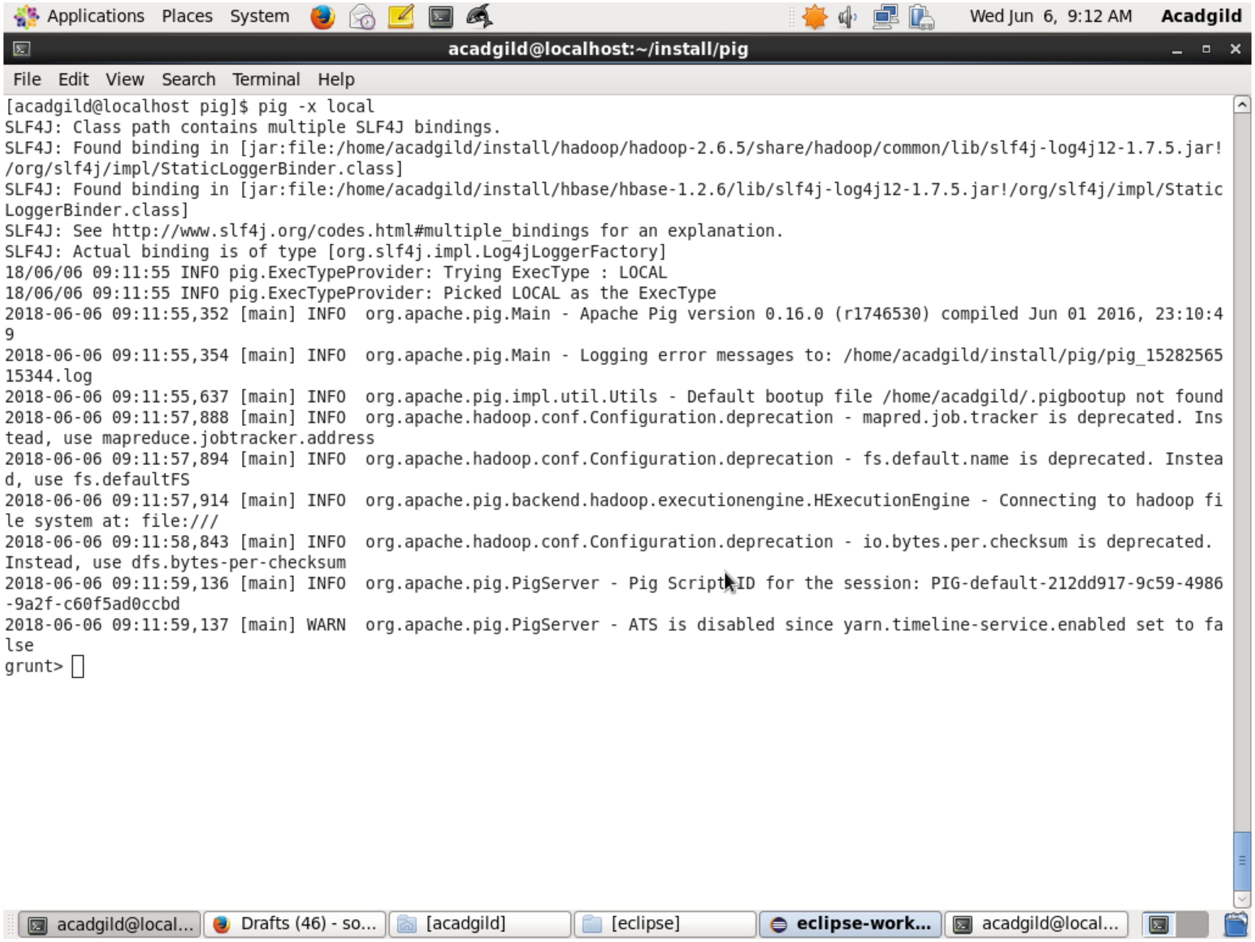
Assignment 7

Common Step:

- **Start Pig in Local**
  Command:
  Pig -x local



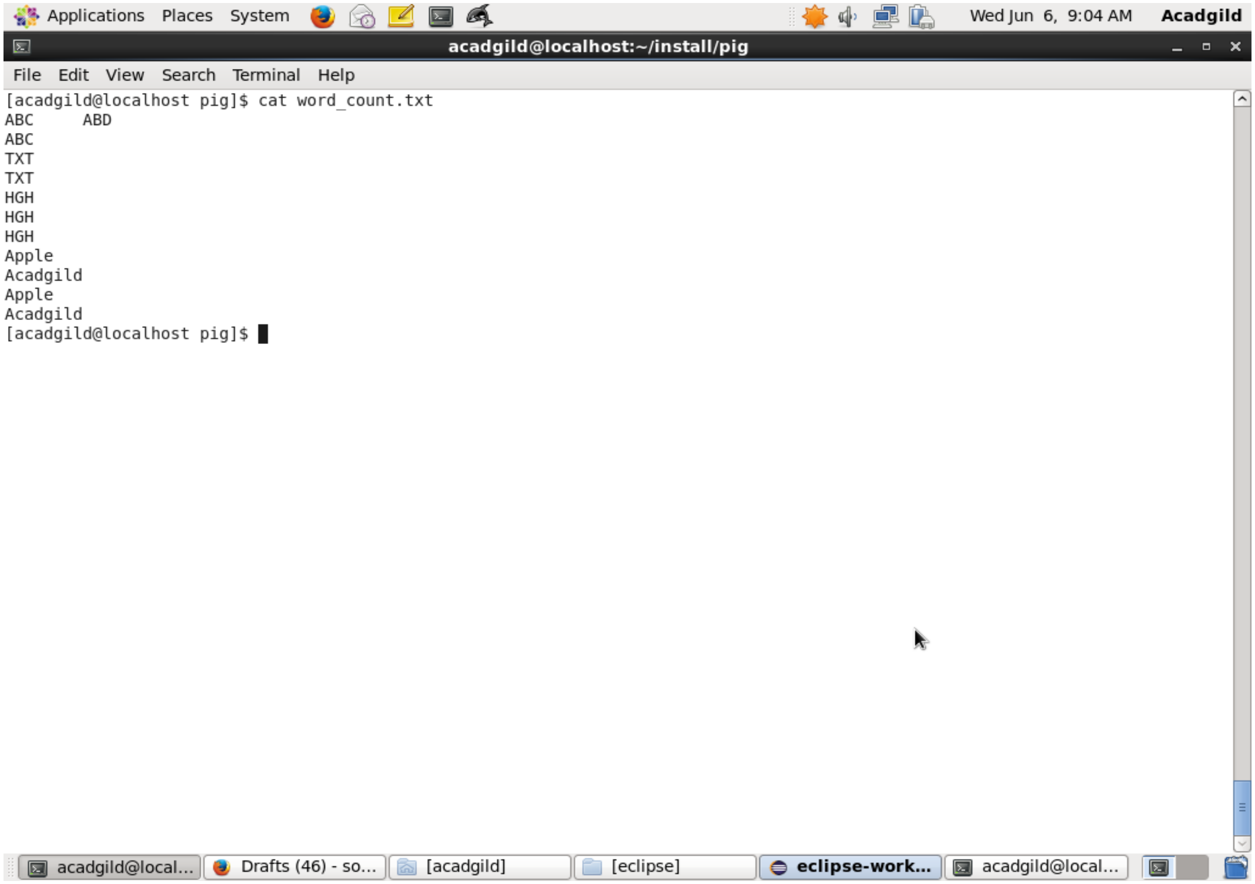
Applications Places System
Wed Jun 6, 9:12 AM Acadgild
acadgild@localhost:~/install/pig
File Edit View Search Terminal Help
acadgild@local... Drafts (46) - so... [acadgild] [eclipse] eclipse-work... acadgild@local...

```
[acadgild@localhost pig]$ 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]
18/06/06 09:11:55 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/06/06 09:11:55 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-06-06 09:11:55,352 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:4
9
2018-06-06 09:11:55,354 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/install/pig/pig_15282565
15344.log
2018-06-06 09:11:55,637 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-06-06 09:11:57,888 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2018-06-06 09:11:57,894 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-06 09:11:57,914 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
le system at: file:///
2018-06-06 09:11:58,843 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-06 09:11:59,136 [main] INFO  org.apache.pig.PigServer - Pig ScriptID for the session: PIG-default-212dd917-9c59-4986
-9a2f-c60f5ad0ccbd
2018-06-06 09:11:59,137 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to fa
lse
grunt> 
```

Task 1:
Write a program to implement wordcount using Pig.

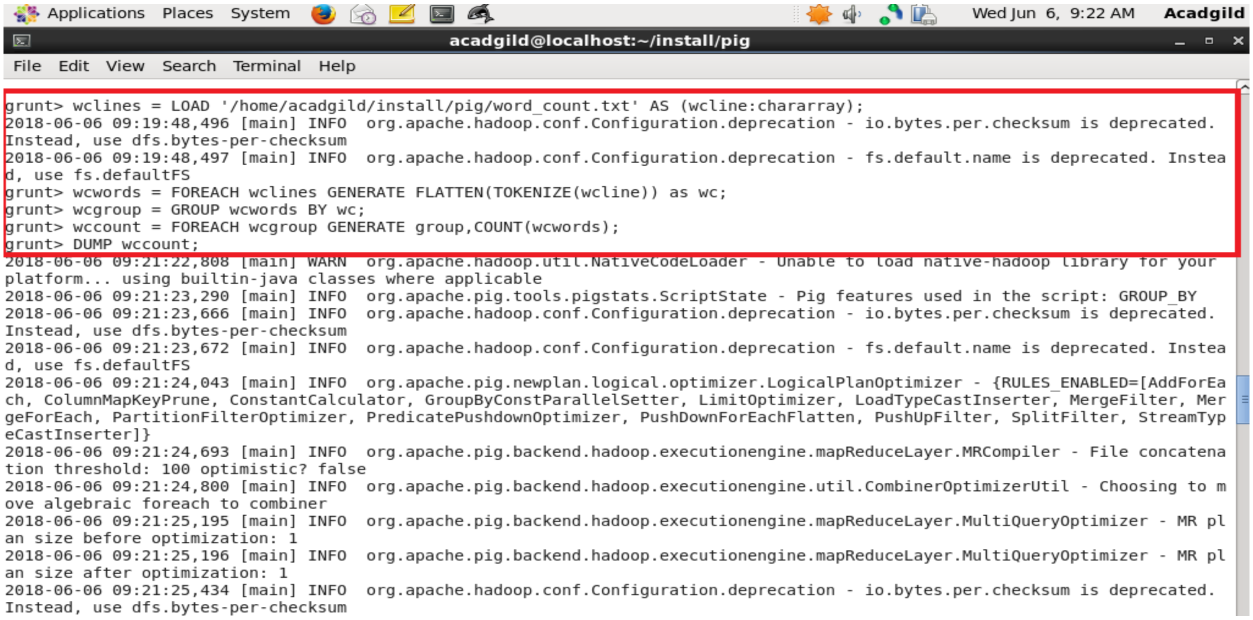
Step 1:

- **Create word_count.txt in local File system.**
  Command:
  Vi word_count.txt
- **Display the contents of word_count.txt**
  Command:
  Cat word_count.txt

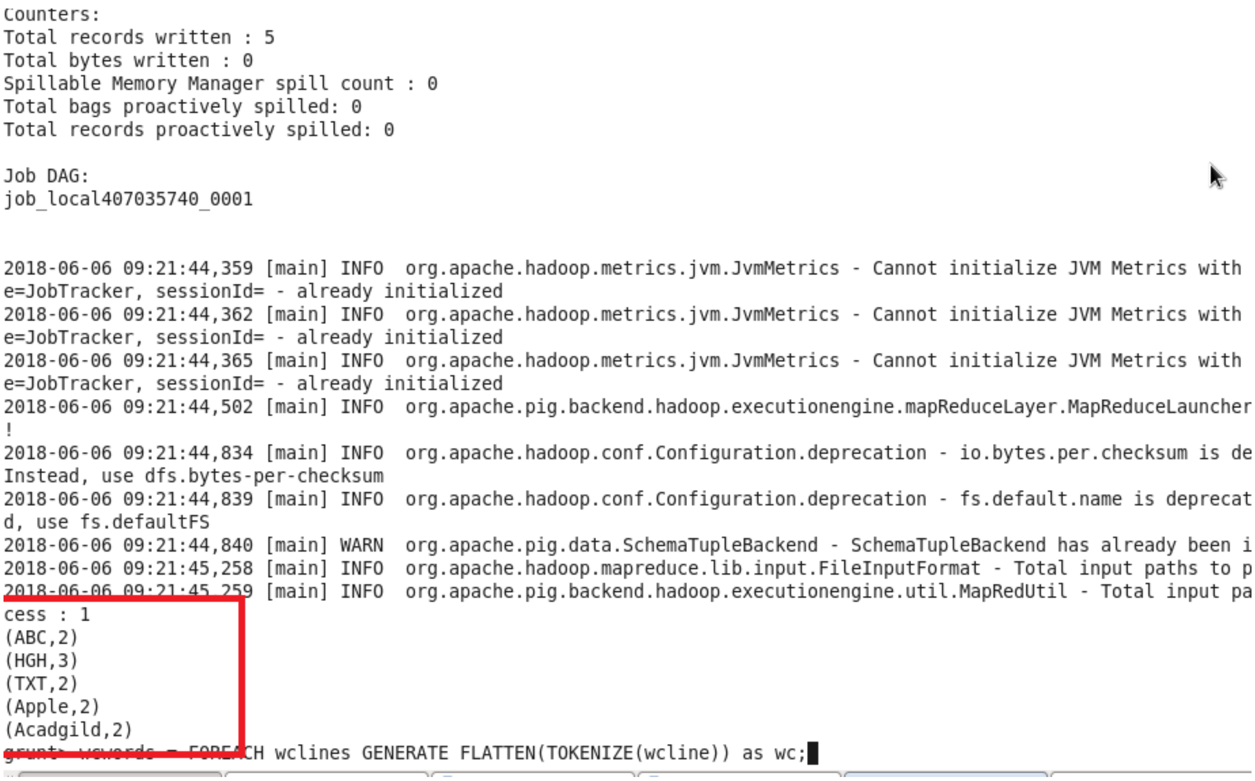

Applications Places System
Wed Jun 6, 9:04 AM Acadgild
acadgild@localhost:~/install/pig
File Edit View Search Terminal Help
acadgild@local... Drafts (46) - so... [acadgild] [eclipse] eclipse-work... acadgild@local...

```
[acadgild@localhost pig]$ cat word_count.txt
ABC     ABD
ABC
TXT
TXT
HGH
HGH
HGH
Apple
Acadgild
Apple
Acadgild
[acadgild@localhost pig]$
```

Step 2:


Applications Places System
Wed Jun 6, 9:22 AM Acadgild
acadgild@localhost:~/install/pig
File Edit View Search Terminal Help

```
grunt> wclines = LOAD '/home/acadgild/install/pig/word_count.txt' AS (wcline:chararray);
2018-06-06 09:19:48,496 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-06 09:19:48,497 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> wcwords = FOREACH wclines GENERATE FLATTEN(TOKENIZE(wcline)) as wc;
grunt> wcgroup = GROUP wcwords BY wc;
grunt> wccount = FOREACH wcgroup GENERATE group,COUNT(wcwords);
grunt> DUMP wccount;
2018-06-06 09:21:22,808 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
2018-06-06 09:21:23,290 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2018-06-06 09:21:23,666 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-06 09:21:23,672 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-06 09:21:24,043 [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]}
2018-06-06 09:21:24,693 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatena
tion threshold: 100 optimistic? false
2018-06-06 09:21:24,800 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to m
ove algebraic foreach to combiner
2018-06-06 09:21:25,195 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size before optimization: 1
2018-06-06 09:21:25,196 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size after optimization: 1
2018-06-06 09:21:25,434 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
```



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

Job DAG:
job_local407035740_0001


2018-06-06 09:21:44,359 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with
e=JobTracker, sessionId= - already initialized
2018-06-06 09:21:44,362 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with
e=JobTracker, sessionId= - already initialized
2018-06-06 09:21:44,365 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with
e=JobTracker, sessionId= - already initialized
2018-06-06 09:21:44,502 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher
!
2018-06-06 09:21:44,834 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is de
Instead, use dfs.bytes-per-checksum
2018-06-06 09:21:44,839 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecat
d, use fs.defaultFS
2018-06-06 09:21:44,840 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been i
2018-06-06 09:21:45,258 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to p
2018-06-06 09:21:45,259 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input pa
cess : 1
(ABC,2)
(HGH,3)
(TXT,2)
(Apple,2)
(Acadgild,2)
grunt> wcwords = FOREACH wclines GENERATE FLATTEN(TOKENIZE(wcline)) as wc;
```

Code:

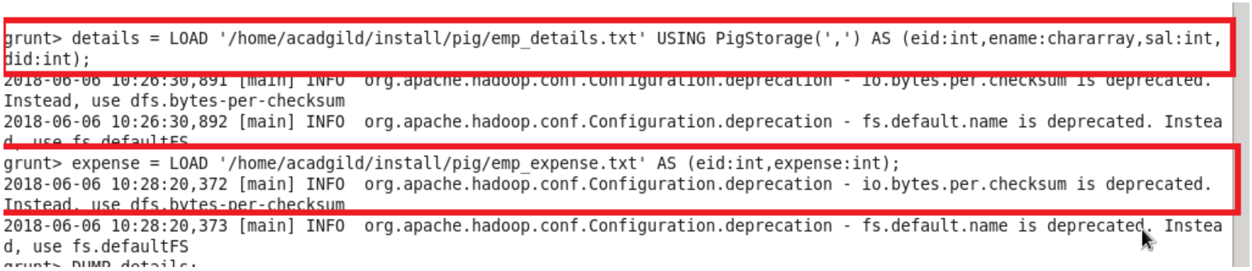
```
Wclines = LOAD '/home/acadgild/install/pig/word_count.txt' AS (wcline:chararray);
Wcwords = FOREACH wclines GENERATE FLATTEN(TOKENIZE(wcline)) as wc;
Wcgroup = GROUP wcwords BY wc;
Wccount = FOREACH wcgroup GENERATE group,COUNT(wcwords);
DUMP wccount;
```

Task 2:

Employee_details and Employee_expense

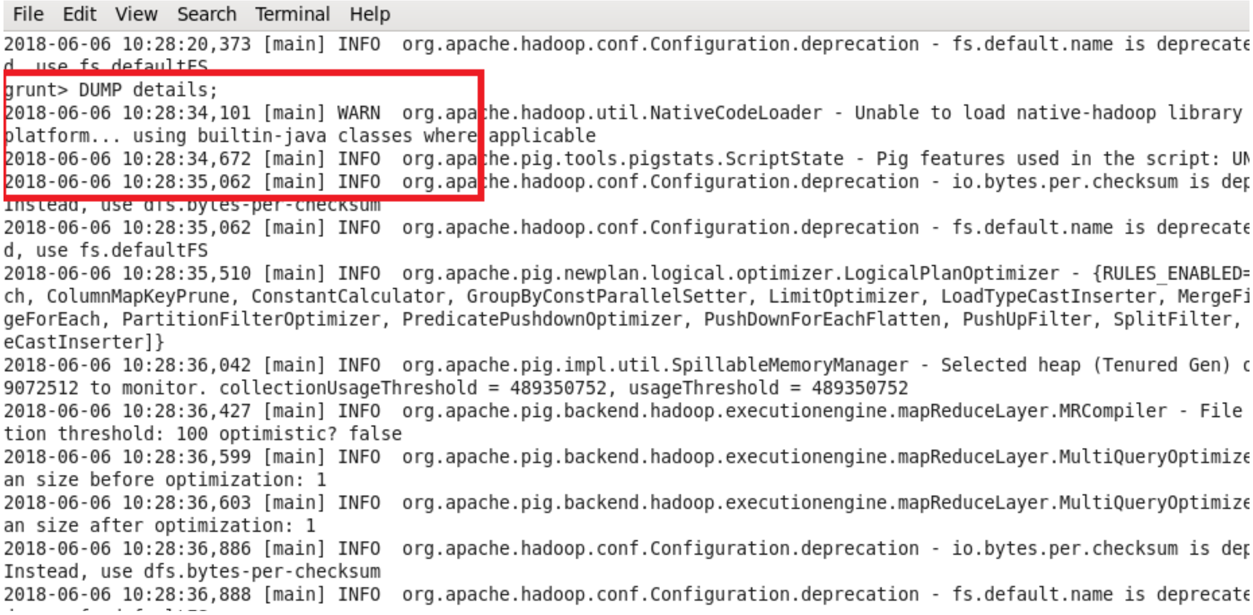
Step 1:

Load Employee details and employee expense

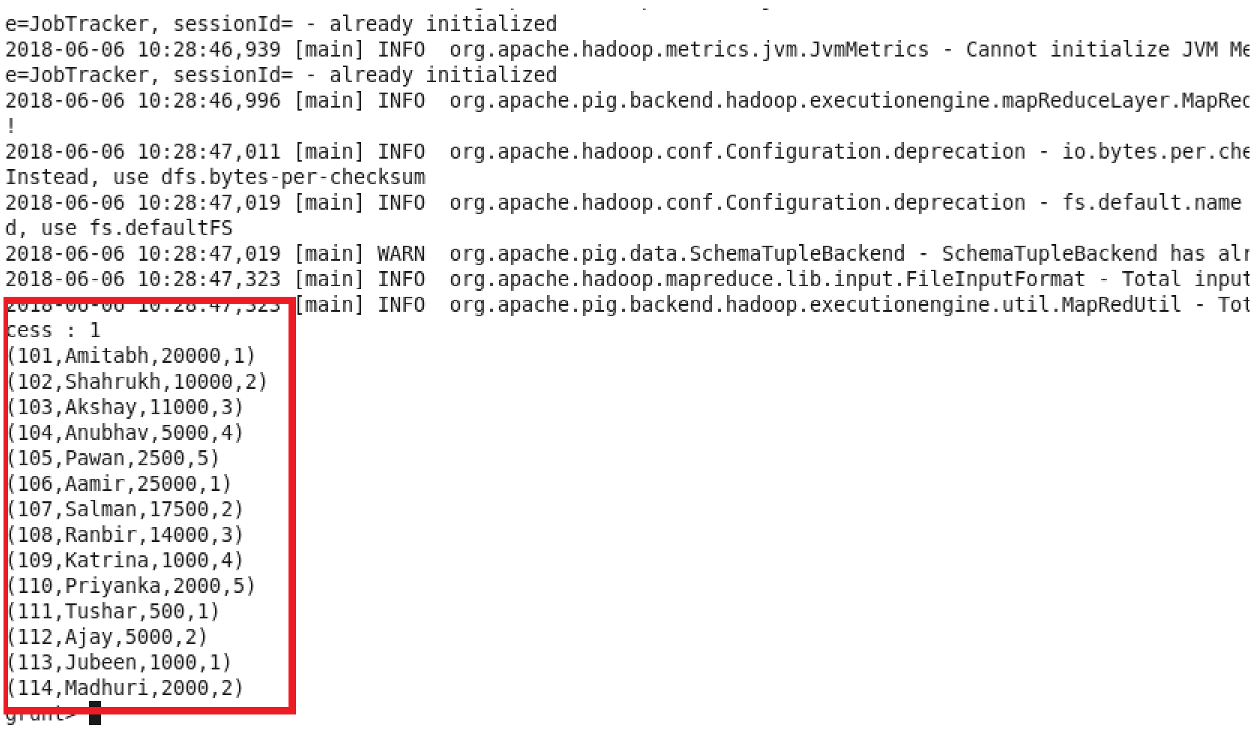


```
grunt> details = LOAD '/home/acadgild/install/pig/emp_details.txt' USING PigStorage(',') AS (eid:int,ename:chararray,sal:int,
did:int);
2018-06-06 10:26:30,891 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-06 10:26:30,892 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> expense = LOAD '/home/acadgild/install/pig/emp_expense.txt' AS (eid:int,expense:int);
2018-06-06 10:28:20,372 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-06 10:28:20,373 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> DUMP details;
```

Dump employee details

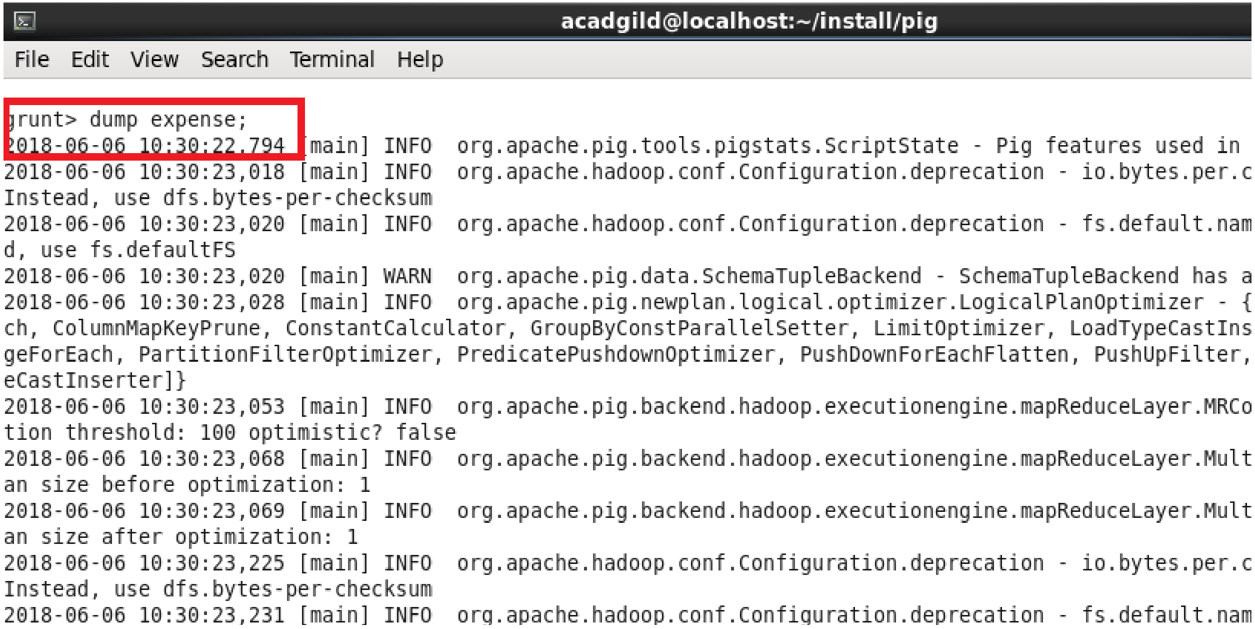


```
2018-06-06 10:28:20,373 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecate
d, use fs.defaultFS
grunt> DUMP details;
2018-06-06 10:28:34,101 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library
platform... using builtin-java classes where applicable
2018-06-06 10:28:34,672 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UN
2018-06-06 10:28:35,062 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is dep
Instead, use dfs.bytes-per-checksum
2018-06-06 10:28:35,062 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecate
d, use fs.defaultFS
2018-06-06 10:28:35,510 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=
ch, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFi
geForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter,
eCastInserter]}
2018-06-06 10:28:36,042 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) c
9072512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 489350752
2018-06-06 10:28:36,427 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File
tion threshold: 100 optimistic? false
2018-06-06 10:28:36,599 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimize
an size before optimization: 1
2018-06-06 10:28:36,603 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimize
an size after optimization: 1
2018-06-06 10:28:36,886 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is dep
Instead, use dfs.bytes-per-checksum
2018-06-06 10:28:36,888 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecate
```



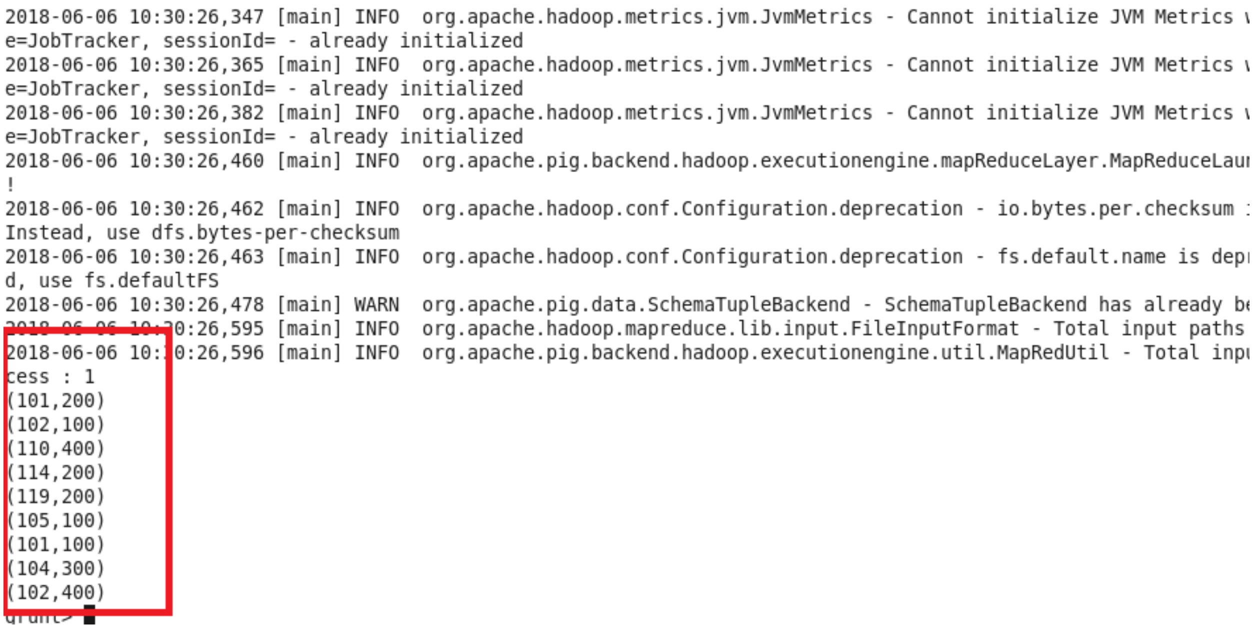
```
e=JobTracker, sessionId= - already initialized
2018-06-06 10:28:46,939 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Me
e=JobTracker, sessionId= - already initialized
2018-06-06 10:28:46,996 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRec
!
2018-06-06 10:28:47,011 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.che
Instead, use dfs.bytes-per-checksum
2018-06-06 10:28:47,019 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name
d, use fs.defaultFS
2018-06-06 10:28:47,019 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has alr
2018-06-06 10:28:47,323 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input
2018-06-06 10:28:47,323 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Tot
cess : 1
(101,Amitabh,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,Ranbir,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)
grunt> 
```

Dump employee expense



acadgild@localhost:~/install/pig

```
grunt> dump expense;
2018-06-06 10:30:22,794 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in
2018-06-06 10:30:23,018 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.c
Instead, use dfs.bytes-per-checksum
2018-06-06 10:30:23,020 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.nam
d, use fs.defaultFS
2018-06-06 10:30:23,020 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has a
2018-06-06 10:30:23,028 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {
ch, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastIns
geForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter,
eCastInserter]}
2018-06-06 10:30:23,053 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCo
tion threshold: 100 optimistic? false
2018-06-06 10:30:23,068 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Mult
an size before optimization: 1
2018-06-06 10:30:23,069 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Mult
an size after optimization: 1
2018-06-06 10:30:23,225 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.c
Instead, use dfs.bytes-per-checksum
2018-06-06 10:30:23,231 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.nam
```



```
2018-06-06 10:30:26,347 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
e=JobTracker, sessionId= - already initialized
2018-06-06 10:30:26,365 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
e=JobTracker, sessionId= - already initialized
2018-06-06 10:30:26,382 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
e=JobTracker, sessionId= - already initialized
2018-06-06 10:30:26,460 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLau
!
2018-06-06 10:30:26,462 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum i
Instead, use dfs.bytes-per-checksum
2018-06-06 10:30:26,463 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depr
d, use fs.defaultFS
2018-06-06 10:30:26,478 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already be
2018-06-06 10:30:26,595 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths
2018-06-06 10:30:26,596 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total inpu
cess : 1
(101,200)
(102,100)
(110,400)
(114,200)
(119,200)
(105,100)
(101,100)
(104,300)
(102,400)
grunt>
```

Code:

```
Details = LOAD '/home/acadgild/install/pig/emp_details.txt' USING PigStorage(',') AS
(eid:int,ename:chararray,sal:int,did:int);

expense = LOAD '/home/acadgild/install/pig/emp_expense.txt' USING PigStorage(',') AS
(eid:int,expense:int);

DUMP Details;
DUMP expense;
```

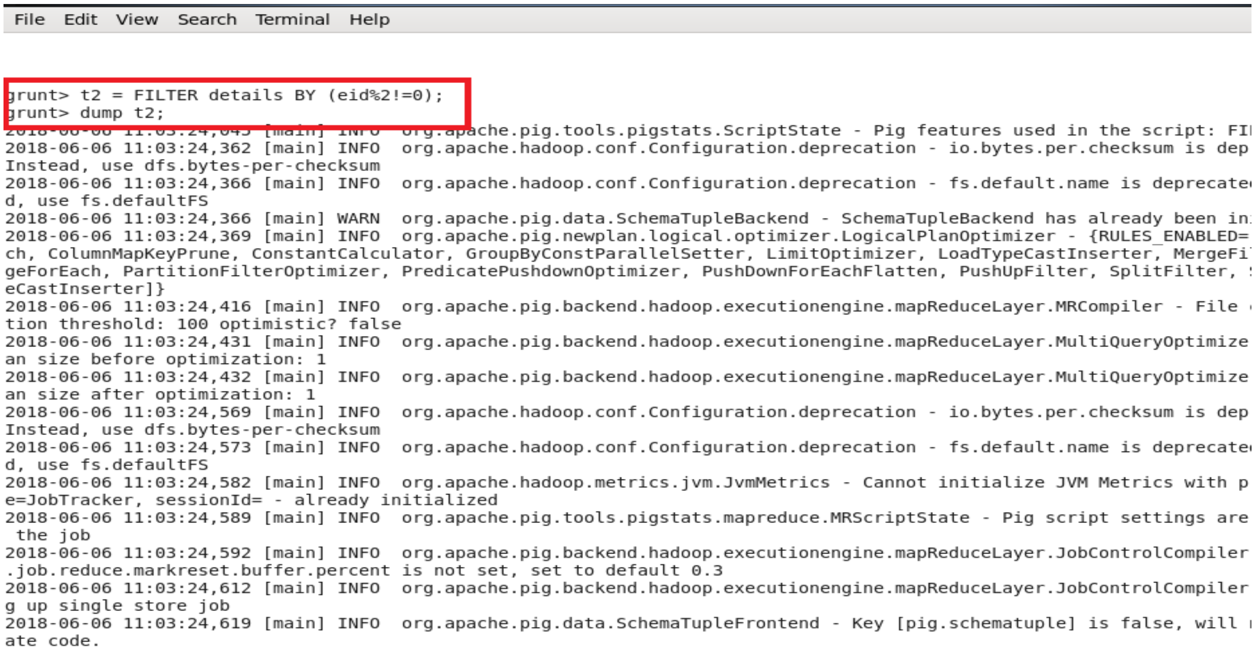
Step 2:

Top 5 Employees with highest Rating:

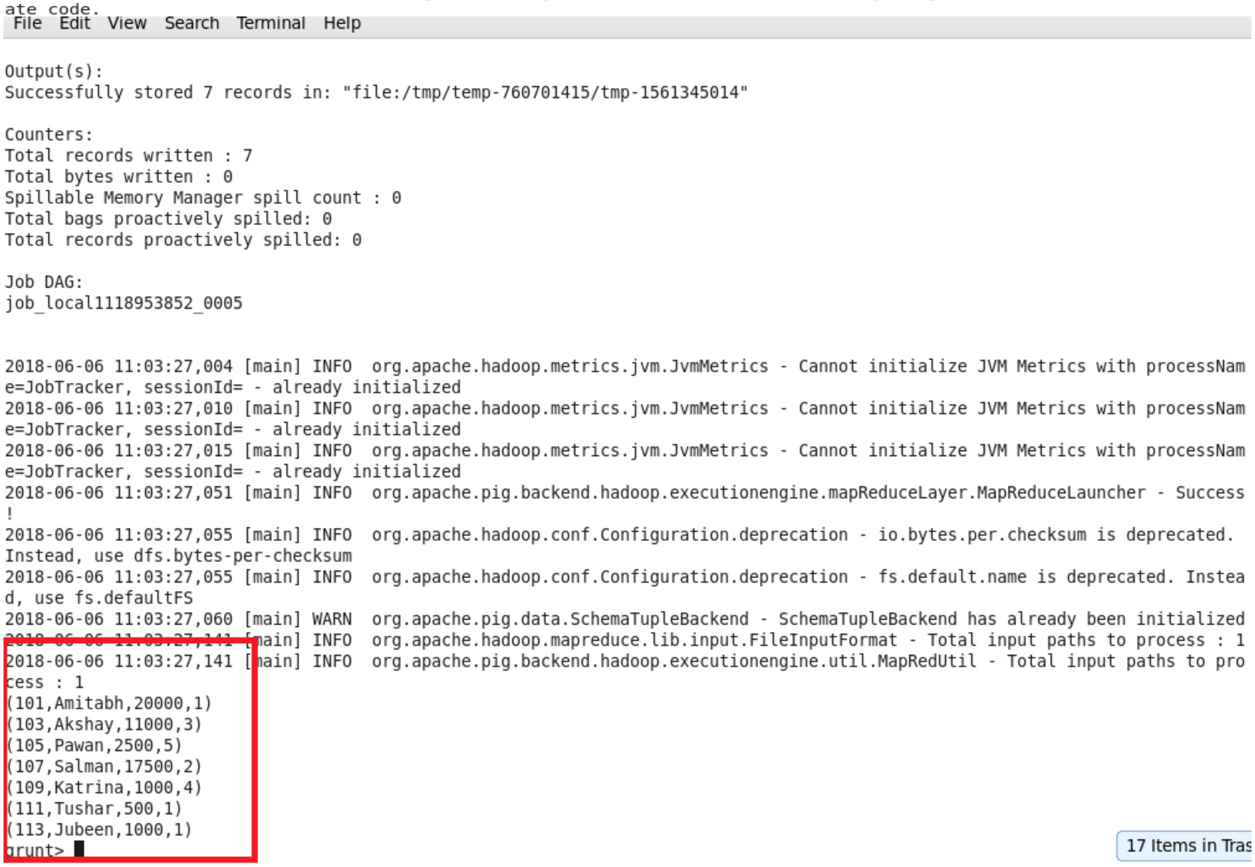
Rating Field is not present in the provided data set.

Step 3:

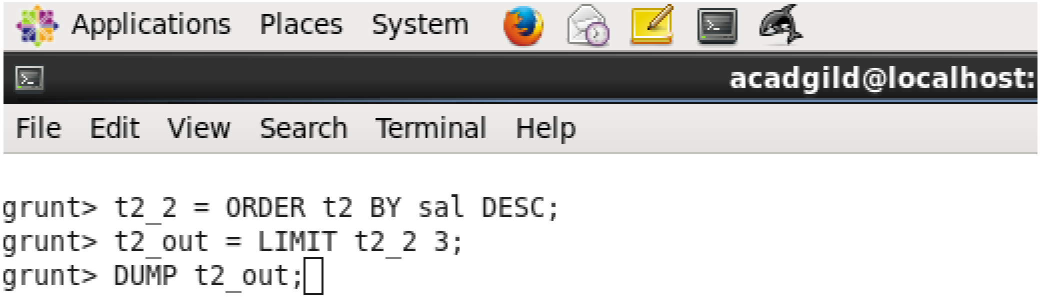
Top 3 employees with highest salary whose employee id is an odd number:



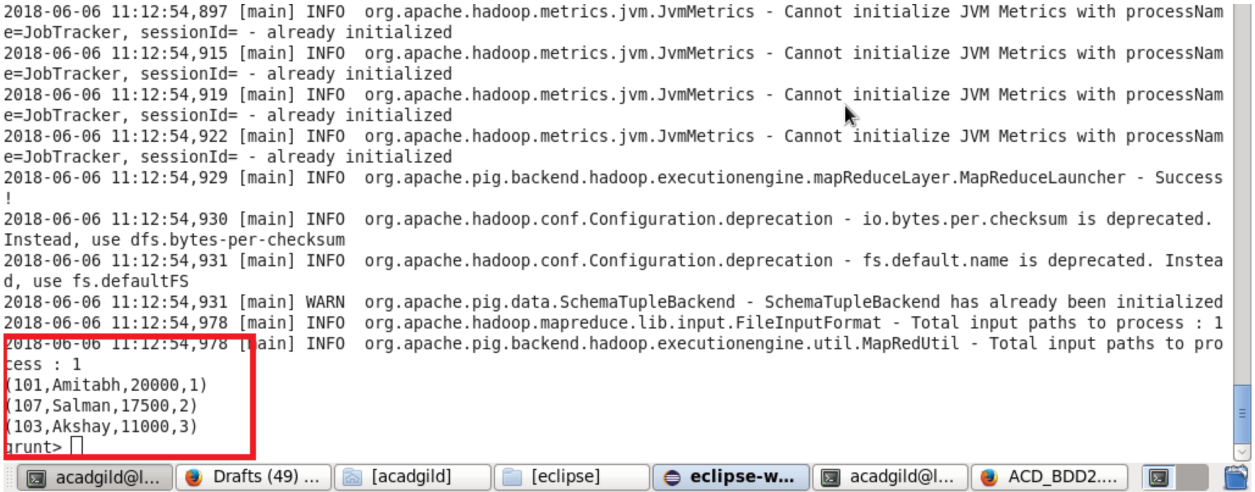
```
grunt> t2 = FILTER details BY (eid%2!=0);
grunt> dump t2;
2018-06-06 11:03:24,345 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FI
2018-06-06 11:03:24,362 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is dep
Instead, use dfs.bytes-per-checksum
2018-06-06 11:03:24,366 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecate
d, use fs.defaultFS
2018-06-06 11:03:24,366 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been in
2018-06-06 11:03:24,369 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=
ch, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFi
geForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, 
eCastInserter]}
2018-06-06 11:03:24,416 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File 
tion threshold: 100 optimistic? false
2018-06-06 11:03:24,431 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimize
an size before optimization: 1
2018-06-06 11:03:24,432 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimize
an size after optimization: 1
2018-06-06 11:03:24,569 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is dep
Instead, use dfs.bytes-per-checksum
2018-06-06 11:03:24,573 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecate
d, use fs.defaultFS
2018-06-06 11:03:24,582 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with p
e=JobTracker, sessionId= - already initialized
2018-06-06 11:03:24,589 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are
 the job
2018-06-06 11:03:24,592 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-06-06 11:03:24,612 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
g up single store job
2018-06-06 11:03:24,619 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will 
ate code.
```



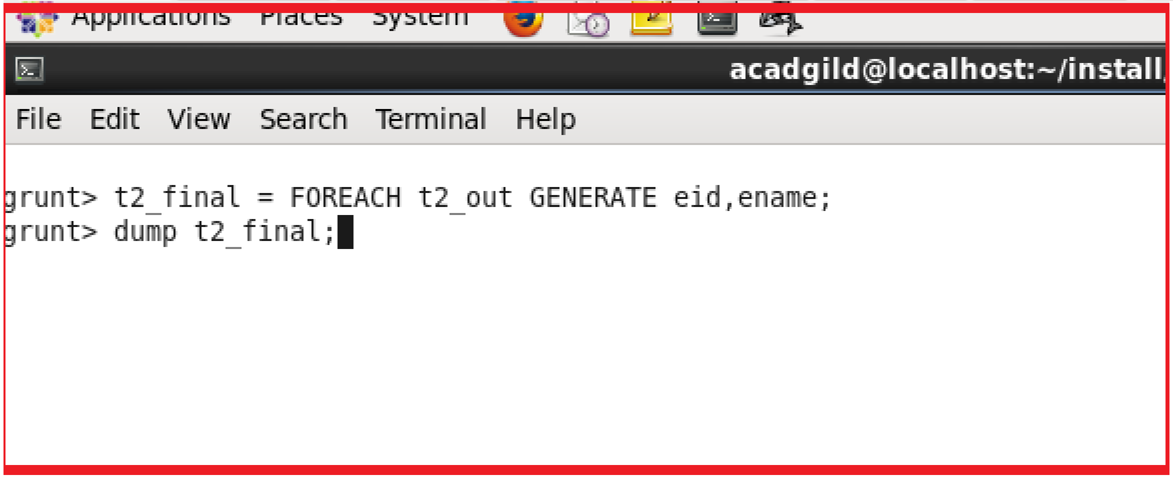
```
Output(s):
Successfully stored 7 records in: "file:/tmp/temp-760701415/tmp-1561345014"

Counters:
Total records written : 7
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1118953852_0005


2018-06-06 11:03:27,004 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-06 11:03:27,010 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-06 11:03:27,015 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-06 11:03:27,051 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-06-06 11:03:27,055 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-06 11:03:27,055 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-06 11:03:27,060 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-06 11:03:27,141 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-06 11:03:27,141 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(101,Amitabh,20000,1)
(103,Akshay,11000,3)
(105,Pawan,2500,5)
(107,Salman,17500,2)
(109,Katrina,1000,4)
(111,Tushar,500,1)
(113,Jubeen,1000,1)
grunt> 
```


Applications Places System
acadgild@localhost:
File Edit View Search Terminal Help

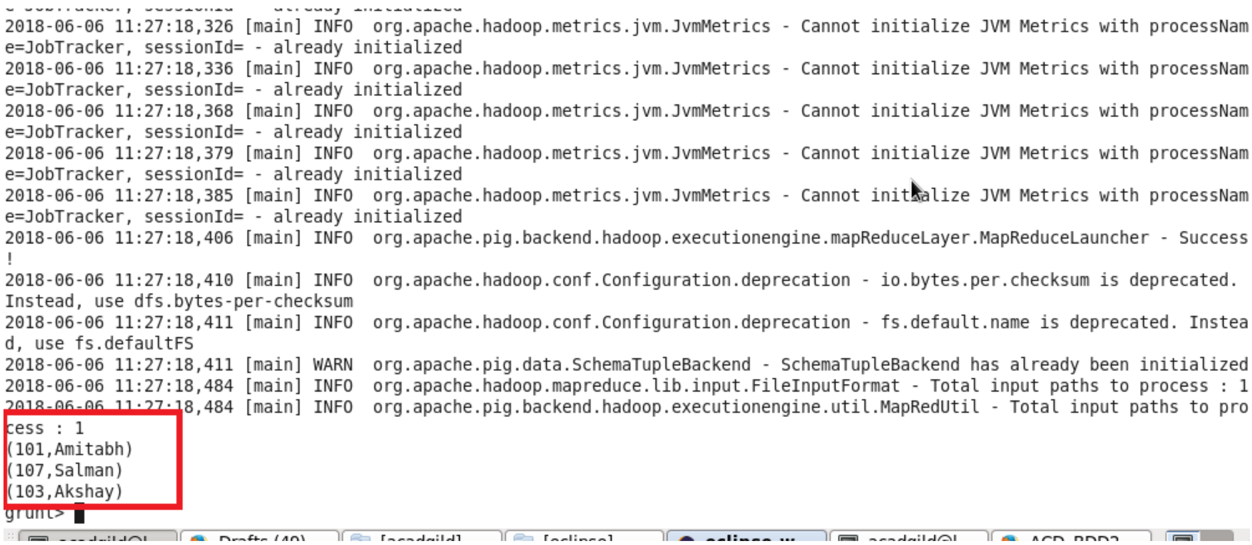
```
grunt> t2_2 = ORDER t2 BY sal DESC;
grunt> t2_out = LIMIT t2_2 3;
grunt> DUMP t2_out;
```



```
2018-06-06 11:12:54,897 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-06 11:12:54,915 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-06 11:12:54,919 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-06 11:12:54,922 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-06 11:12:54,929 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-06-06 11:12:54,930 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-06 11:12:54,931 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-06 11:12:54,931 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-06 11:12:54,978 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-06 11:12:54,978 [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>
```


Applications Places System
acadgild@localhost:~/install
File Edit View Search Terminal Help

```
grunt> t2_final = FOREACH t2_out GENERATE eid,ename;
grunt> dump t2_final;
```



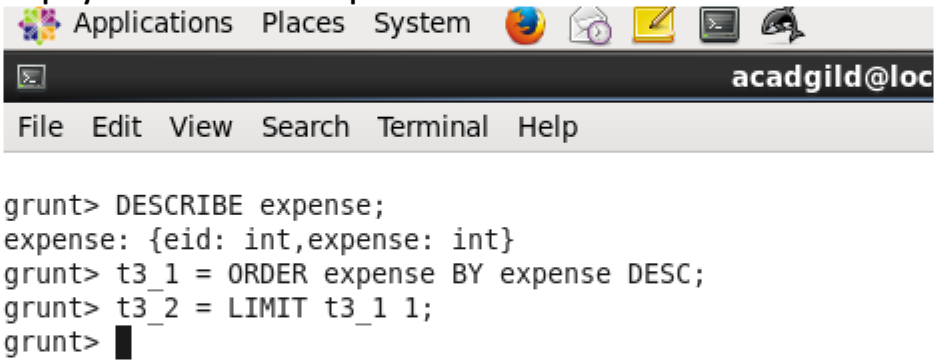
```
e=JobTracker, sessionId= - already initialized
2018-06-06 11:27:18,326 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-06 11:27:18,336 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-06 11:27:18,368 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-06 11:27:18,379 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-06 11:27:18,385 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-06 11:27:18,406 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-06-06 11:27:18,410 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-06 11:27:18,411 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-06 11:27:18,411 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-06 11:27:18,484 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-06 11:27:18,484 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(101,Amitabh)
(107,Salman)
(103,Akshay)
grunt> 
```

Code:

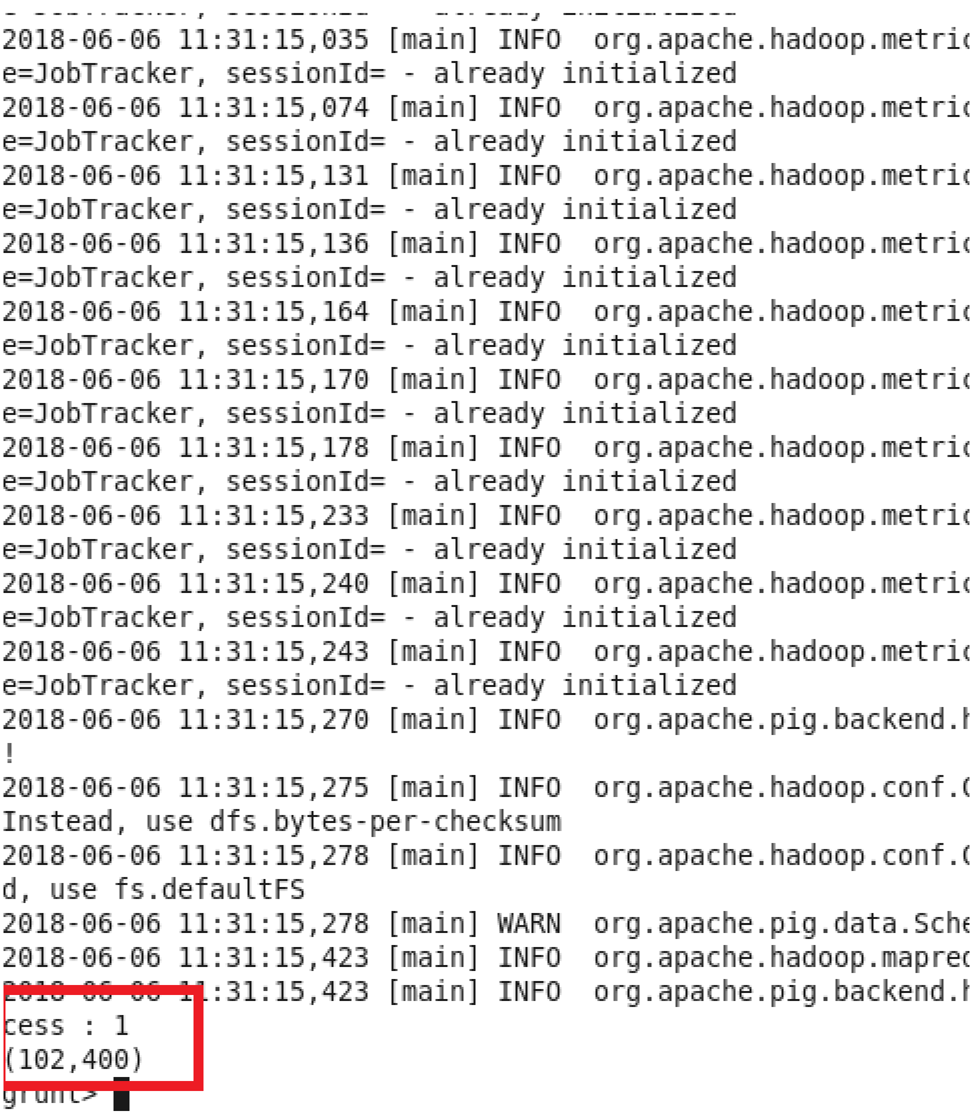
T2 = FILTER details BY (eid%2!=0);
DUMP t2;
T2_2 = ORDER t2 BY sal DESC;
T2_out = LIMIT t2_2 3;
DUMP t2_out;
T2_final = FOREACH t2_out GENERATE eid,ename ;
DUMP t2_final;

Step4:

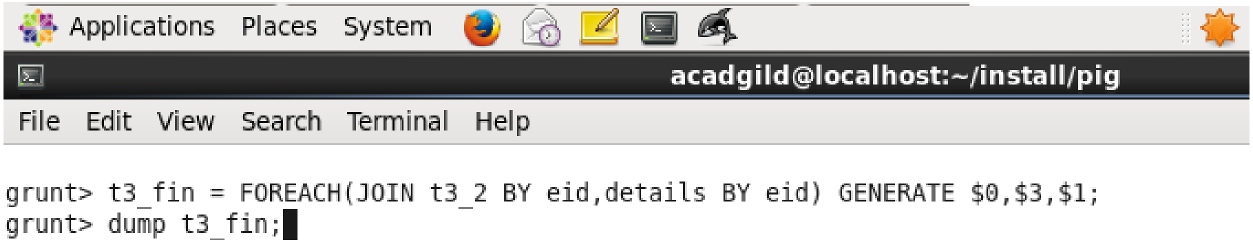
Employee with maximum expense:


Applications Places System
acadgild@loc
File Edit View Search Terminal Help

```
grunt> DESCRIBE expense;
expense: {eid: int,expense: int}
grunt> t3_1 = ORDER expense BY expense DESC;
grunt> t3_2 = LIMIT t3_1 1;
grunt> 
```

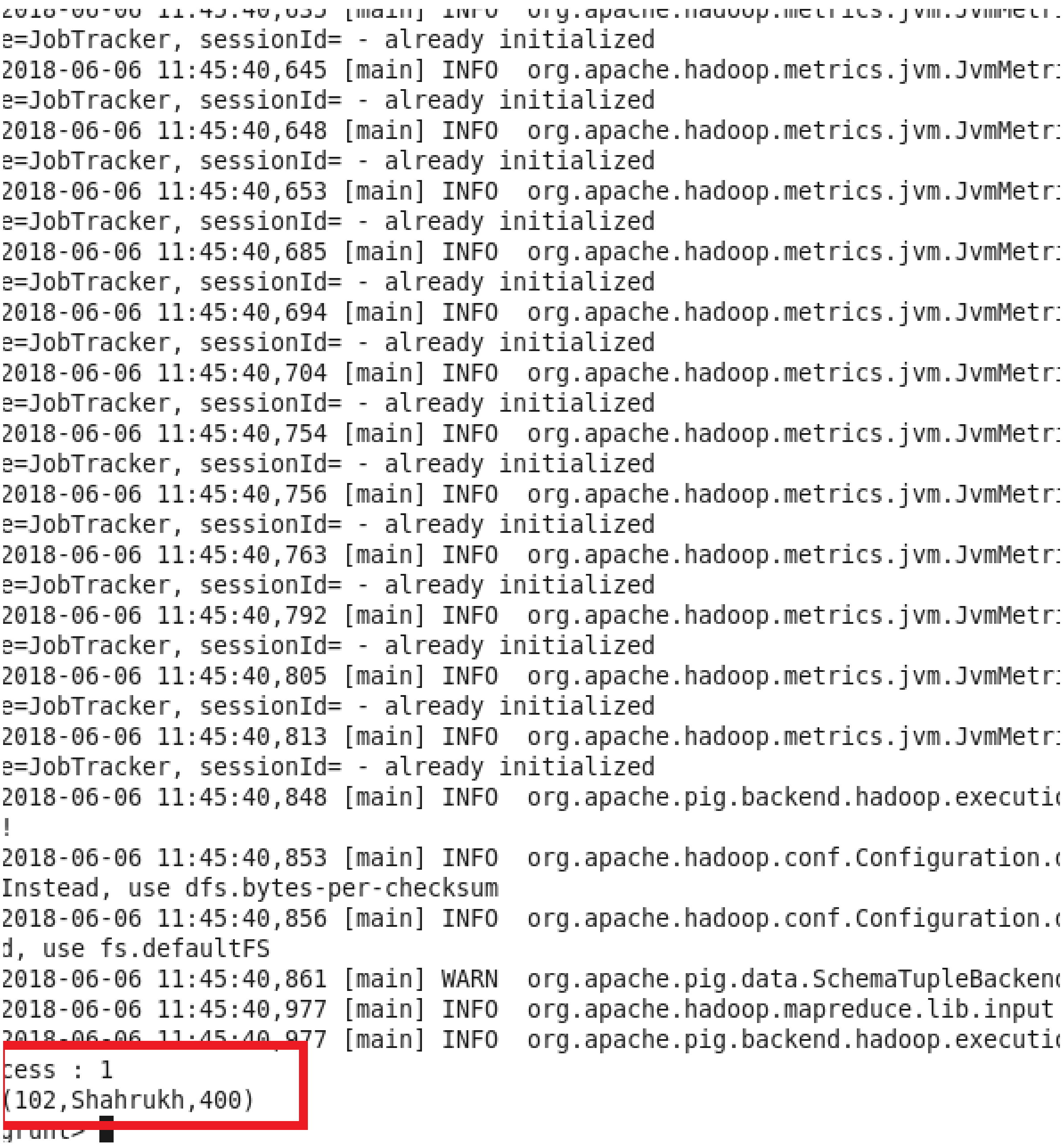


```
2018-06-06 11:31:15,035 [main] INFO  org.apache.hadoop.metric
e=JobTracker, sessionId= - already initialized
2018-06-06 11:31:15,074 [main] INFO  org.apache.hadoop.metric
e=JobTracker, sessionId= - already initialized
2018-06-06 11:31:15,131 [main] INFO  org.apache.hadoop.metric
e=JobTracker, sessionId= - already initialized
2018-06-06 11:31:15,136 [main] INFO  org.apache.hadoop.metric
e=JobTracker, sessionId= - already initialized
2018-06-06 11:31:15,164 [main] INFO  org.apache.hadoop.metric
e=JobTracker, sessionId= - already initialized
2018-06-06 11:31:15,170 [main] INFO  org.apache.hadoop.metric
e=JobTracker, sessionId= - already initialized
2018-06-06 11:31:15,178 [main] INFO  org.apache.hadoop.metric
e=JobTracker, sessionId= - already initialized
2018-06-06 11:31:15,233 [main] INFO  org.apache.hadoop.metric
e=JobTracker, sessionId= - already initialized
2018-06-06 11:31:15,240 [main] INFO  org.apache.hadoop.metric
e=JobTracker, sessionId= - already initialized
2018-06-06 11:31:15,243 [main] INFO  org.apache.hadoop.metric
e=JobTracker, sessionId= - already initialized
2018-06-06 11:31:15,270 [main] INFO  org.apache.pig.backend.h
!
2018-06-06 11:31:15,275 [main] INFO  org.apache.hadoop.conf.C
Instead, use dfs.bytes-per-checksum
2018-06-06 11:31:15,278 [main] INFO  org.apache.hadoop.conf.C
d, use fs.defaultFS
2018-06-06 11:31:15,278 [main] WARN  org.apache.pig.data.Sche
2018-06-06 11:31:15,423 [main] INFO  org.apache.hadoop.mapred
2018-06-06 11:31:15,423 [main] INFO  org.apache.pig.backend.h
cess : 1
(102,400)
grunt> 
```



Applications Places System
acadgild@localhost:~/install/pig
File Edit View Search Terminal Help

```
grunt> t3_fin = FOREACH(JOIN t3_2 BY eid,details BY eid) GENERATE $0,$3,$1;
grunt> dump t3_fin;
```



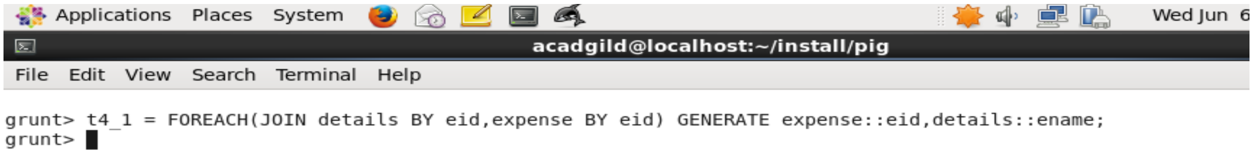
```
2018-06-06 11:45:40,635 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,645 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,648 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,653 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,685 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,694 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,704 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,754 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,756 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,763 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,792 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,805 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,813 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetr:
e=JobTracker, sessionId= - already initialized
2018-06-06 11:45:40,848 [main] INFO  org.apache.pig.backend.hadoop.executi
!
2018-06-06 11:45:40,853 [main] INFO  org.apache.hadoop.conf.Configuration.
Instead, use dfs.bytes-per-checksum
2018-06-06 11:45:40,856 [main] INFO  org.apache.hadoop.conf.Configuration.
d, use fs.defaultFS
2018-06-06 11:45:40,861 [main] WARN  org.apache.pig.data.SchemaTupleBacken
2018-06-06 11:45:40,977 [main] INFO  org.apache.hadoop.mapreduce.lib.input
2018-06-06 11:45:40,977 [main] INFO  org.apache.pig.backend.hadoop.executi
cess : 1
(102,Shahrukh,400)
grunt> 
```

<u>Code:</u>

```
T3_1 = ORDER expense BY expense DESC;
T3_2 = LIMIT t3_1 1;
DUMP T3_2;
T3_fin = FOREACH(JOIN t3_2 BY eid,details BY eid) GENERATE $1,$3,$1;
DUMP t3_fin;
```

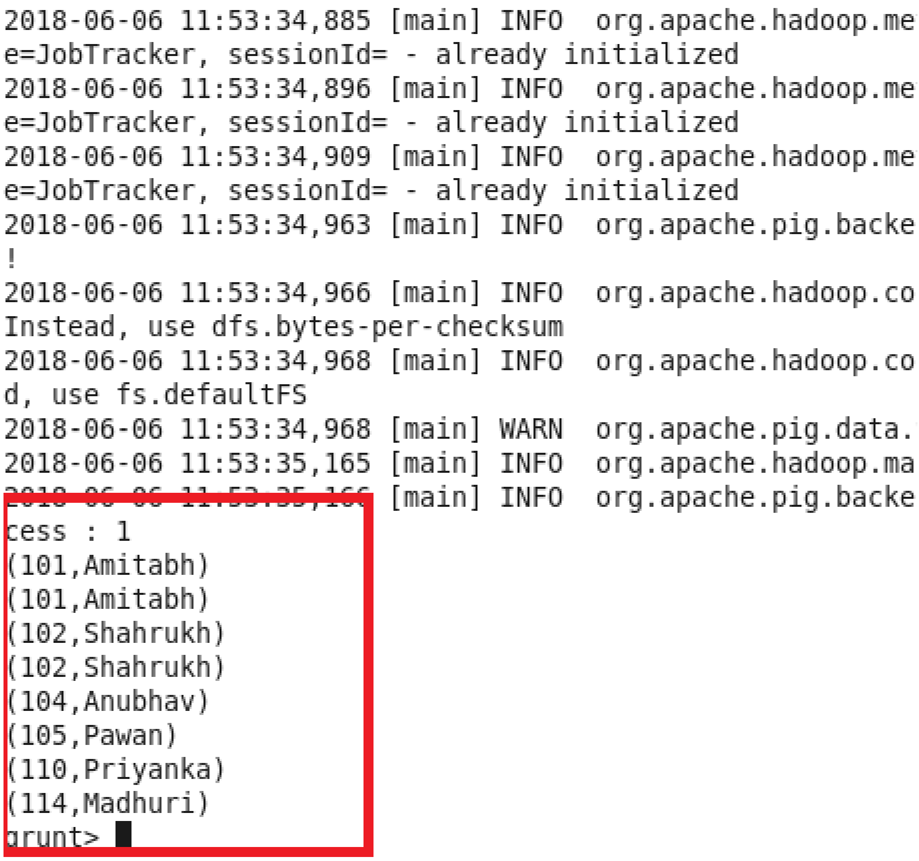
Step 5:

List of employees in Employee expense

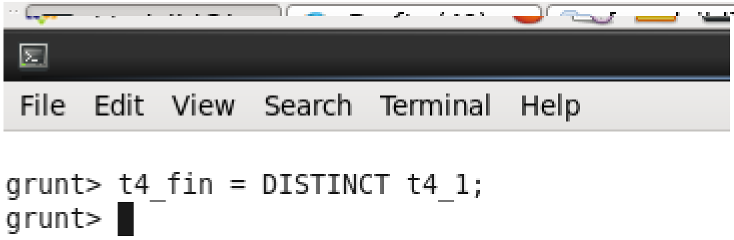


Applications Places System
Wed Jun 6
acadgild@localhost:~/install/pig
File Edit View Search Terminal Help

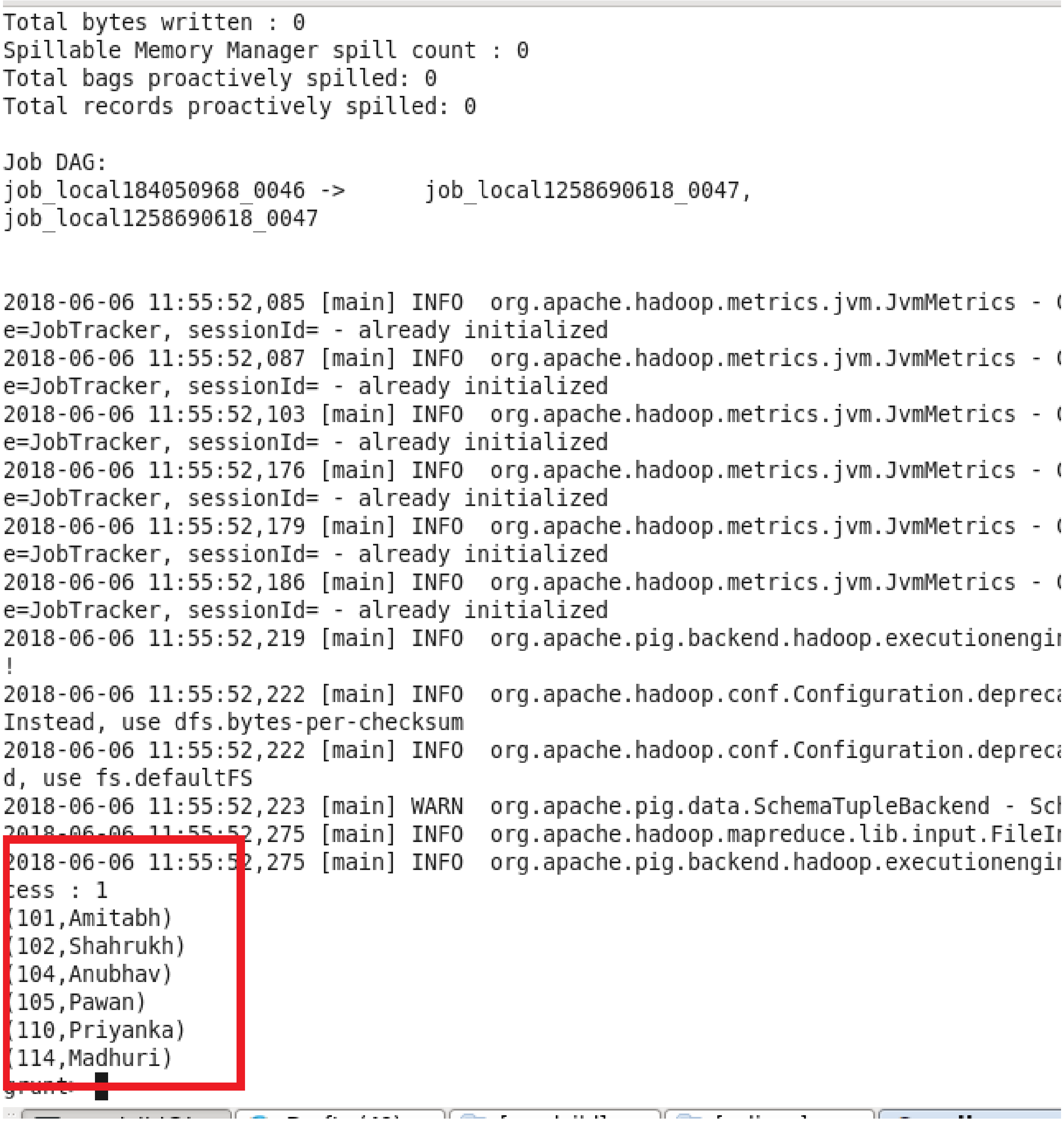
```
grunt> t4_1 = FOREACH(JOIN details BY eid,expense BY eid) GENERATE expense::eid,details::ename;
grunt> 
```



```
2018-06-06 11:53:34,885 [main] INFO  org.apache.hadoop.me
e=JobTracker, sessionId= - already initialized
2018-06-06 11:53:34,896 [main] INFO  org.apache.hadoop.me
e=JobTracker, sessionId= - already initialized
2018-06-06 11:53:34,909 [main] INFO  org.apache.hadoop.me
e=JobTracker, sessionId= - already initialized
2018-06-06 11:53:34,963 [main] INFO  org.apache.pig.backe
!
2018-06-06 11:53:34,966 [main] INFO  org.apache.hadoop.co
Instead, use dfs.bytes-per-checksum
2018-06-06 11:53:34,968 [main] INFO  org.apache.hadoop.co
d, use fs.defaultFS
2018-06-06 11:53:34,968 [main] WARN  org.apache.pig.data.
2018-06-06 11:53:35,165 [main] INFO  org.apache.hadoop.ma
2018-06-06 11:53:35,166 [main] INFO  org.apache.pig.backe
cess : 1
(101,Amitabh)
(101,Amitabh)
(102,Shahrukh)
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
grunt> 
```



```
grunt> t4_fin = DISTINCT t4_1;
grunt> 
```



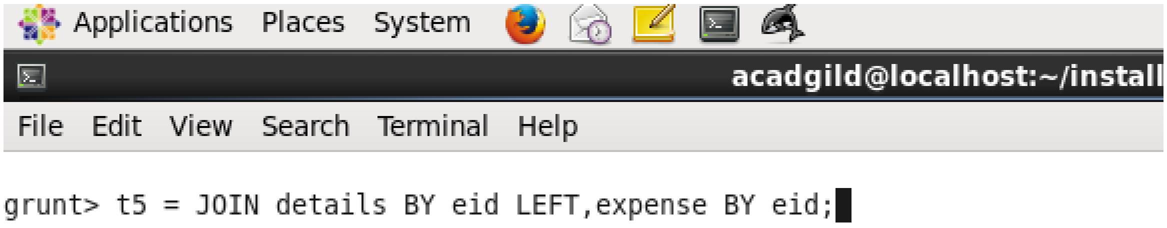
```
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local184050968_0046 ->      job_local1258690618_0047,
job_local1258690618_0047


2018-06-06 11:55:52,085 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - (
e=JobTracker, sessionId= - already initialized
2018-06-06 11:55:52,087 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - (
e=JobTracker, sessionId= - already initialized
2018-06-06 11:55:52,103 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - (
e=JobTracker, sessionId= - already initialized
2018-06-06 11:55:52,176 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - (
e=JobTracker, sessionId= - already initialized
2018-06-06 11:55:52,179 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - (
e=JobTracker, sessionId= - already initialized
2018-06-06 11:55:52,186 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - (
e=JobTracker, sessionId= - already initialized
2018-06-06 11:55:52,219 [main] INFO  org.apache.pig.backend.hadoop.executionengi
!
2018-06-06 11:55:52,222 [main] INFO  org.apache.hadoop.conf.Configuration.depreca
Instead, use dfs.bytes-per-checksum
2018-06-06 11:55:52,222 [main] INFO  org.apache.hadoop.conf.Configuration.depreca
d, use fs.defaultFS
2018-06-06 11:55:52,223 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sch
2018-06-06 11:55:52,275 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileIn
2018-06-06 11:55:52,275 [main] INFO  org.apache.pig.backend.hadoop.executionengi
cess : 1
(101,Amitabh)
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
grunt>
```

Code:

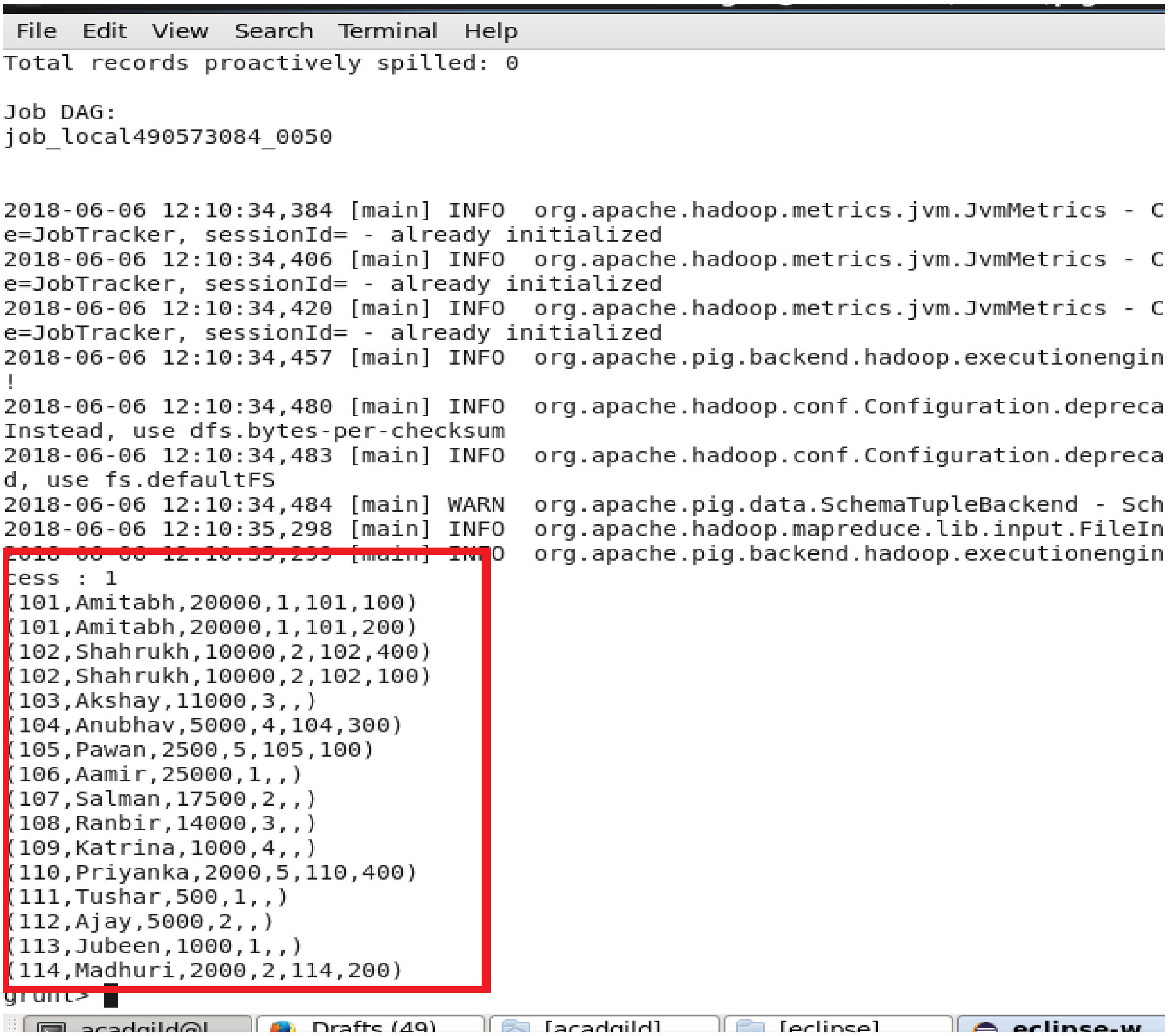
```
T4_1 = FOREACH(JOIN details BY eid,expense BY eid) GENERATE expense::eid,details::ename;
DUMP T4_1;
T4_fin = DISTINCT t4_1;
DUMP T4_fin;
```

Step6:

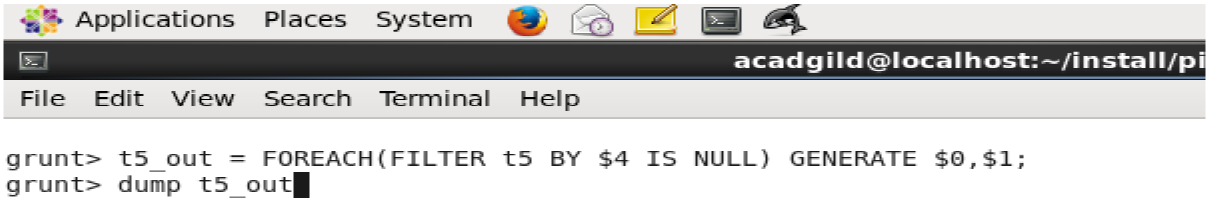
List of Employees not in Expense:



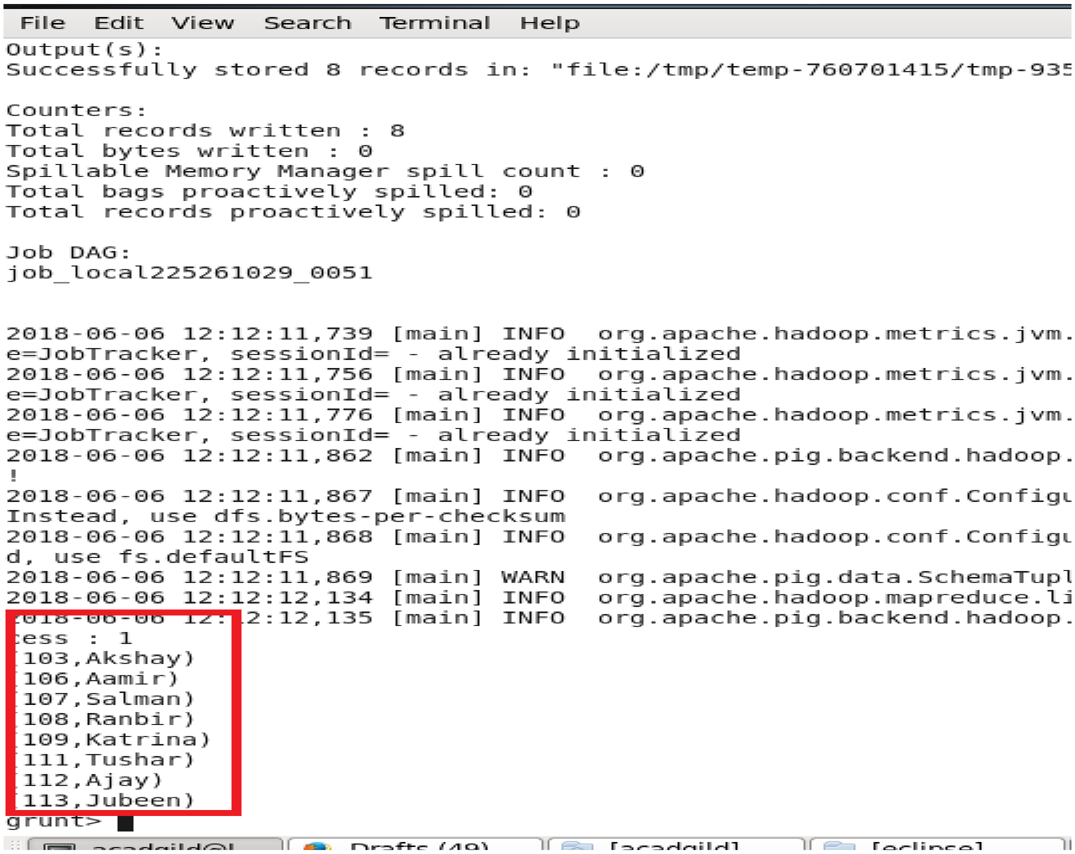
```
grunt> t5 = JOIN details BY eid LEFT,expense BY eid;
```



```
Total records proactively spilled: 0

Job DAG:
job_local490573084_0050


2018-06-06 12:10:34,384 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - C
e=JobTracker, sessionId= - already initialized
2018-06-06 12:10:34,406 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - C
e=JobTracker, sessionId= - already initialized
2018-06-06 12:10:34,420 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - C
e=JobTracker, sessionId= - already initialized
2018-06-06 12:10:34,457 [main] INFO  org.apache.pig.backend.hadoop.executionengin
!
2018-06-06 12:10:34,480 [main] INFO  org.apache.hadoop.conf.Configuration.depreca
Instead, use dfs.bytes-per-checksum
2018-06-06 12:10:34,483 [main] INFO  org.apache.hadoop.conf.Configuration.depreca
d, use fs.defaultFS
2018-06-06 12:10:34,484 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sch
2018-06-06 12:10:35,298 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileIn
2018-06-06 12:10:35,299 [main] INFO  org.apache.pig.backend.hadoop.executionengin
cess : 1
(101,Amitabh,20000,1,101,100)
(101,Amitabh,20000,1,101,200)
(102,Shahrukh,10000,2,102,400)
(102,Shahrukh,10000,2,102,100)
(103,Akshay,11000,3,,)
(104,Anubhav,5000,4,104,300)
(105,Pawan,2500,5,105,100)
(106,Aamir,25000,1,,)
(107,Salman,17500,2,,)
(108,Ranbir,14000,3,,)
(109,Katrina,1000,4,,)
(110,Priyanka,2000,5,110,400)
(111,Tushar,500,1,,)
(112,Ajay,5000,2,,)
(113,Jubeen,1000,1,,)
(114,Madhuri,2000,2,114,200)
grunt>
```



```
grunt> t5_out = FOREACH(FILTER t5 BY $4 IS NULL) GENERATE $0,$1;
grunt> dump t5_out
```



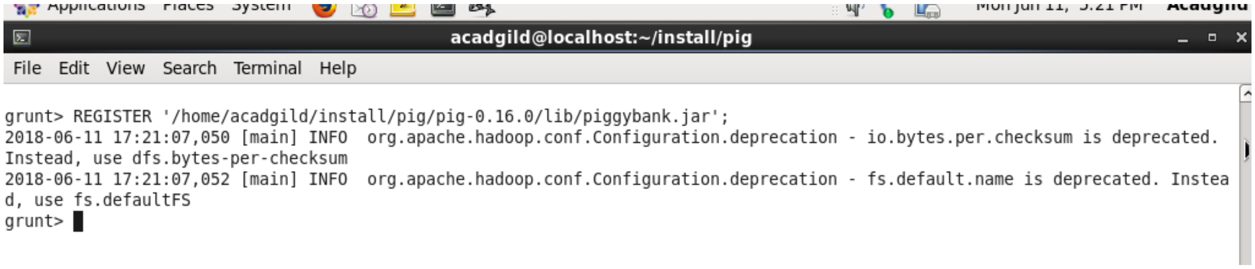
```
Output(s):
Successfully stored 8 records in: "file:/tmp/temp-760701415/tmp-935

Counters:
Total records written : 8
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local225261029_0051


2018-06-06 12:12:11,739 [main] INFO  org.apache.hadoop.metrics.jvm.
e=JobTracker, sessionId= - already initialized
2018-06-06 12:12:11,756 [main] INFO  org.apache.hadoop.metrics.jvm.
e=JobTracker, sessionId= - already initialized
2018-06-06 12:12:11,776 [main] INFO  org.apache.hadoop.metrics.jvm.
e=JobTracker, sessionId= - already initialized
2018-06-06 12:12:11,862 [main] INFO  org.apache.pig.backend.hadoop.
!
2018-06-06 12:12:11,867 [main] INFO  org.apache.hadoop.conf.Configu
Instead, use dfs.bytes-per-checksum
2018-06-06 12:12:11,868 [main] INFO  org.apache.hadoop.conf.Configu
d, use fs.defaultFS
2018-06-06 12:12:11,869 [main] WARN  org.apache.pig.data.SchemaTupl
2018-06-06 12:12:12,134 [main] INFO  org.apache.hadoop.mapreduce.li
2018-06-06 12:12:12,135 [main] INFO  org.apache.pig.backend.hadoop.
cess : 1
(103,Akshay)
(106,Aamir)
(107,Salman)
(108,Ranbir)
(109,Katrina)
(111,Tushar)
(112,Ajay)
(113,Jubeen)
grunt>
```

Code:

T5 = JOIN details BY eidLEFT,expense BY eid;
DUMP T5;
T5_out = FOREACH(FILTER t5 BY $4 IS NULL) GENERATE $0,$1;
DUMP T5_out;

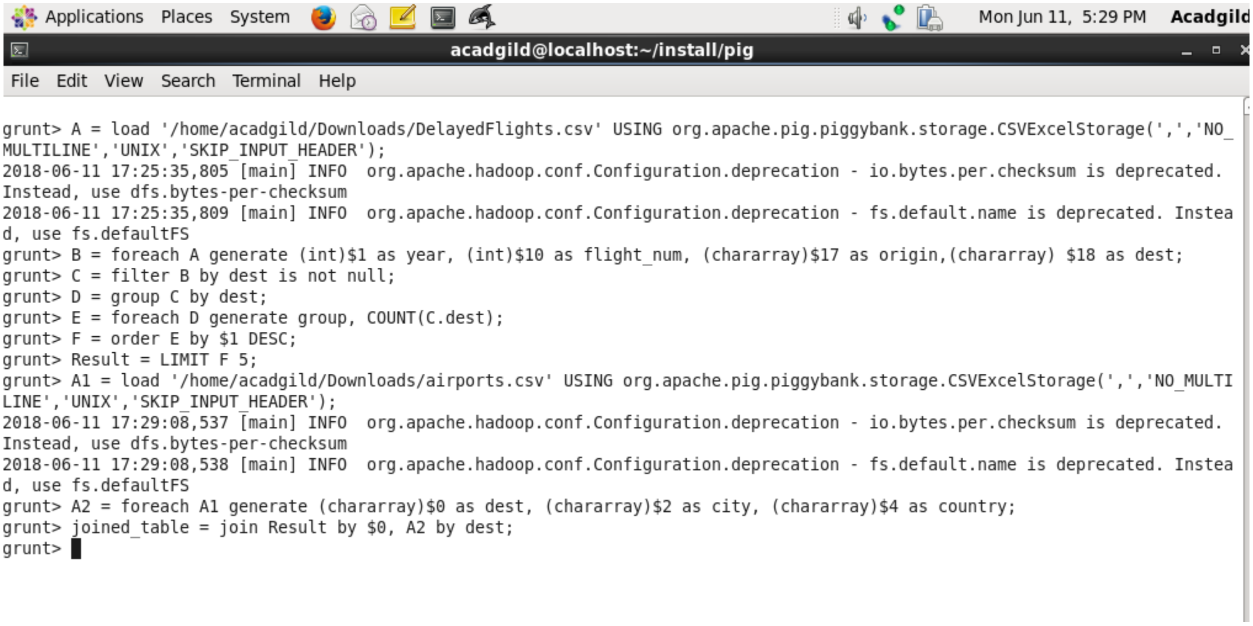
Task 3:
Implement the use case present in blog.

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

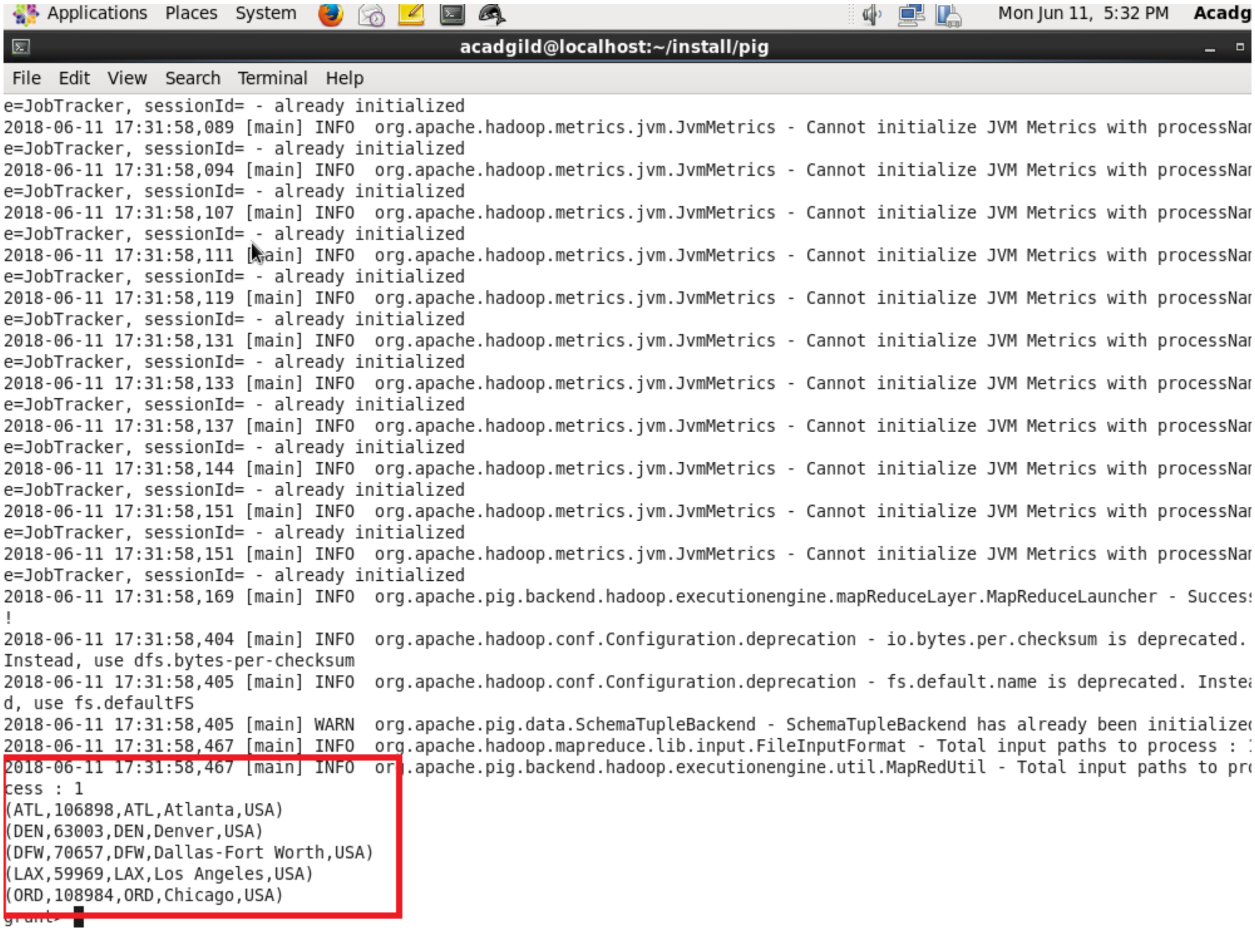

acadgild@localhost:~/install/pig
File Edit View Search Terminal Help

```
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
2018-06-11 17:21:07,050 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-06-11 17:21:07,052 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> 
```

Step 1:
Top 5 most visited destinations:


Applications Places System
Mon Jun 11, 5:29 PM
Acadgild
acadgild@localhost:~/install/pig
File Edit View Search Terminal Help

```
grunt> A = load '/home/acadgild/Downloads/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-06-11 17:25:35,805 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-06-11 17:25:35,809 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18 as dest;
grunt> C = filter B by dest is not null;
grunt> D = group C by dest;
grunt> E = foreach D generate group, COUNT(C.dest);
grunt> F = order E by $1 DESC;
grunt> Result = LIMIT F 5;
grunt> A1 = load '/home/acadgild/Downloads/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-06-11 17:29:08,537 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-06-11 17:29:08,538 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
grunt> joined_table = join Result by $0, A2 by dest;
grunt> 
```


Applications Places System
Mon Jun 11, 5:32 PM
Acadg
acadgild@localhost:~/install/pig
File Edit View Search Terminal Help

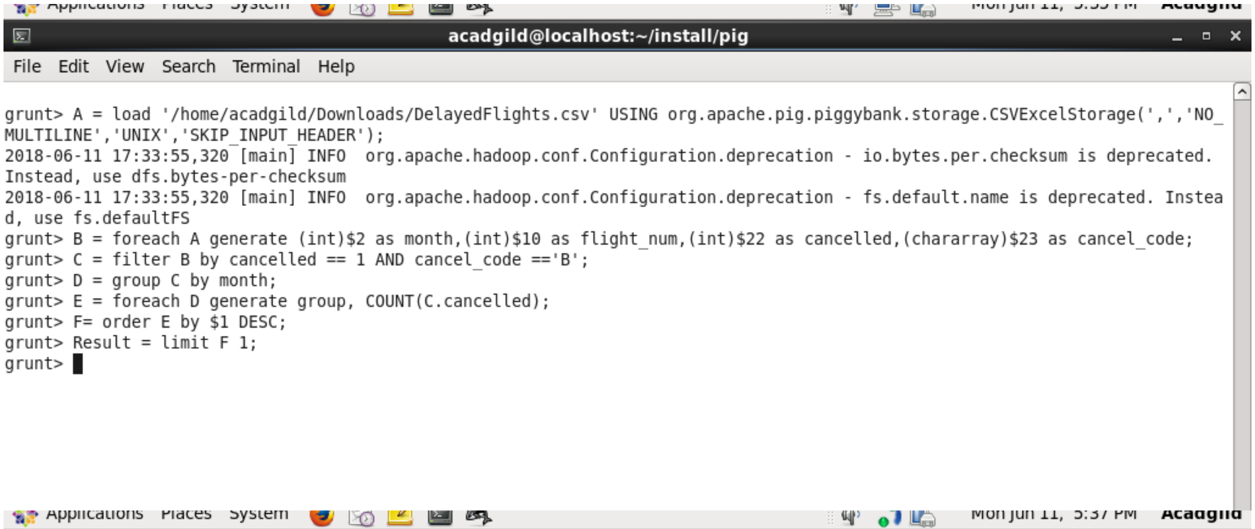
```
e=JobTracker, sessionId= - already initialized
2018-06-11 17:31:58,089 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
e=JobTracker, sessionId= - already initialized
2018-06-11 17:31:58,094 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
e=JobTracker, sessionId= - already initialized
2018-06-11 17:31:58,107 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
e=JobTracker, sessionId= - already initialized
2018-06-11 17:31:58,111 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
e=JobTracker, sessionId= - already initialized
2018-06-11 17:31:58,119 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
e=JobTracker, sessionId= - already initialized
2018-06-11 17:31:58,131 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
e=JobTracker, sessionId= - already initialized
2018-06-11 17:31:58,133 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
e=JobTracker, sessionId= - already initialized
2018-06-11 17:31:58,137 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
e=JobTracker, sessionId= - already initialized
2018-06-11 17:31:58,144 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
e=JobTracker, sessionId= - already initialized
2018-06-11 17:31:58,151 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
e=JobTracker, sessionId= - already initialized
2018-06-11 17:31:58,151 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
e=JobTracker, sessionId= - already initialized
2018-06-11 17:31:58,169 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succes
!
2018-06-11 17:31:58,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-11 17:31:58,405 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-11 17:31:58,405 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialize
2018-06-11 17:31:58,467 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process :
2018-06-11 17:31:58,467 [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)
grunt>
```

Code:

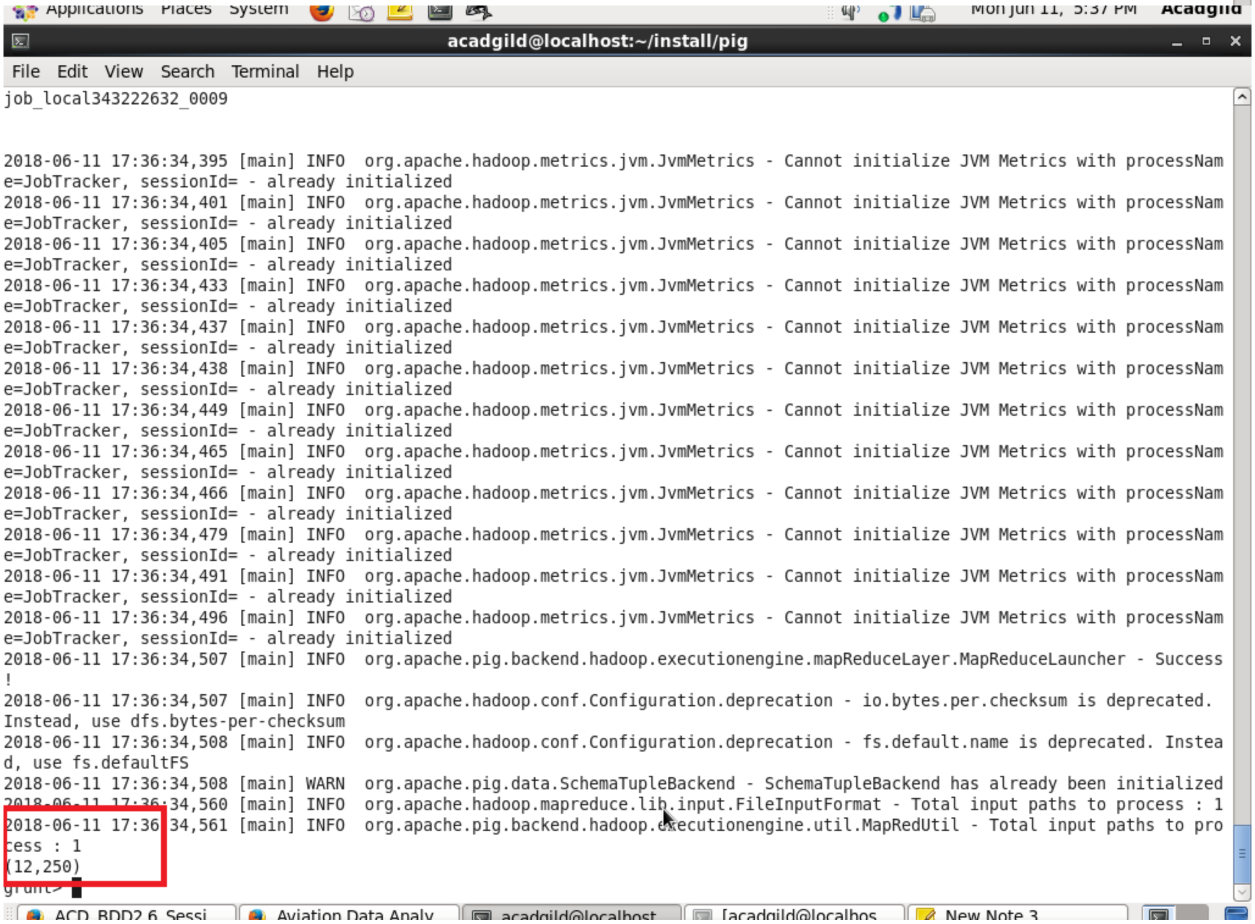
```
A = load '/home/acadgild/Downloads/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18 as dest;
C = filter B by dest is not 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/Downloads/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;
```

Step 2:

Month that has seen most number of cancellations due to bad weather:


Applications Places System
Mon Jun 11, 5:35 PM
Acadgild
acadgild@localhost:~/install/pig
File Edit View Search Terminal Help

```
grunt> A = load '/home/acadgild/Downloads/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_
MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-06-11 17:33:55,320 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-11 17:33:55,320 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
grunt> C = filter B by cancelled == 1 AND cancel_code =='B';
grunt> D = group C by month;
grunt> E = foreach D generate group, COUNT(C.cancelled);
grunt> F= order E by $1 DESC;
grunt> Result = limit F 1;
grunt> 
```



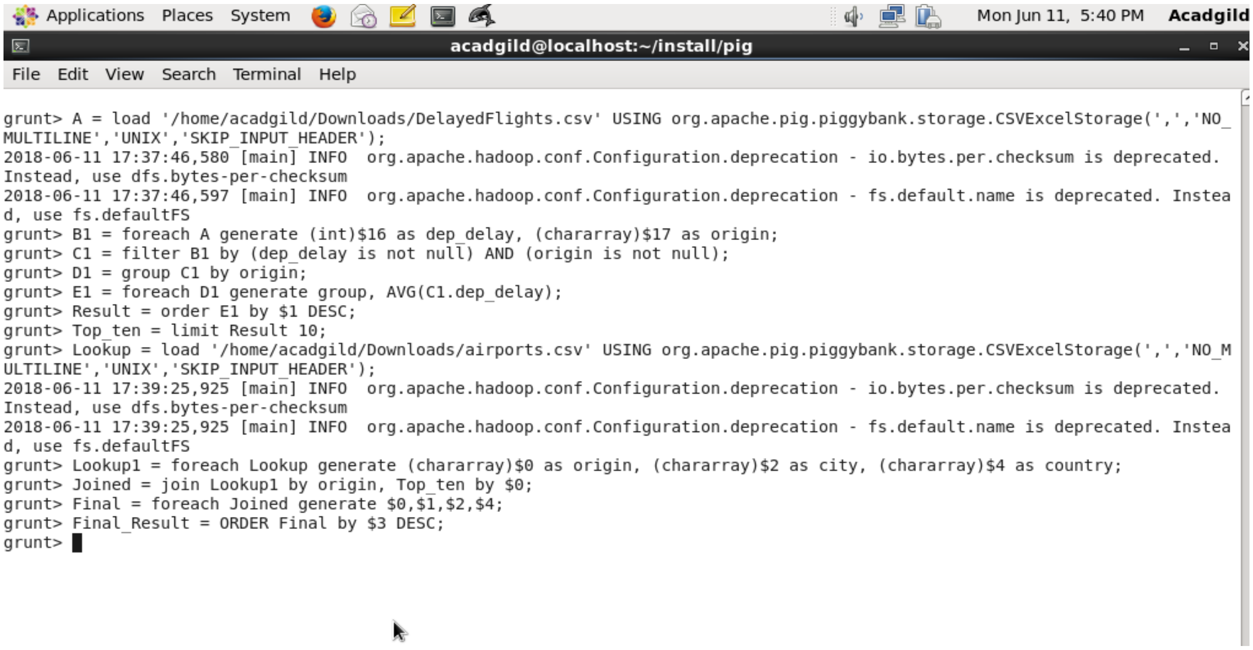
```
job_local343222632_0009


2018-06-11 17:36:34,395 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-11 17:36:34,401 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-11 17:36:34,405 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-11 17:36:34,433 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-11 17:36:34,437 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-11 17:36:34,438 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-11 17:36:34,449 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-11 17:36:34,465 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-11 17:36:34,466 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-11 17:36:34,479 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-11 17:36:34,491 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-11 17:36:34,496 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-11 17:36:34,507 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-06-11 17:36:34,507 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-11 17:36:34,508 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-11 17:36:34,508 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-06-11 17:36:34,560 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-06-11 17:36:34,561 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(12,250)
grunt> 
```

Code:

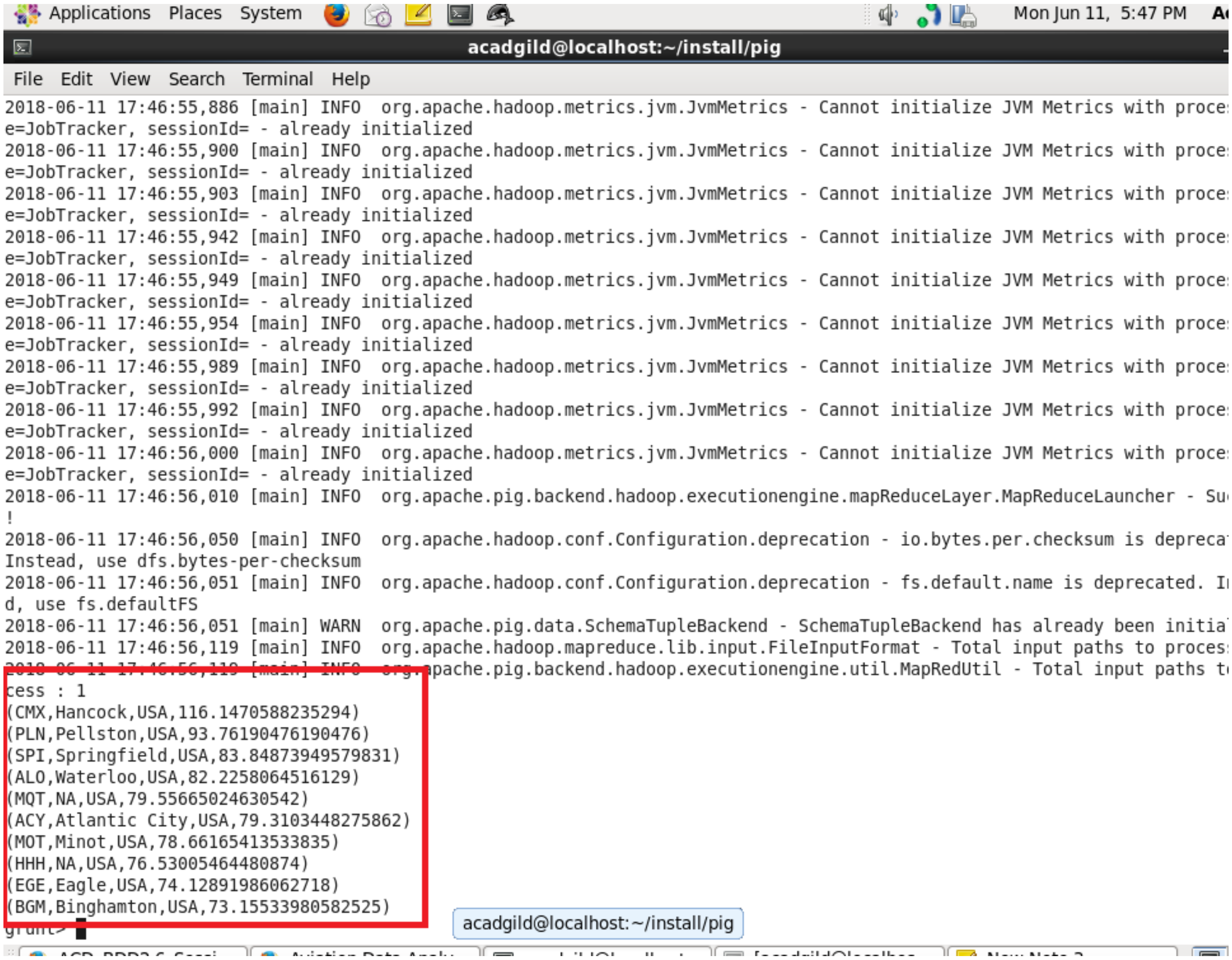
```
A = load '/home/acadgild/Downloads/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_H
EADER');
B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as
cancelled,(chararray)$23 as cancel_code;
C = filter B by cancelled == 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;
```

Step 3:

Top 10 origins with highest avg departure delay



```
grunt> A = load '/home/acadgild/Downloads/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_
MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-06-11 17:37:46,580 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-11 17:37:46,597 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, 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 '/home/acadgild/Downloads/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_M
ULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-06-11 17:39:25,925 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-11 17:39:25,925 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, 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> 
```


Applications Places System
Mon Jun 11, 5:47 PM
acadgild@localhost:~/install/pig
File Edit View Search Terminal Help
acadgild@localhost:~/install/pig

```
2018-06-11 17:46:55,886 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:46:55,900 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:46:55,903 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:46:55,942 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:46:55,949 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:46:55,954 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:46:55,989 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:46:55,992 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:46:56,000 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:46:56,010 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Su
!
2018-06-11 17:46:56,050 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depreca
Instead, use dfs.bytes-per-checksum
2018-06-11 17:46:56,051 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. I
d, use fs.defaultFS
2018-06-11 17:46:56,051 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initia
2018-06-11 17:46:56,119 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to proces
2018-06-11 17:46:56,119 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths t
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)
grunt> 
```

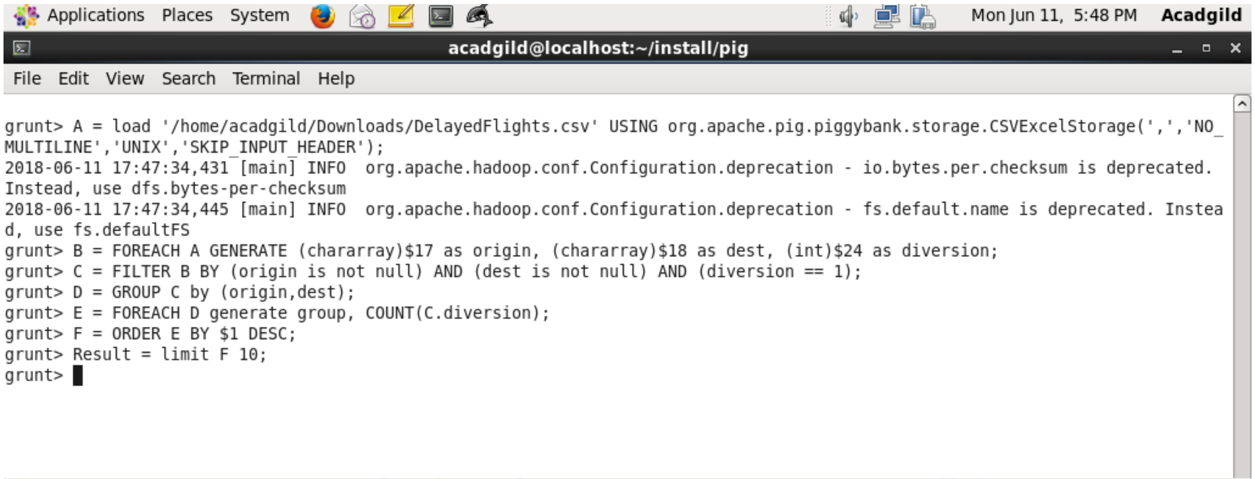
Code:

```
A = load '/home/acadgild/Downloads/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/Downloads/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;
```

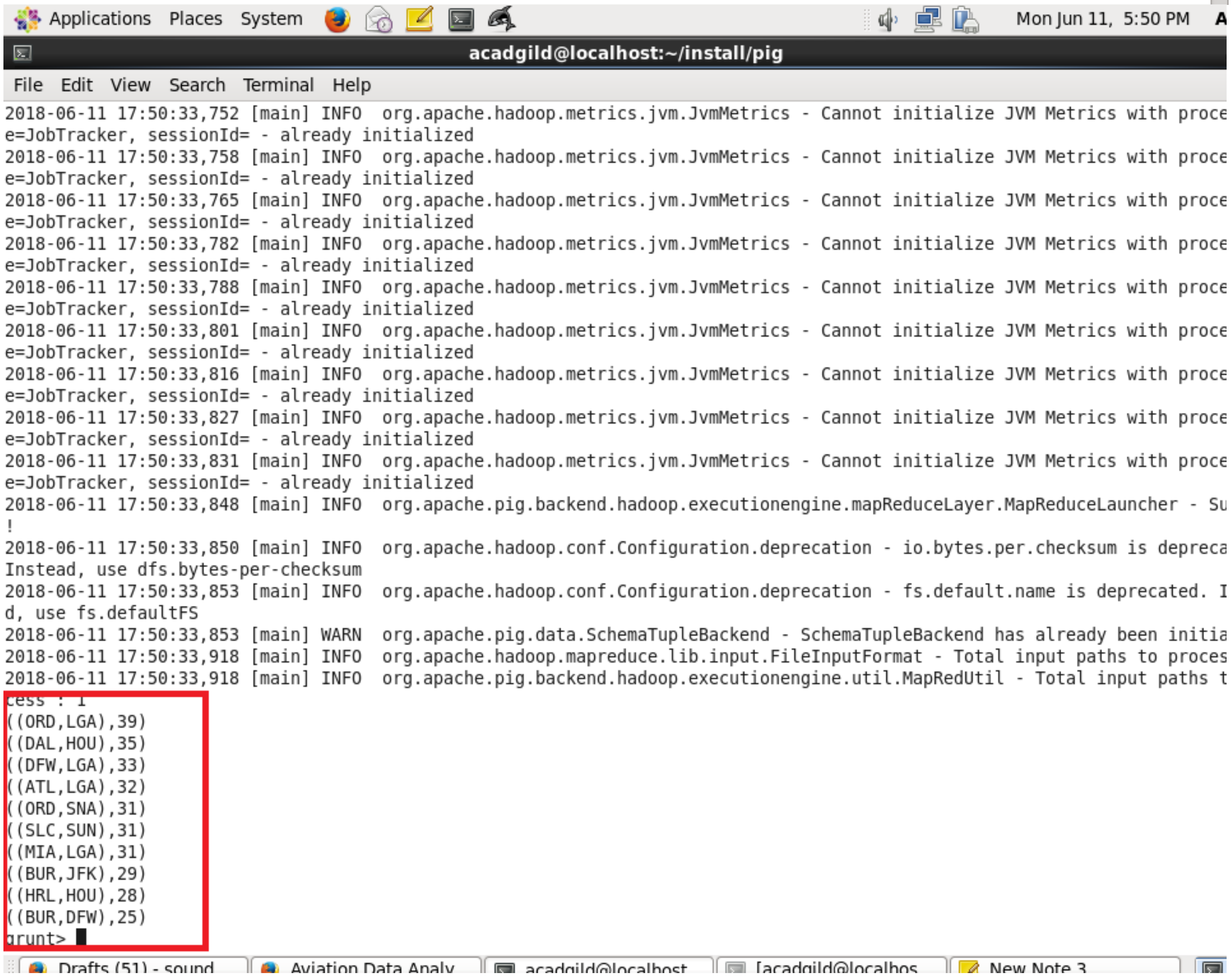
Final_Result = ORDER Final by $3 DESC;
dump Final_Result;

Step 4:

Which route has seen max diversions


Applications Places System
Mon Jun 11, 5:48 PM Acadgild
acadgild@localhost:~/install/pig
File Edit View Search Terminal Help

```
grunt> A = load '/home/acadgild/Downloads/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_
MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-06-11 17:47:34,431 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-11 17:47:34,445 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;
grunt> C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);
grunt> D = GROUP C by (origin,dest);
grunt> E = FOREACH D generate group, COUNT(C.diversion);
grunt> F = ORDER E BY $1 DESC;
grunt> Result = limit F 10;
grunt> 
```



```
2018-06-11 17:50:33,752 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:50:33,758 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:50:33,765 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:50:33,782 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:50:33,788 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:50:33,801 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:50:33,816 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:50:33,827 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:50:33,831 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proce
e=JobTracker, sessionId= - already initialized
2018-06-11 17:50:33,848 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Su
!
2018-06-11 17:50:33,850 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depreca
Instead, use dfs.bytes-per-checksum
2018-06-11 17:50:33,853 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. I
d, use fs.defaultFS
2018-06-11 17:50:33,853 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initia
2018-06-11 17:50:33,918 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to proces
2018-06-11 17:50:33,918 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths t
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)
grunt> 
```

Code:

```
A = load '/home/acadgild/Downloads/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;
```