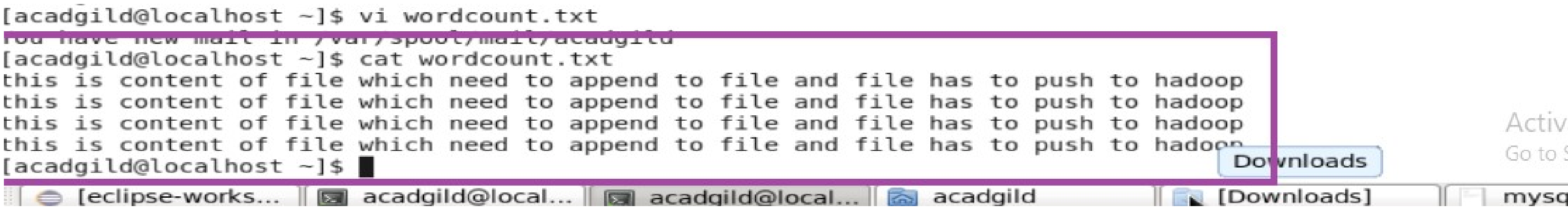
PIG Assignment

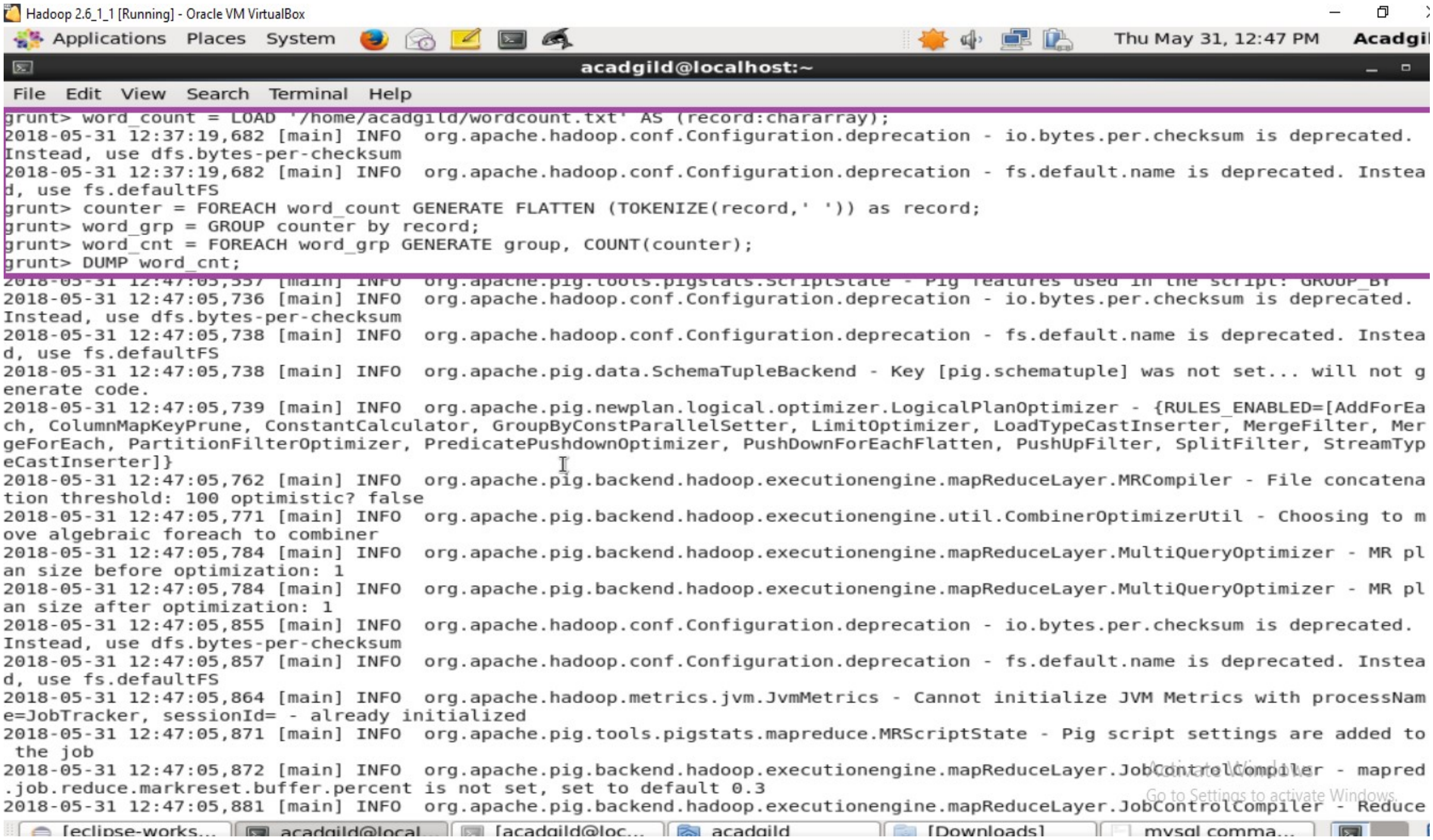
TASK-1

Write a program to implement wordcount using Pig.



```
[acadgild@localhost ~]$ vi wordcount.txt
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ cat wordcount.txt
this is content of file which need to append to file and file has to push to hadoop
this is content of file which need to append to file and file has to push to hadoop
this is content of file which need to append to file and file has to push to hadoop
this is content of file which need to append to file and file has to push to hadoop
[acadgild@localhost ~]$ 
```

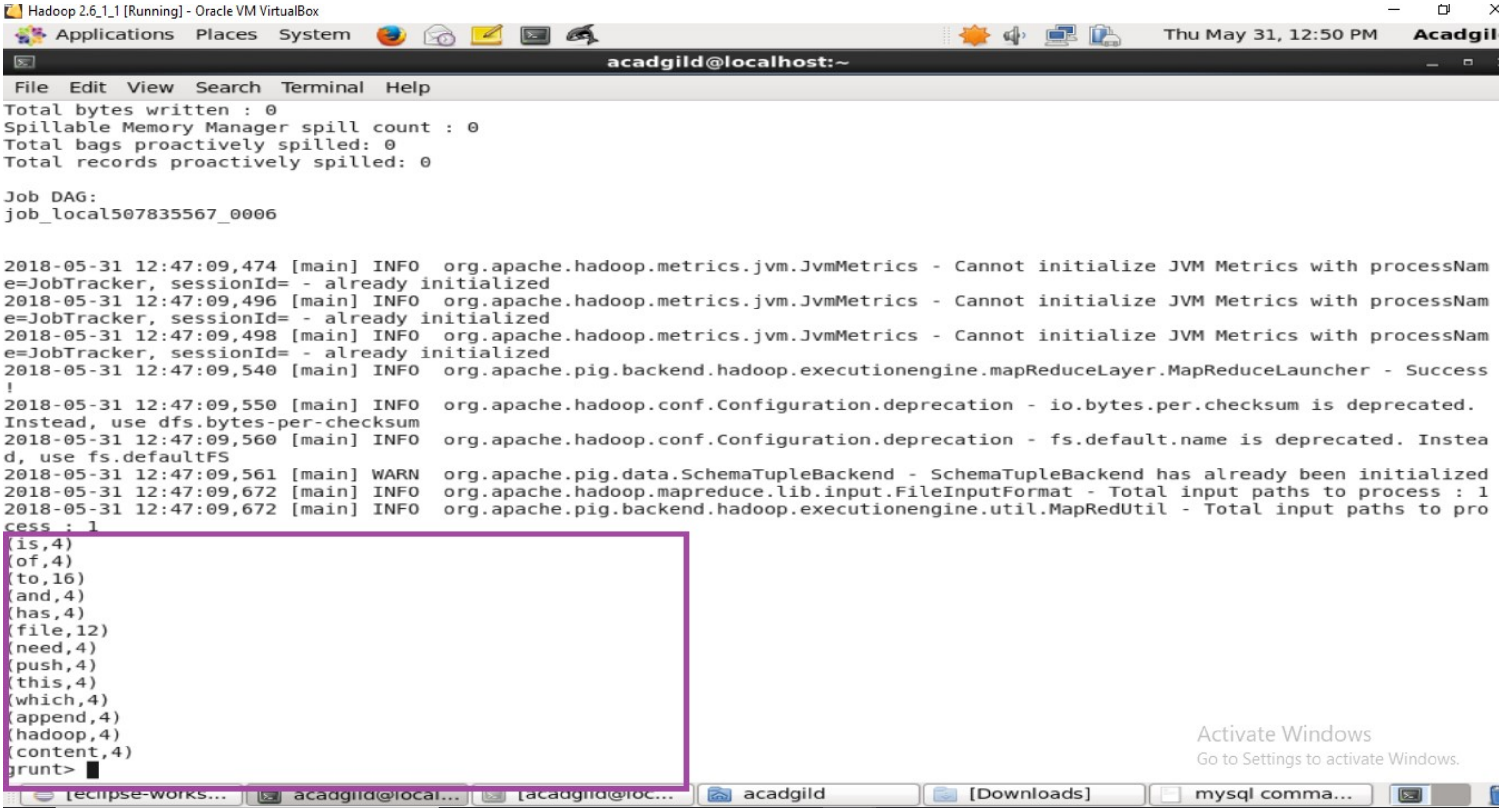
Processing



Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
Applications Places System
Thu May 31, 12:47 PM
Acadgi
acadgild@localhost:~
File Edit View Search Terminal Help

```
grunt> word_count = LOAD '/home/acadgild/wordcount.txt' AS (record:chararray);
2018-05-31 12:37:19,682 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-31 12:37:19,682 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> counter = FOREACH word_count GENERATE FLATTEN (TOKENIZE(record,' ')) as record;
grunt> word_grp = GROUP counter by record;
grunt> word_cnt = FOREACH word_grp GENERATE group, COUNT(counter);
grunt> DUMP word_cnt;
2018-05-31 12:47:05,557 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2018-05-31 12:47:05,736 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-31 12:47:05,738 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-05-31 12:47:05,738 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not g
enerate code.
2018-05-31 12:47:05,739 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEa
ch, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, Mer
geForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTyp
eCastInserter]}
2018-05-31 12:47:05,762 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatena
tion threshold: 100 optimistic? false
2018-05-31 12:47:05,771 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to m
ove algebraic foreach to combiner
2018-05-31 12:47:05,784 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size before optimization: 1
2018-05-31 12:47:05,784 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size after optimization: 1
2018-05-31 12:47:05,855 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-31 12:47:05,857 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-05-31 12:47:05,864 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 12:47:05,871 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to
 the job
2018-05-31 12:47:05,872 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred
.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-05-31 12:47:05,881 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce
```

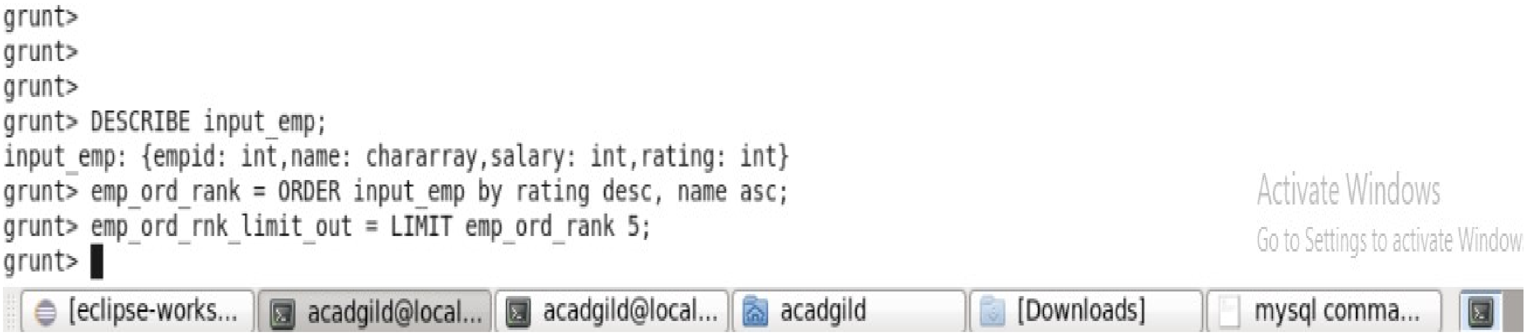
Output



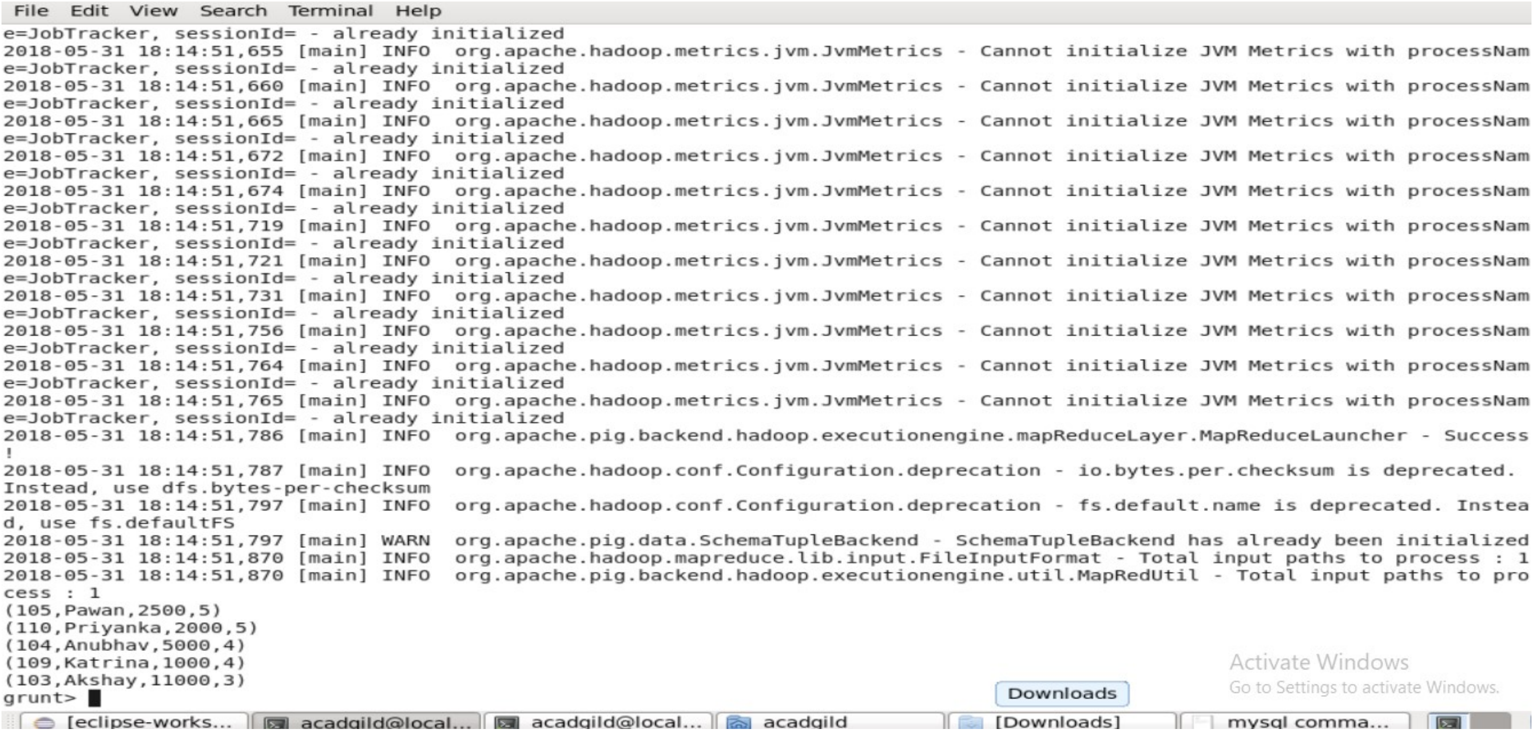
```
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local507835567_0006


2018-05-31 12:47:09,474 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 12:47:09,496 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 12:47:09,498 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 12:47:09,540 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-05-31 12:47:09,550 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-31 12:47:09,560 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-05-31 12:47:09,561 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-31 12:47:09,672 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-31 12:47:09,672 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(is,4)
(of,4)
(to,16)
(and,4)
(has,4)
(file,12)
(need,4)
(push,4)
(this,4)
(which,4)
(append,4)
(hadoop,4)
(content,4)
grunt> █
```

TASK-2(A)

(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)



```
grunt>
grunt>
grunt>
grunt> DESCRIBE input_emp;
input_emp: {empid: int,name: chararray,salary: int,rating: int}
grunt> emp_ord_rank = ORDER input_emp by rating desc, name asc;
grunt> emp_ord_rnk_limit_out = LIMIT emp_ord_rank 5;
grunt> █
```



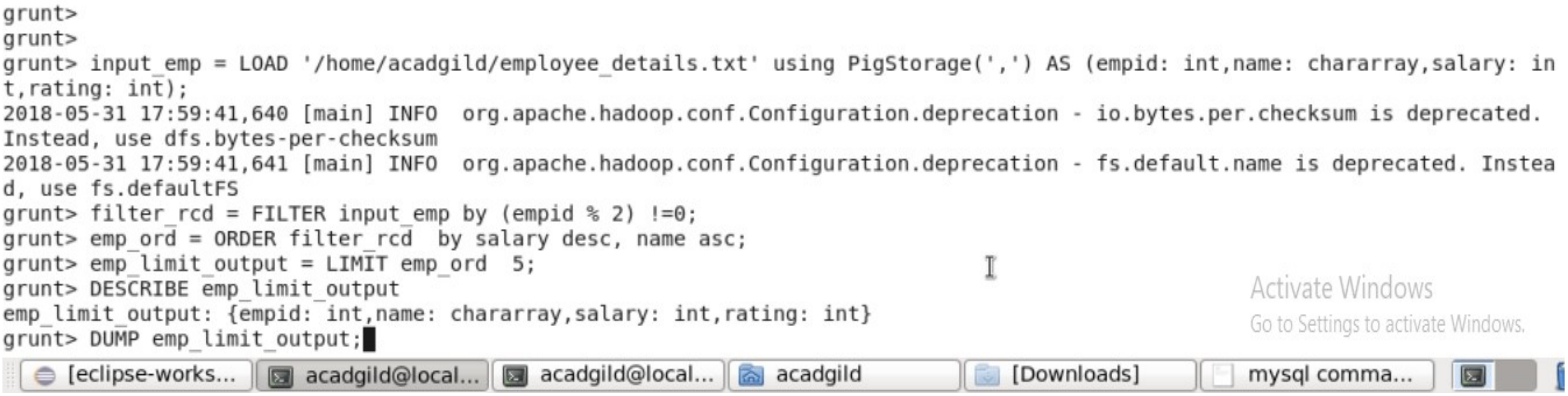
```
e=JobTracker, sessionId= - already initialized
2018-05-31 18:14:51,655 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:14:51,660 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:14:51,665 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:14:51,672 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:14:51,674 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:14:51,719 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:14:51,721 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:14:51,731 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:14:51,756 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:14:51,764 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:14:51,765 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:14:51,786 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-05-31 18:14:51,787 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-31 18:14:51,797 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-05-31 18:14:51,797 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-31 18:14:51,870 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-31 18:14:51,870 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(105,Pawan,2500,5)
(110,Priyanka,2000,5)
(104,Anubhav,5000,4)
(109,Katrina,1000,4)
(103,Akshay,11000,3)
grunt> 
```

TASK-2-(b)

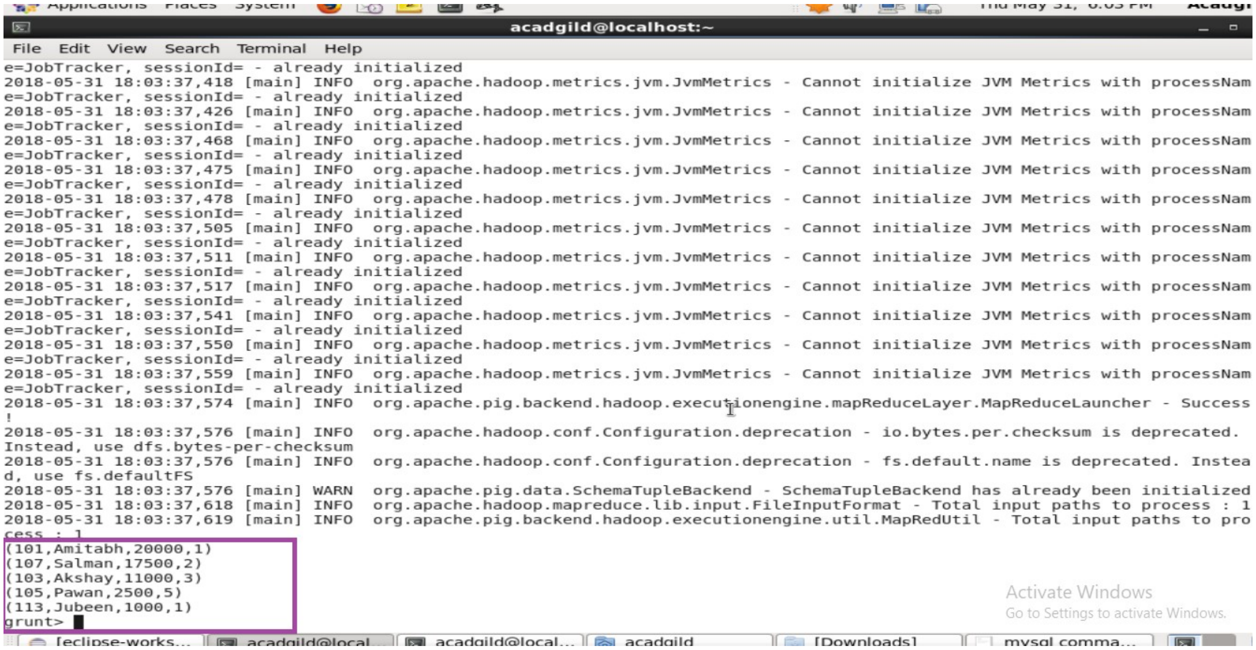
(b) Top 3 employees (employee id and employee name) with highest salary, whose employee id

is an odd number. (In case two employees have same salary, employee with name coming first

in dictionary should get preference)

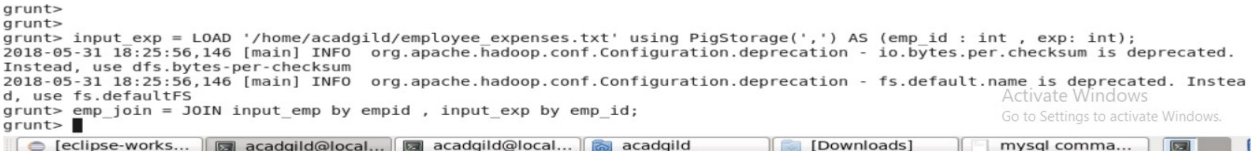


```
grunt>
grunt>
grunt> input_emp = LOAD '/home/acadgild/employee_details.txt' using PigStorage(',') AS (empid: int,name: chararray,salary: in
t,rating: int);
2018-05-31 17:59:41,640 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-31 17:59:41,641 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> filter_rcd = FILTER input_emp by (empid % 2) !=0;
grunt> emp_ord = ORDER filter_rcd  by salary desc, name asc;
grunt> emp_limit_output = LIMIT emp_ord  5;
grunt> DESCRIBE emp_limit_output
emp_limit_output: {empid: int,name: chararray,salary: int,rating: int}
grunt> DUMP emp_limit_output;
```

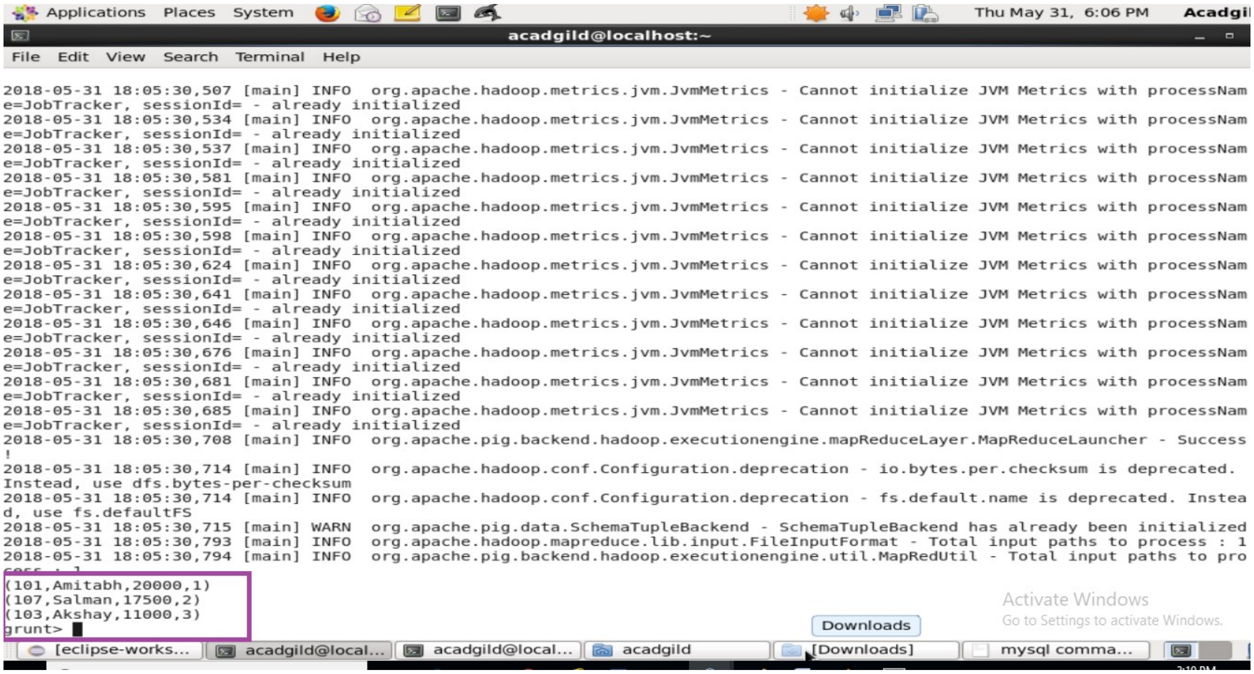


```
e=JobTracker, sessionId= - already initialized
2018-05-31 18:03:37,418 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:03:37,426 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:03:37,468 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:03:37,475 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:03:37,478 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:03:37,505 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:03:37,511 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:03:37,517 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:03:37,541 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:03:37,550 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:03:37,559 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:03:37,574 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-05-31 18:03:37,576 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-31 18:03:37,576 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-05-31 18:03:37,576 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-31 18:03:37,618 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-31 18:03:37,619 [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)
(105,Pawan,2500,5)
(113,Jubeen,1000,1)
grunt> █
```

Highest 3 person



```
grunt>
grunt>
grunt> input_exp = LOAD '/home/acadgild/employee_expenses.txt' using PigStorage(',') AS (emp_id : int , exp: int);
2018-05-31 18:25:56,146 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-31 18:25:56,146 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> emp_join = JOIN input_emp by empid , input_exp by emp_id;
grunt> █
```

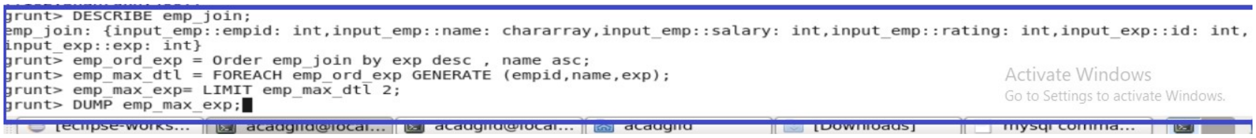


```
2018-05-31 18:05:30,507 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:05:30,534 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:05:30,537 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:05:30,581 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:05:30,595 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:05:30,598 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:05:30,624 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:05:30,641 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:05:30,646 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:05:30,676 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:05:30,681 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:05:30,685 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-05-31 18:05:30,708 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-05-31 18:05:30,714 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-31 18:05:30,714 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-05-31 18:05:30,715 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-31 18:05:30,793 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-31 18:05:30,794 [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)
grunt> █
```

Task 2 c

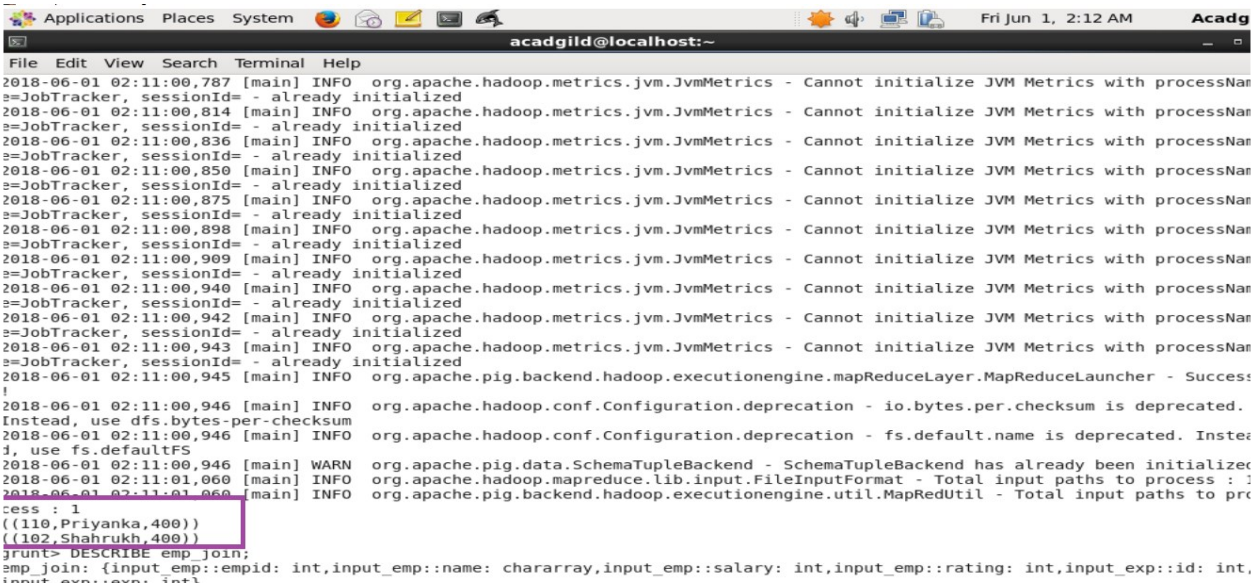
(c) Employee (employee id and employee name) with maximum expense (In case two employees have same expense, employee with name coming first in dictionary should get

preference)



```
grunt> DESCRIBE emp_join;
emp_join: {input_emp::empid: int,input_emp::name: chararray,input_emp::salary: int,input_emp::rating: int,input_exp::id: int,
input_exp::exp: int}
grunt> emp_ord_exp = Order emp_join by exp desc , name asc;
grunt> emp_max_dtl = FOREACH emp_ord_exp GENERATE (empid,name,exp);
grunt> emp_max_exp= LIMIT emp_max_dtl 2;
grunt> DUMP emp_max_exp;
```

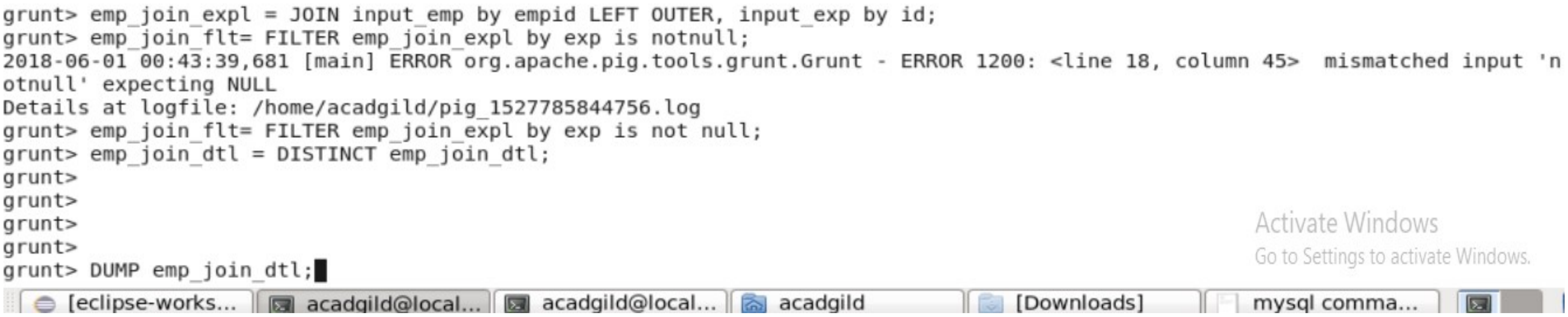
Output


Applications Places System
Fri Jun 1, 2:12 AM
Acadg
acadgild@localhost:~
File Edit View Search Terminal Help

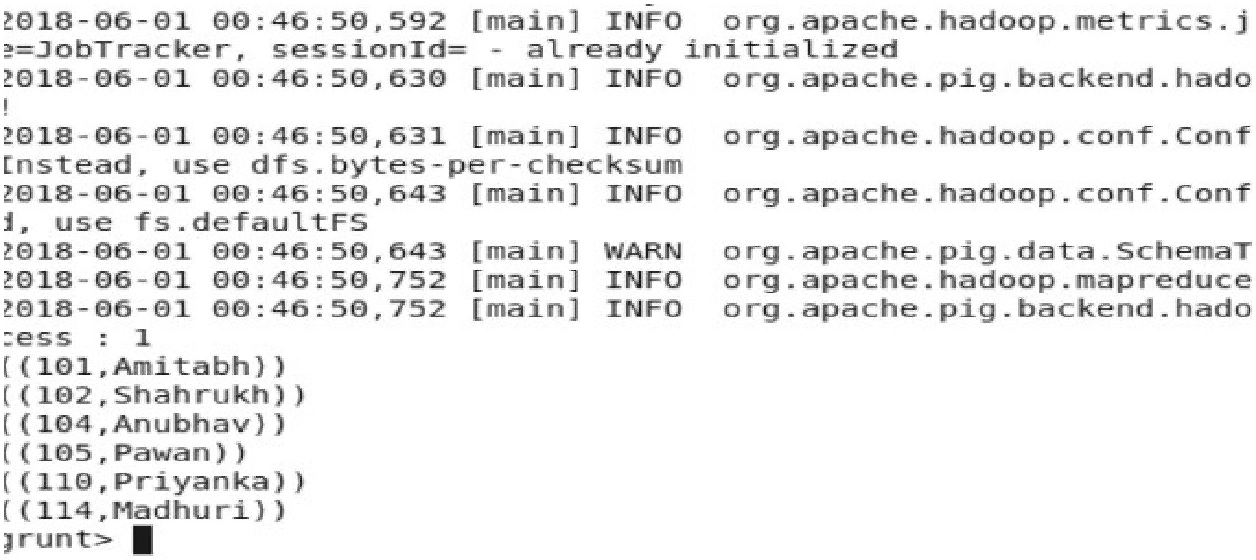
```
2018-06-01 02:11:00,787 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:11:00,814 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:11:00,836 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:11:00,850 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:11:00,875 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:11:00,898 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:11:00,909 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:11:00,940 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:11:00,942 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:11:00,943 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:11:00,945 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-06-01 02:11:00,946 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-01 02:11:00,946 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-01 02:11:00,946 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-01 02:11:01,060 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-01 02:11:01,060 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
((110,Priyanka,400))
((102,Shahrukh,400))
grunt> DESCRIBE emp_join;
emp_join: {input_emp::empid: int,input_emp::name: chararray,input_emp::salary: int,input_emp::rating: int,input_exp::id: int,
input_exp::exp: int}
```

TASK-2 (d)

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



```
grunt> emp_join_expl = JOIN input_emp by empid LEFT OUTER, input_exp by id;
grunt> emp_join_flt= FILTER emp_join_expl by exp is notnull;
2018-06-01 00:43:39,681 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 18, column 45>  mismatched input 'n
otnull' expecting NULL
Details at logfile: /home/acadgild/pig_1527785844756.log
grunt> emp_join_flt= FILTER emp_join_expl by exp is not null;
grunt> emp_join_dtl = DISTINCT emp_join_dtl;
grunt>
grunt>
grunt>
grunt>
grunt> DUMP emp_join_dtl;
```

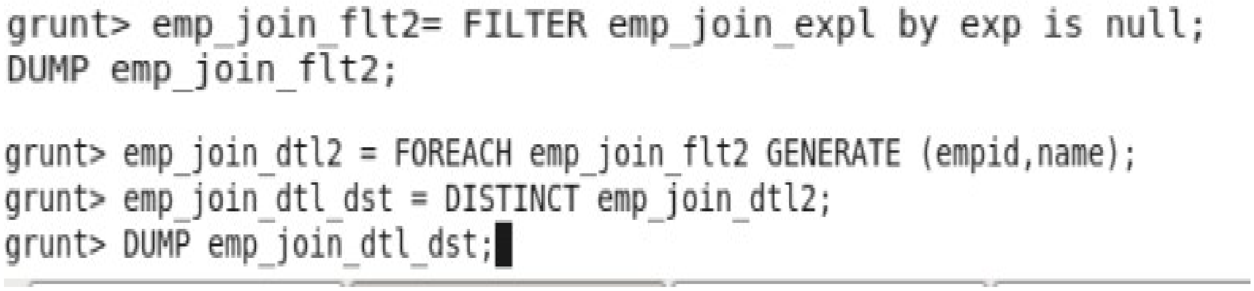


```
2018-06-01 00:46:50,592 [main] INFO  org.apache.hadoop.metrics.j
e=JobTracker, sessionId= - already initialized
2018-06-01 00:46:50,630 [main] INFO  org.apache.pig.backend.hado
!
2018-06-01 00:46:50,631 [main] INFO  org.apache.hadoop.conf.Conf
Instead, use dfs.bytes-per-checksum
2018-06-01 00:46:50,643 [main] INFO  org.apache.hadoop.conf.Conf
d, use fs.defaultFS
2018-06-01 00:46:50,643 [main] WARN  org.apache.pig.data.SchemaT
2018-06-01 00:46:50,752 [main] INFO  org.apache.hadoop.mapreduce
2018-06-01 00:46:50,752 [main] INFO  org.apache.pig.backend.hado
cess : 1
((101,Amitabh))
((102,Shahrukh))
((104,Anubhav))
((105,Pawan))
((110,Priyanka))
((114,Madhuri))
grunt> █
```

TASK-2 (E)

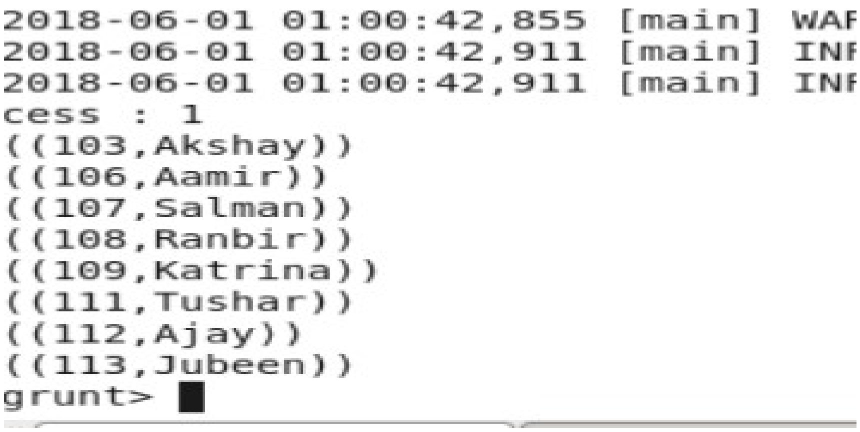
List of employees (employee id and employee name) having no entry in employee_expenses file.

Execution



```
grunt> emp_join_flt2= FILTER emp_join_expl by exp is null;
DUMP emp_join_flt2;

grunt> emp_join_dtl2 = FOREACH emp_join_flt2 GENERATE (empid,name);
grunt> emp_join_dtl_dst = DISTINCT emp_join_dtl2;
grunt> DUMP emp_join_dtl_dst;█
```

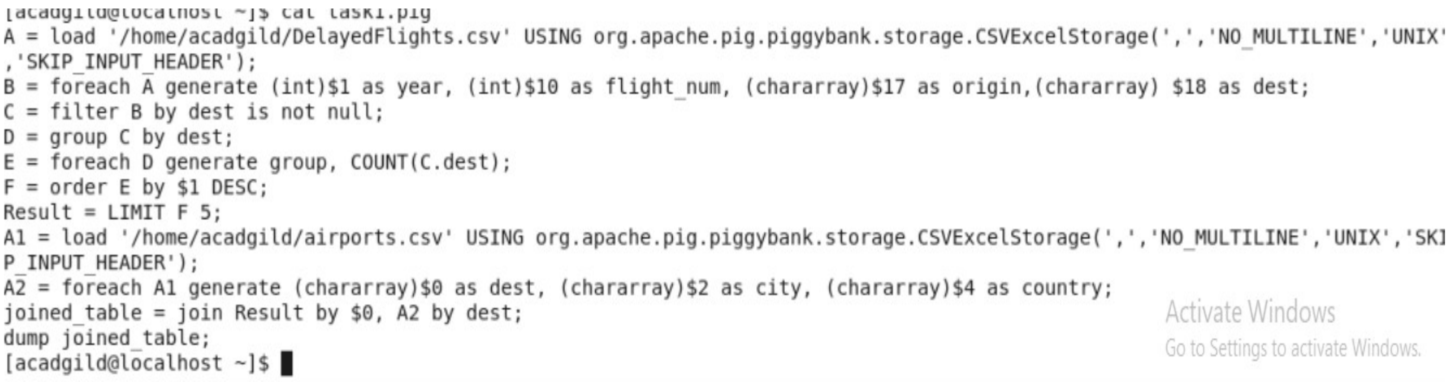
OUTPUT



```
2018-06-01 01:00:42,855 [main] WAF
2018-06-01 01:00:42,911 [main] INF
2018-06-01 01:00:42,911 [main] INF
cess : 1
((103,Akshay))
((106,Aamir))
((107,Salman))
((108,Ranbir))
((109,Katrina))
((111,Tushar))
((112,Ajay))
((113,Jubeen))
grunt> █
```

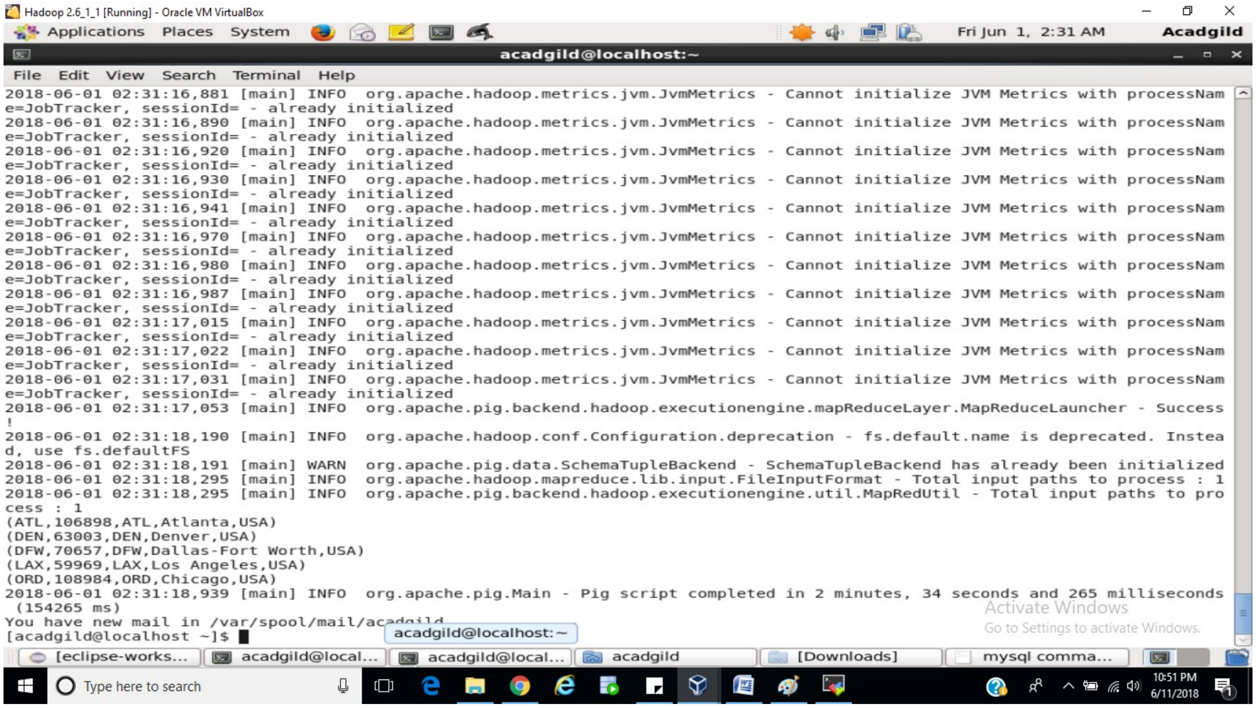
Task-3

Sub task-1



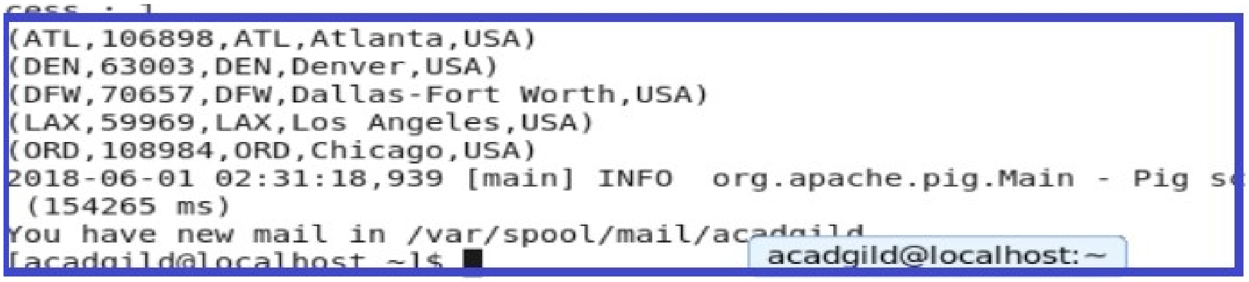
```
[acadgild@localhost ~]$ cat task1.pig
A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX'
,'SKIP_INPUT_HEADER');
B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18 as dest;
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/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKI
P_INPUT_HEADER');
A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
joined_table = join Result by $0, A2 by dest;
dump joined_table;
[acadgild@localhost ~]$ █
```

Activate Windows
Go to Settings to activate Windows.



Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
Applications Places System
Fri Jun 1, 2:31 AM
Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help

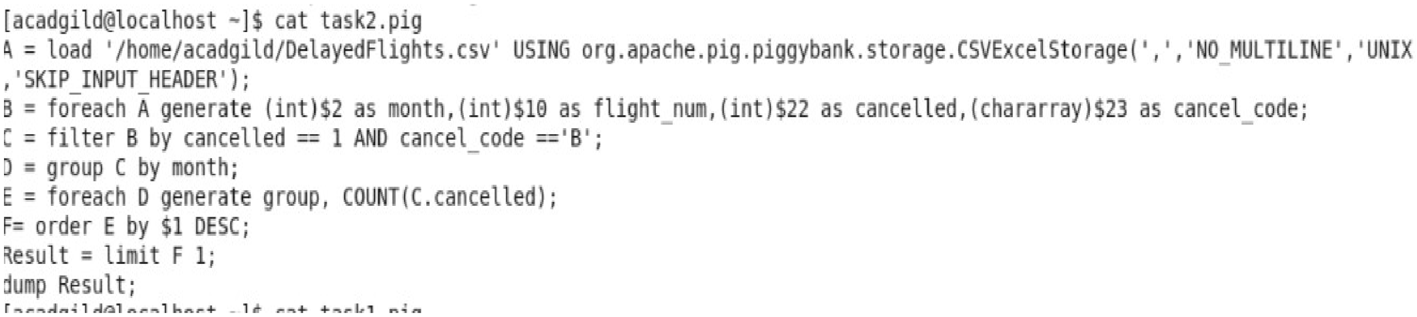
```
2018-06-01 02:31:16,881 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:31:16,890 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:31:16,920 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:31:16,930 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:31:16,941 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:31:16,970 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:31:16,980 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:31:16,987 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:31:17,015 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:31:17,022 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:31:17,031 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-01 02:31:17,053 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-06-01 02:31:18,190 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-01 02:31:18,191 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-01 02:31:18,295 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-01 02:31:18,295 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 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)
2018-06-01 02:31:18,939 [main] INFO  org.apache.pig.Main - Pig script completed in 2 minutes, 34 seconds and 265 milliseconds
 (154265 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ █
```



```
cess : 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)
2018-06-01 02:31:18,939 [main] INFO  org.apache.pig.Main - Pig sc
 (154265 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ █
```

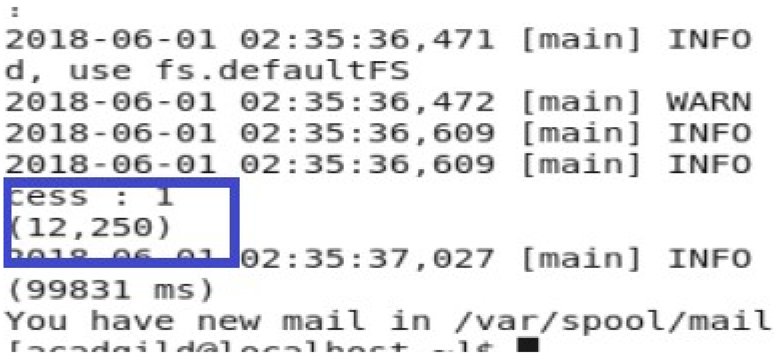
SubTask-2

Script



```
[acadgild@localhost ~]$ cat task2.pig
A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX
,'SKIP_INPUT_HEADER');
B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
C = filter B by cancelled == 1 AND cancel_code =='B';
D = group C by month;
E = foreach D generate group, COUNT(C.cancelled);
F= order E by $1 DESC;
Result = limit F 1;
dump Result;
[acadgild@localhost ~]$ cat task1.pig
```

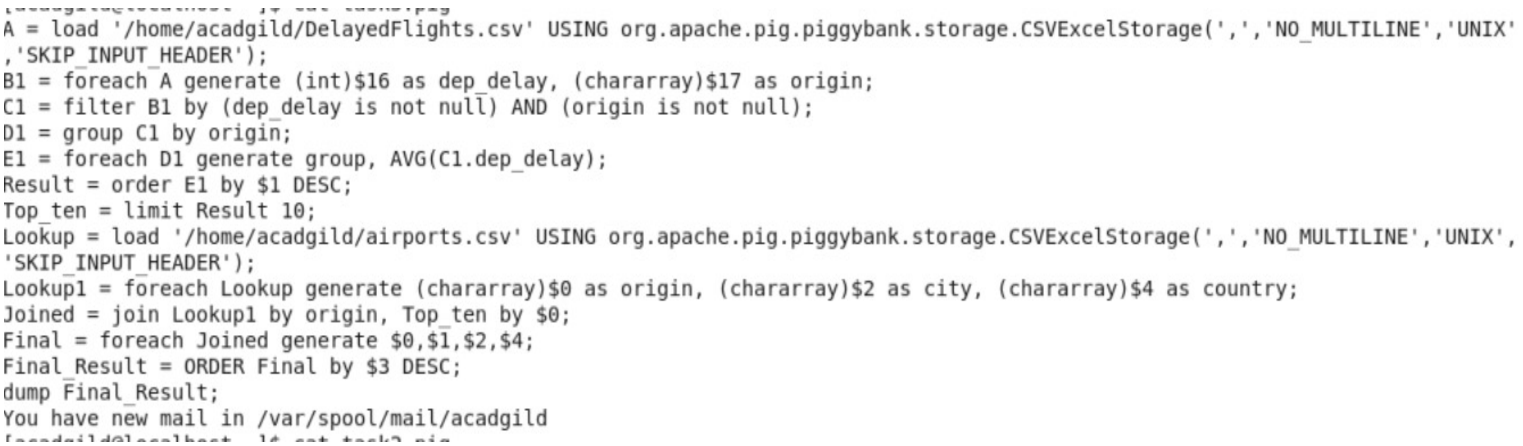
Output



```
:
2018-06-01 02:35:36,471 [main] INFO
d, use fs.defaultFS
2018-06-01 02:35:36,472 [main] WARN
2018-06-01 02:35:36,609 [main] INFO
2018-06-01 02:35:36,609 [main] INFO
cess : 1
(12,250)
2018-06-01 02:35:37,027 [main] INFO
(99831 ms)
You have new mail in /var/spool/mail
[acadgild@localhost ~]$ █
```

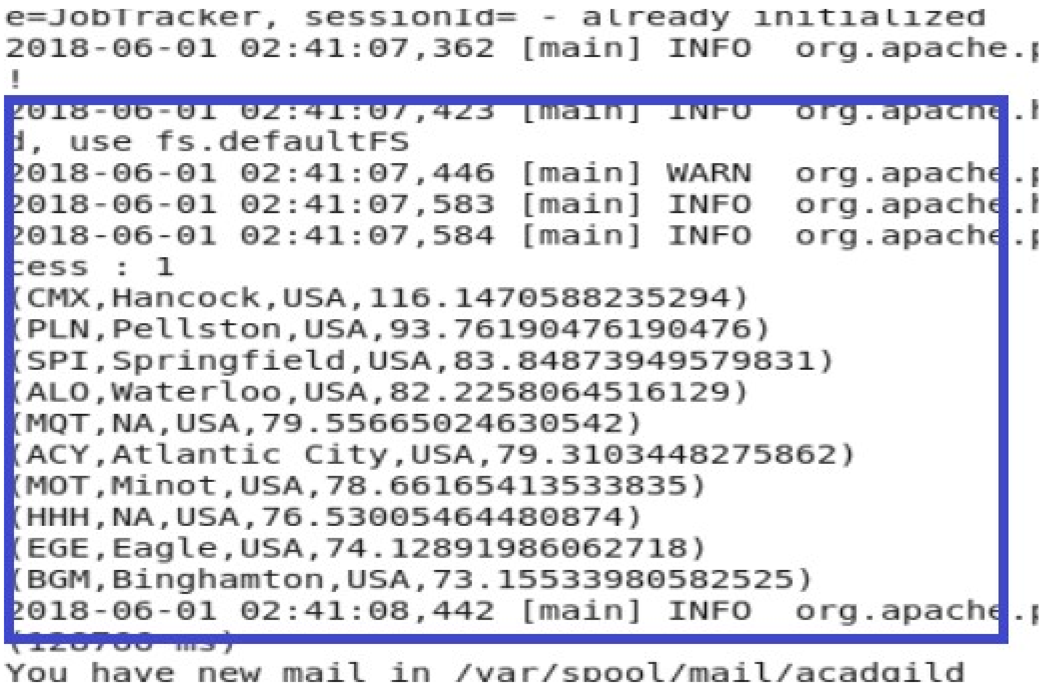
Sub task-3

Script



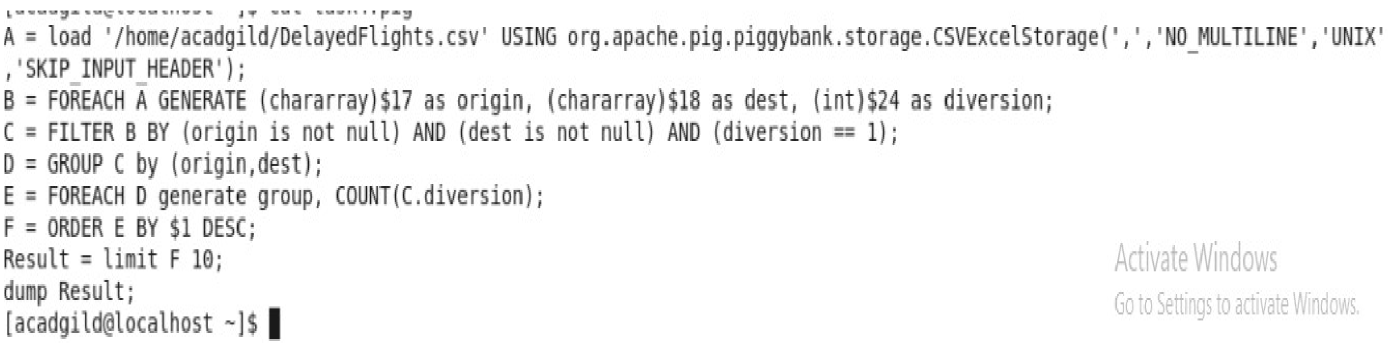
```
[acadgild@localhost ~]$ cat task3.pig
A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX'
,'SKIP_INPUT_HEADER');
B1 = foreach A generate (int)$16 as dep_delay, (chararray)$17 as origin;
C1 = filter B1 by (dep_delay is not null) AND (origin is not null);
D1 = group C1 by origin;
E1 = foreach D1 generate group, AVG(C1.dep_delay);
Result = order E1 by $1 DESC;
Top_ten = limit Result 10;
Lookup = load '/home/acadgild/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX',
'SKIP_INPUT_HEADER');
Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;
Joined = join Lookup1 by origin, Top_ten by $0;
Final = foreach Joined generate $0,$1,$2,$4;
Final_Result = ORDER Final by $3 DESC;
dump Final_Result;
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ cat task2.pig
```

Output



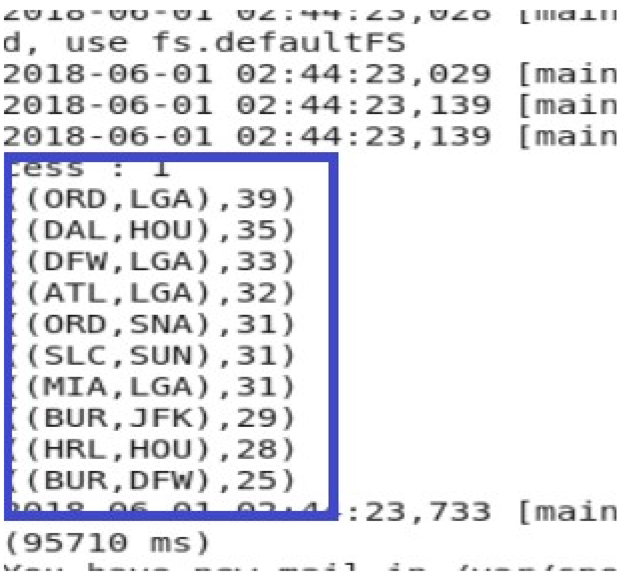
```
e=JobTracker, sessionId= - already initialized
2018-06-01 02:41:07,362 [main] INFO  org.apache.p
!
2018-06-01 02:41:07,423 [main] INFO  org.apache.h
d, use fs.defaultFS
2018-06-01 02:41:07,446 [main] WARN  org.apache.p
2018-06-01 02:41:07,583 [main] INFO  org.apache.h
2018-06-01 02:41:07,584 [main] INFO  org.apache.p
cess : 1
(CMX,Hancock,USA,116.1470588235294)
(PLN,Pellston,USA,93.76190476190476)
(SPI,Springfield,USA,83.84873949579831)
(ALO,Waterloo,USA,82.2258064516129)
(MQT,NA,USA,79.55665024630542)
(ACY,Atlantic City,USA,79.3103448275862)
(MOT,Minot,USA,78.66165413533835)
(HHH,NA,USA,76.53005464480874)
(EGE,Eagle,USA,74.12891986062718)
(BGM,Binghamton,USA,73.15533980582525)
2018-06-01 02:41:08,442 [main] INFO  org.apache.p
(128788 ms)
You have new mail in /var/spool/mail/acadgild
```

Sub task4



```
[acadgild@localhost ~]$ cat task4.pig
A = load '/home/acadgild/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 $1 DESC;
Result = limit F 10;
dump Result;
[acadgild@localhost ~]$
```

Output



```
2018-06-01 02:44:23,028 [main
d, use fs.defaultFS
2018-06-01 02:44:23,029 [main
2018-06-01 02:44:23,139 [main
2018-06-01 02:44:23,139 [main
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)
2018-06-01 02:44:23,733 [main
(95710 ms)
You have new mail in /var/spo
```