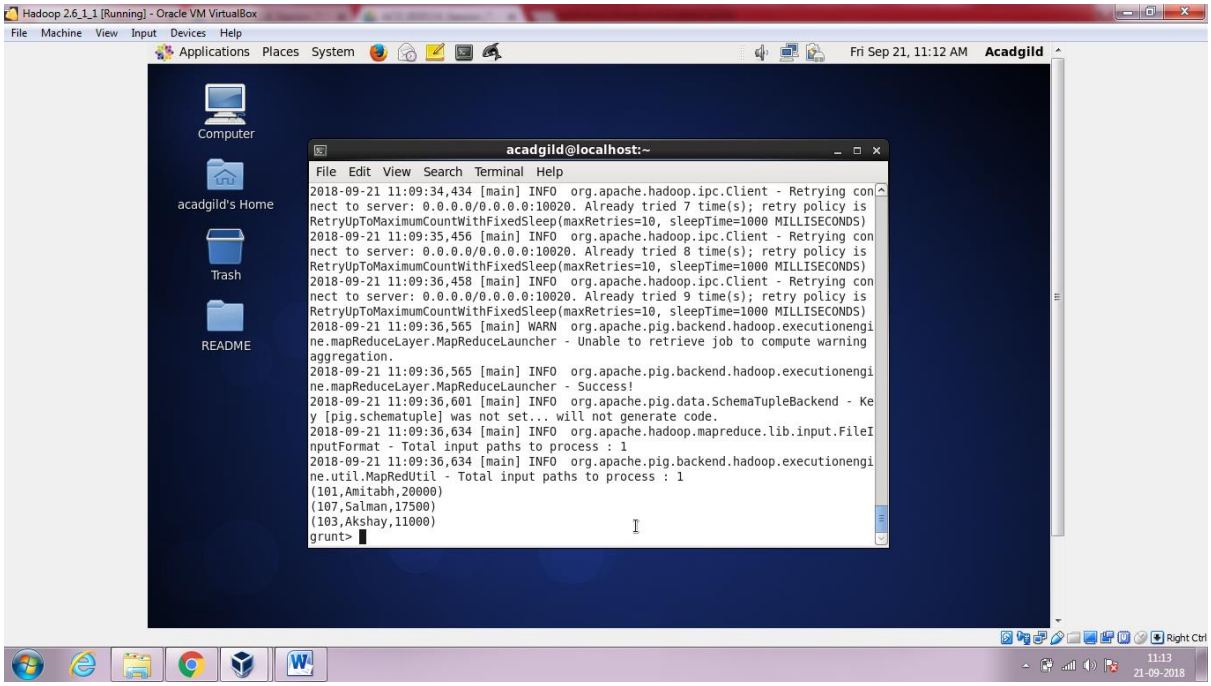



b)Top 3 employees (employee id and employee name) with highest salary,whose employye id is an odd number.

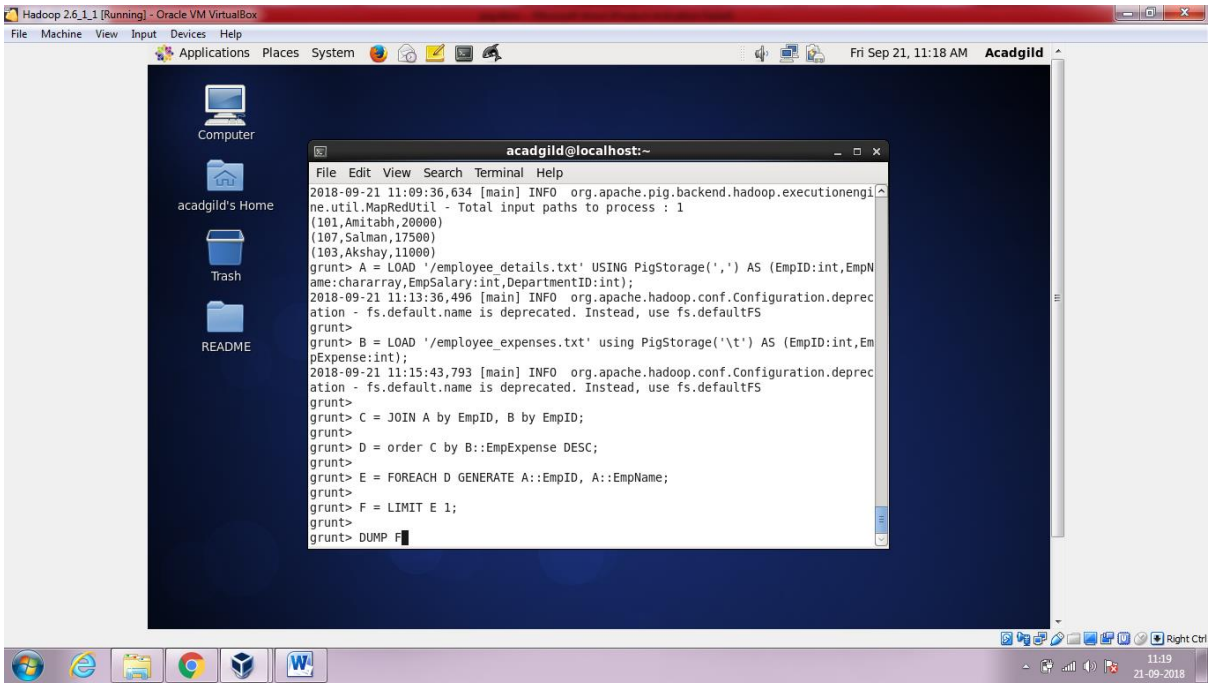

Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Fri Sep 21, 10:55 AM Acadgild
Computer
acadgild's Home
Trash
README
acadgild@localhost:~
File Edit View Search Terminal Help
"run" ...
"exec" ...
"%default" ...
"%declare" ...
"scriptDone" ...
"" ...
"" ...
<EOL> ...
";" ...
Details at logfile: /home/acadgild/pig_1537507327368.log
grunt> A = LOAD '/employee_details.txt' USING PigStorage(',') AS (EmpID:int,EmpName:chararray,EmpSalary:long,DepartmentID:int);
2018-09-21 10:52:49,504 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = order A by EmpSalary DESC;
grunt>
grunt> C = Filter B by EmpID%2==1;
grunt>
grunt> D = FOREACH C GENERATE EmpID,EmpName,EmpSalary;
grunt>
grunt> E = Limit D 3;
grunt>
grunt> Dump E
Right Ctrl
10:56
21-09-2018


Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Fri Sep 21, 10:58 AM Acadgild
Computer
acadgild's Home
Trash
README
acadgild@localhost:~
File Edit View Search Terminal Help
2018-09-21 10:57:57,199 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-21 10:57:57,199 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2018-09-21 10:57:57,199 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2018-09-21 10:57:57,322 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2018-09-21 10:57:57,457 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1537503244061_0006
2018-09-21 10:57:57,466 [JobControl] INFO org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2018-09-21 10:57:57,608 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1537503244061_0006
2018-09-21 10:57:57,618 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8088/proxy/application_1537503244061_0006/
2018-09-21 10:57:57,621 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1537503244061_0006
2018-09-21 10:57:57,621 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases B
2018-09-21 10:57:57,621 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: B[3,4] C: R:
Right Ctrl
10:59
21-09-2018

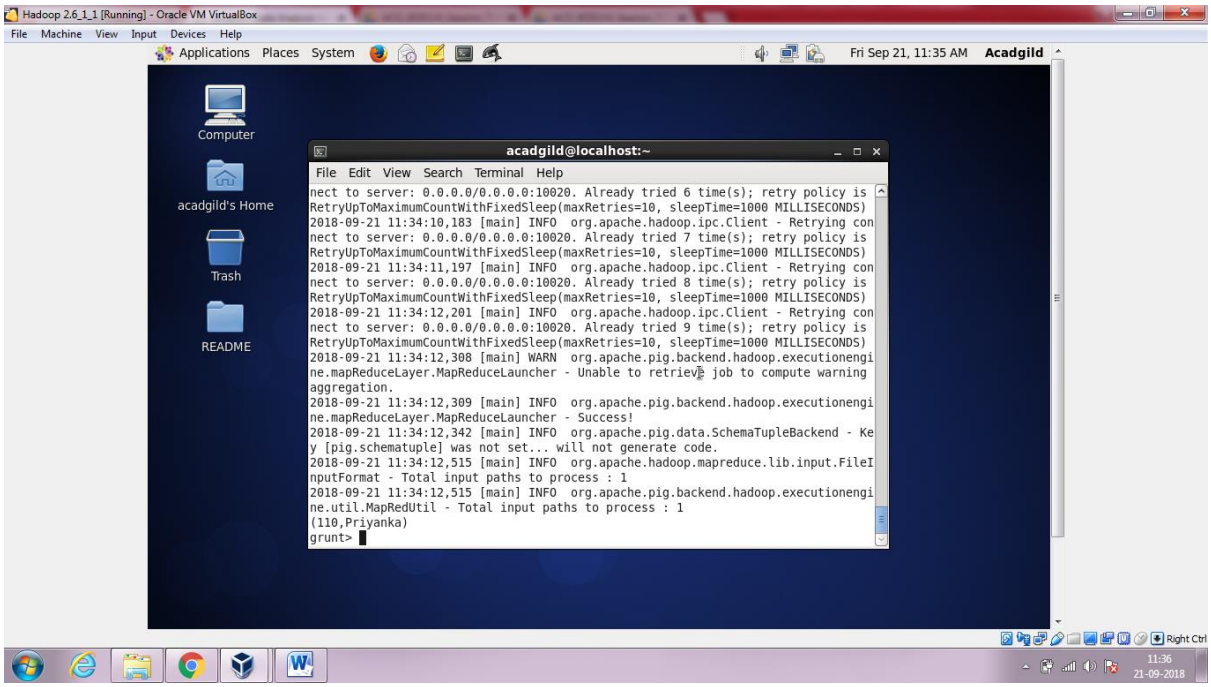
Output:



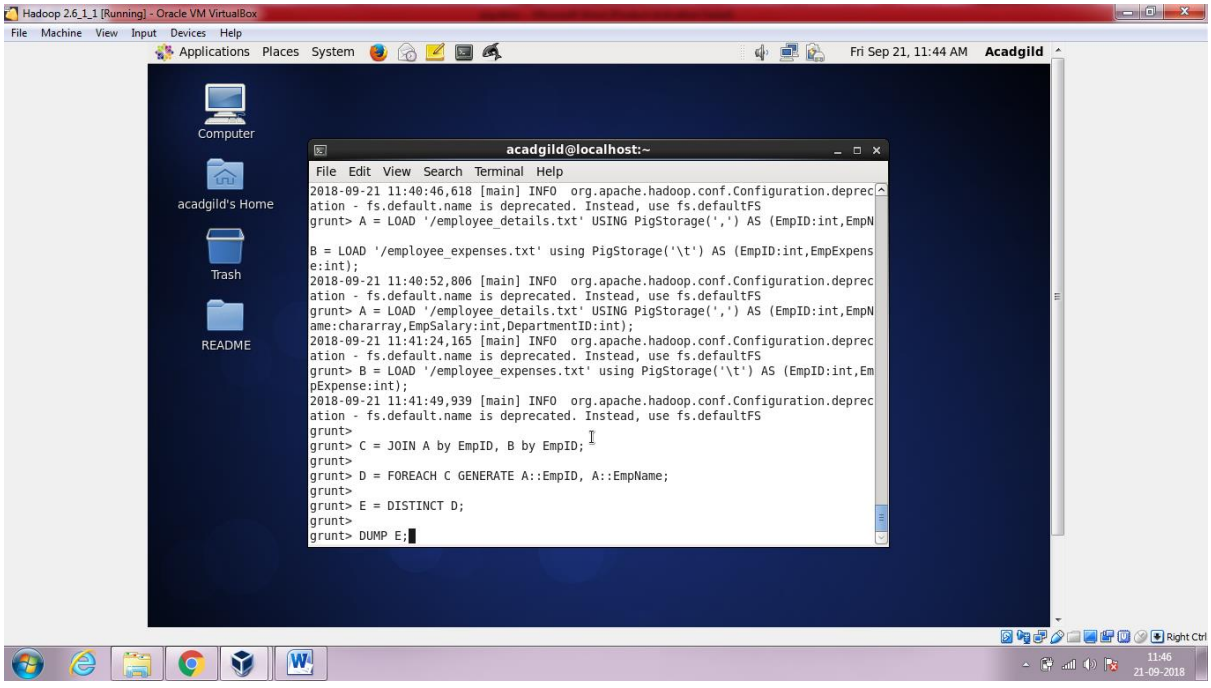

c)Employee(employee id and employee name ) with maximum expense (In case two employee have same expense,employee with name coming first in dictionary should get preference)



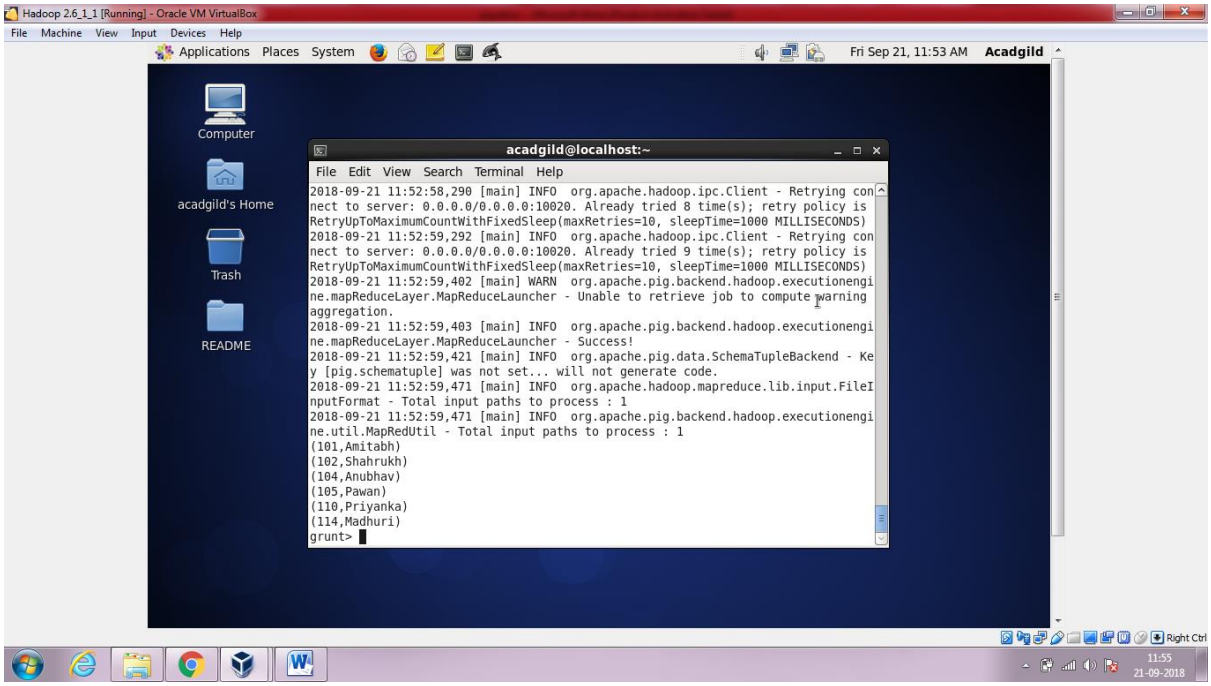
Output:



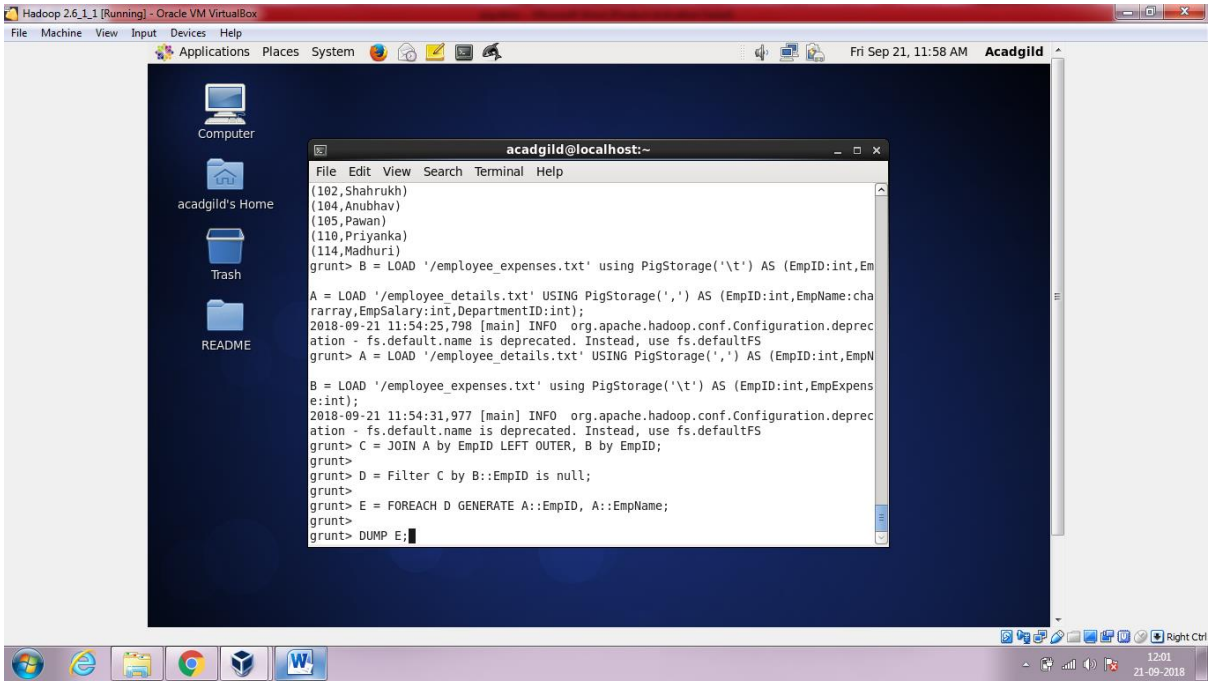

d) List of employees (employee id and employee name ) having entries i9n employee_expenses file.



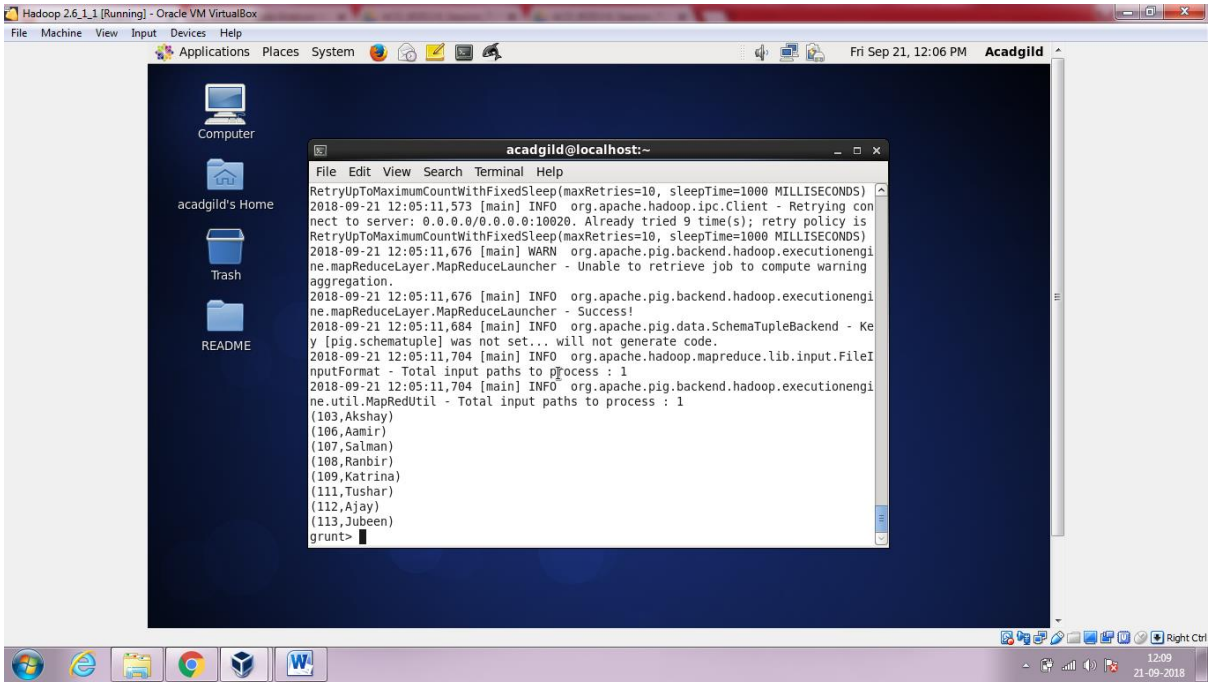
Output:




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

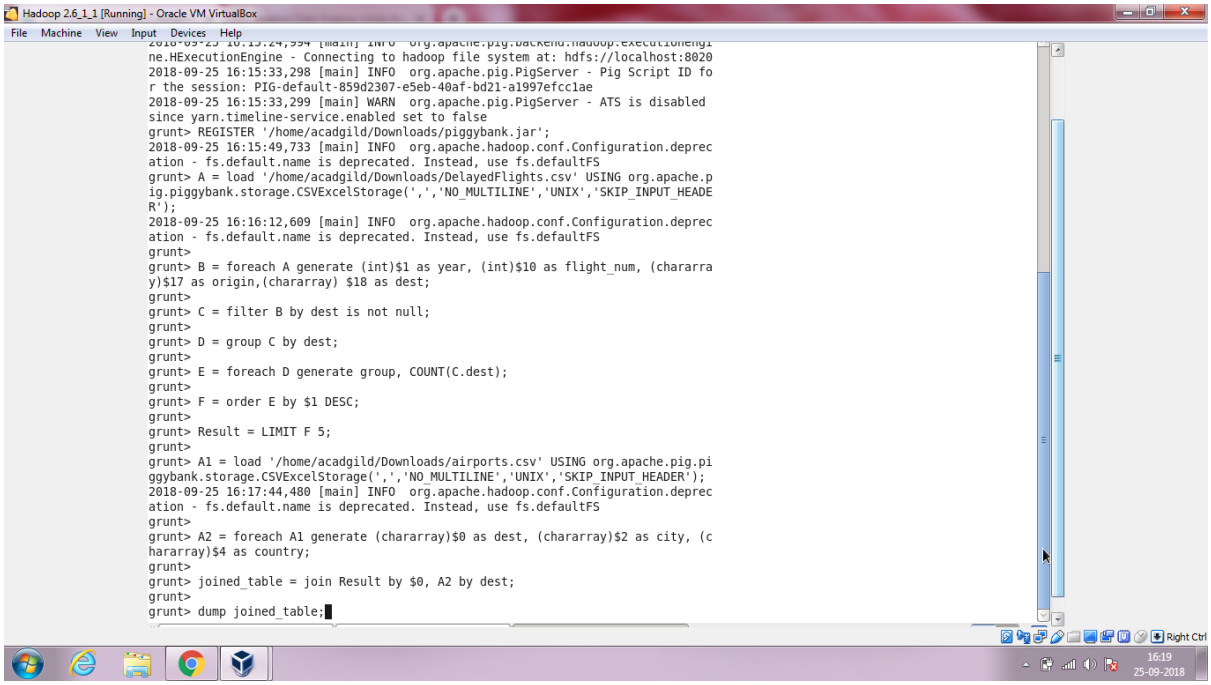



Output:

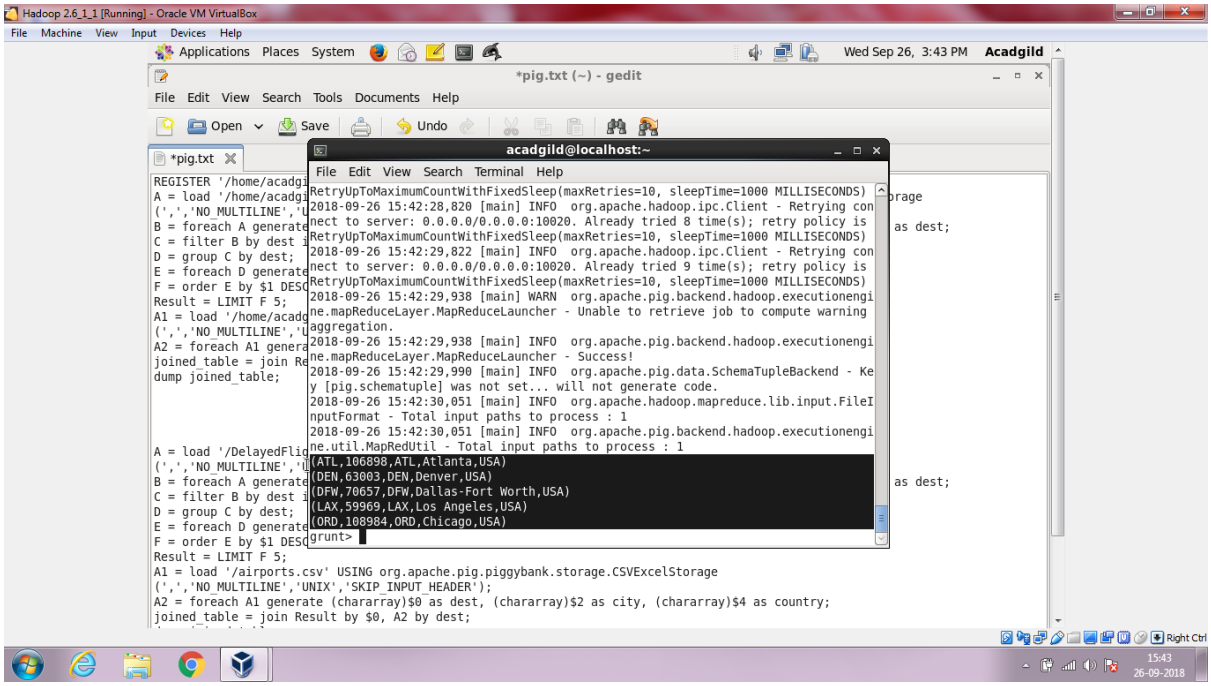



Task3:**Implementing the use case** aviation-data-analysis-using-apache-pig.

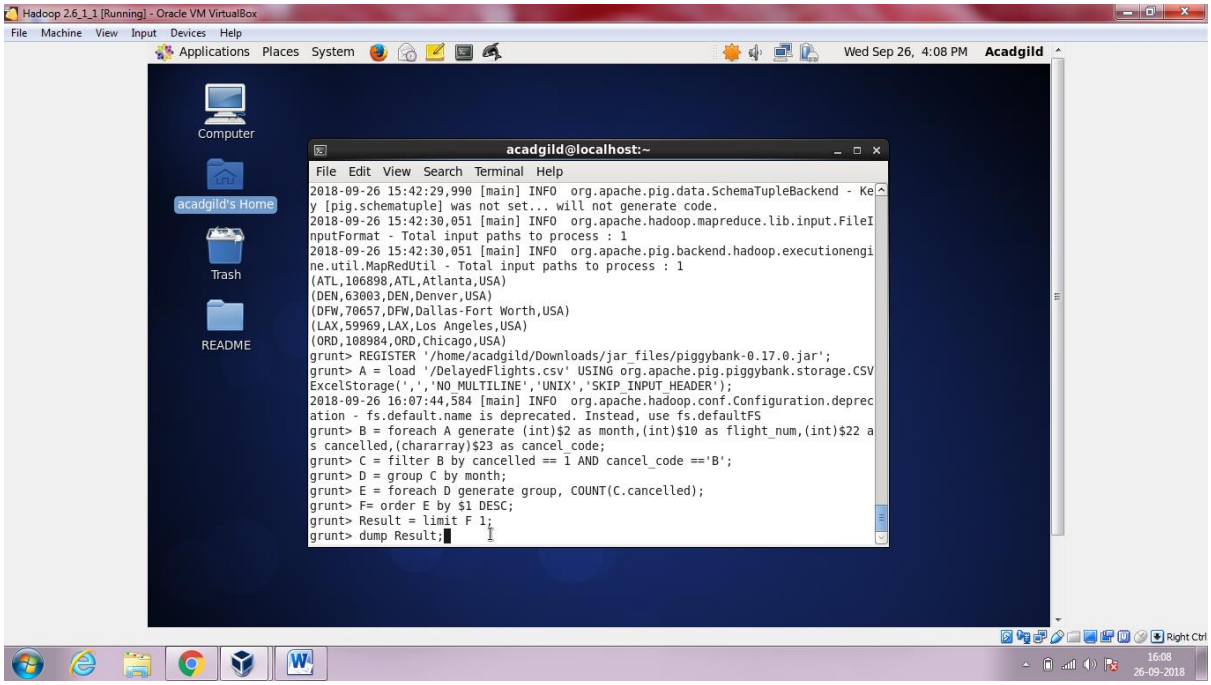
Problem 1:Top 5 most viewed destinations.



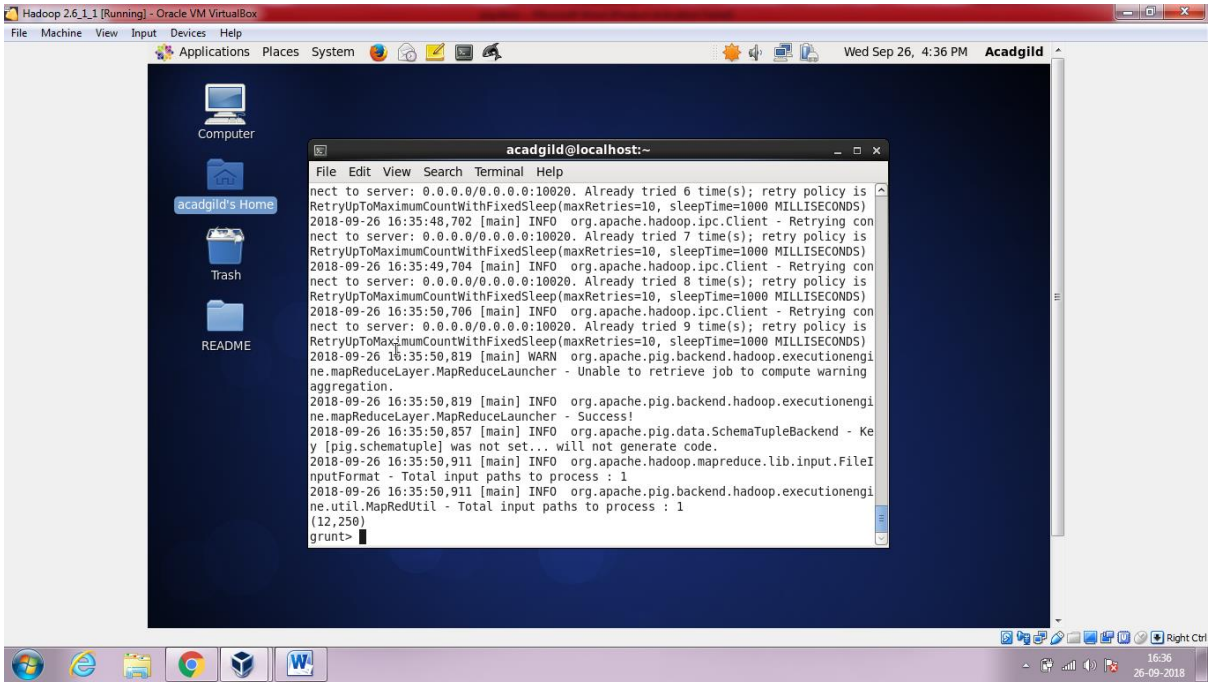

Output:



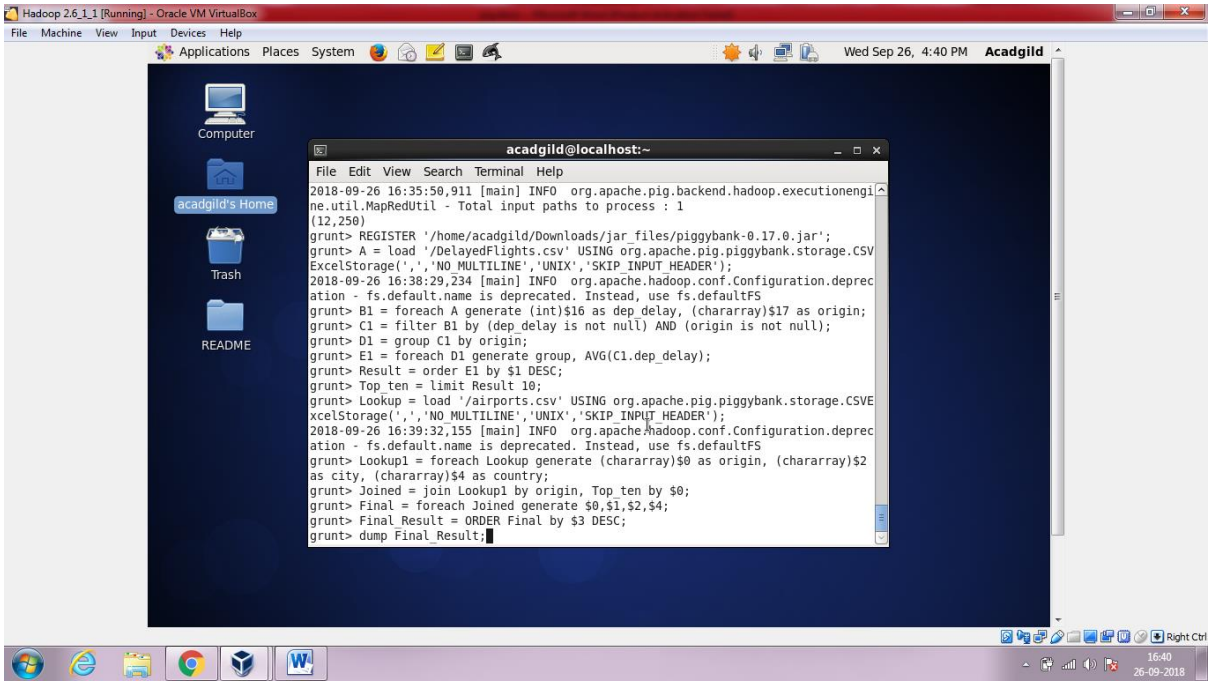

Problem 2:Which month has seen the most number of cancellations due to whether.



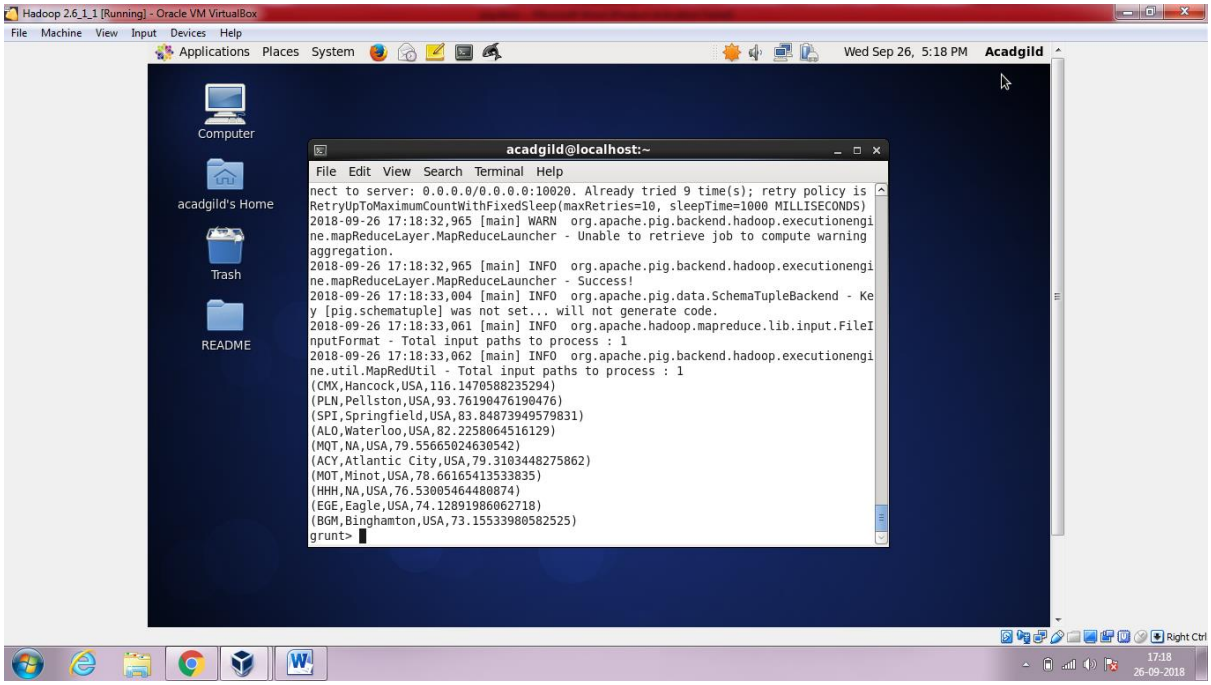

Output:




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

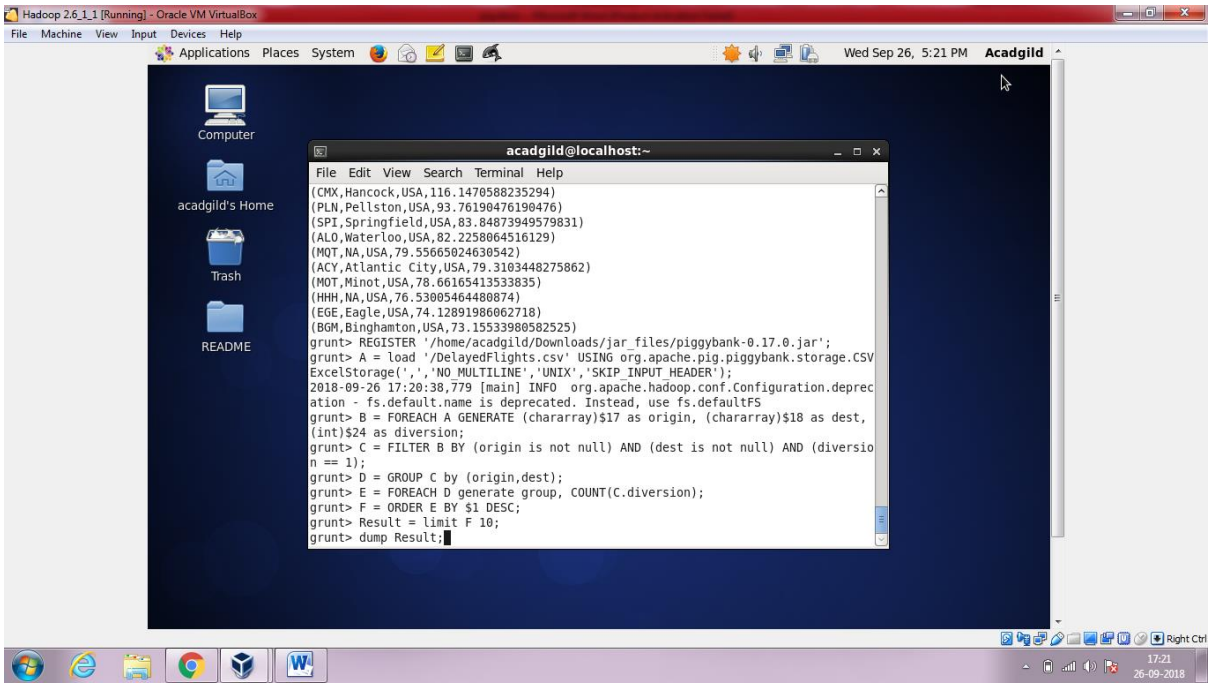

Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Wed Sep 26, 4:40 PM Acadgild
Computer
acadgild's Home
Trash
README
acadgild@localhost:~
File Edit View Search Terminal Help
2018-09-26 16:35:50,911 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
grunt> REGISTER '/home/acadgild/Downloads/jar_files/piggybank-0.17.0.jar';
grunt> A = load '/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-09-26 16:38:29,234 [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 '/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-09-26 16:39:32,155 [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;
Right Ctrl
16:40
26-09-2018

Output:


Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Wed Sep 26, 5:18 PM Acadgild
Computer
acadgild's Home
Trash
README
acadgild@localhost:~
File Edit View Search Terminal Help
nect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2018-09-26 17:18:32,965 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2018-09-26 17:18:32,965 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-26 17:18:33,004 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-09-26 17:18:33,061 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-26 17:18:33,062 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(CMX,Hancock,USA,116.1470588235294)
(PLN,Pellston,USA,93.76190476190476)
(SPI,Springfield,USA,83.84873949579831)
(ALO,Waterloo,USA,82.2258064516129)
(MQT,NA,USA,79.55665024630542)
(ACY,Atlantic City,USA,79.3103448275862)
(MOT,Minot,USA,78.66165413533835)
(HHH,NA,USA,76.53005464480874)
(EGE,Eagle,USA,74.12891986062718)
(BGM,Binghamton,USA,73.15533980582525)
grunt>
Right Ctrl
17:18
26-09-2018

Problem 4:

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




Output: