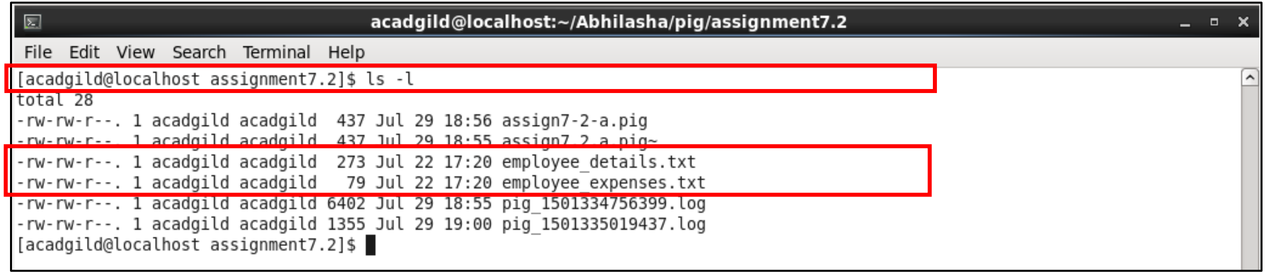
ASSIGNMENT 7

We have employee_details and employee_expenses files. Use local mode while running Pig and write Pig Latin script to get below results:

Input Files:

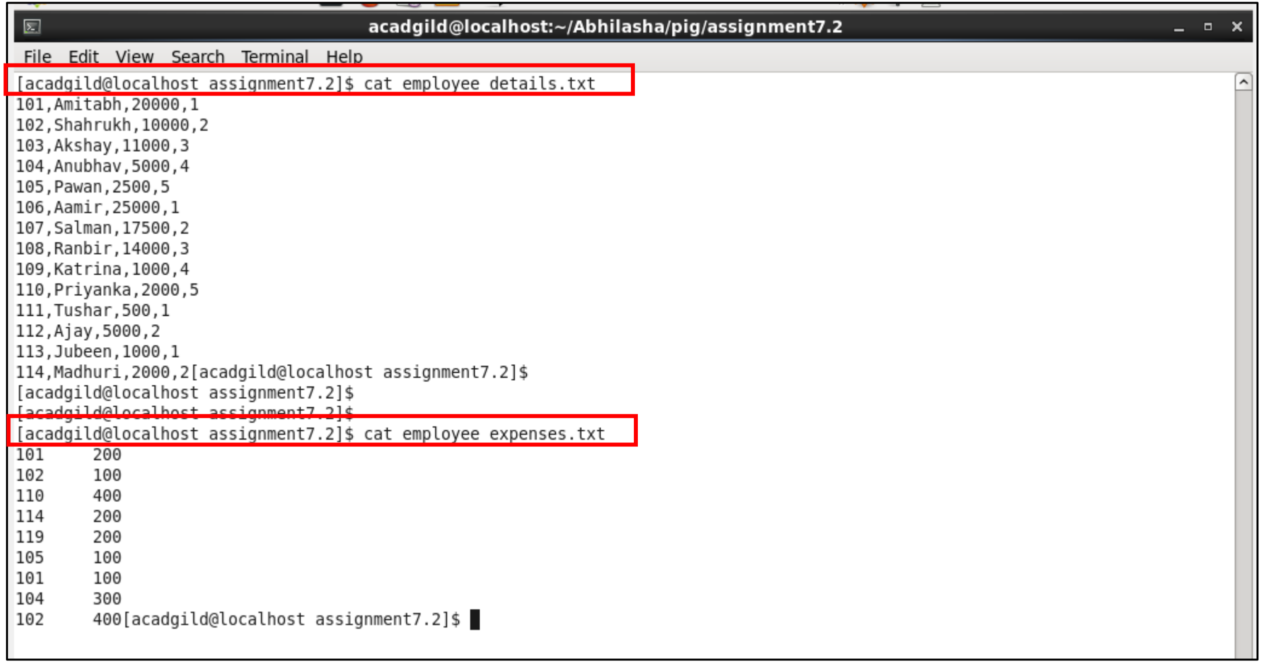
a. Screenshot mentioning the presence of the input files in local directory



acadgild@localhost:~/Abhilasha/pig/assignment7.2
File Edit View Search Terminal Help

```
[acadgild@localhost assignment7.2]$ ls -l
total 28
-rw-rw-r--. 1 acadgild acadgild  437 Jul 29 18:56 assign7-2-a.pig
-rw-rw-r--. 1 acadgild acadgild  437 Jul 29 18:55 assign7_2_a.pig~
-rw-rw-r--. 1 acadgild acadgild  273 Jul 22 17:20 employee_details.txt
-rw-rw-r--. 1 acadgild acadgild   79 Jul 22 17:20 employee_expenses.txt
-rw-rw-r--. 1 acadgild acadgild 6402 Jul 29 18:55 pig_1501334756399.log
-rw-rw-r--. 1 acadgild acadgild 1355 Jul 29 19:00 pig_1501335019437.log
[acadgild@localhost assignment7.2]$
```

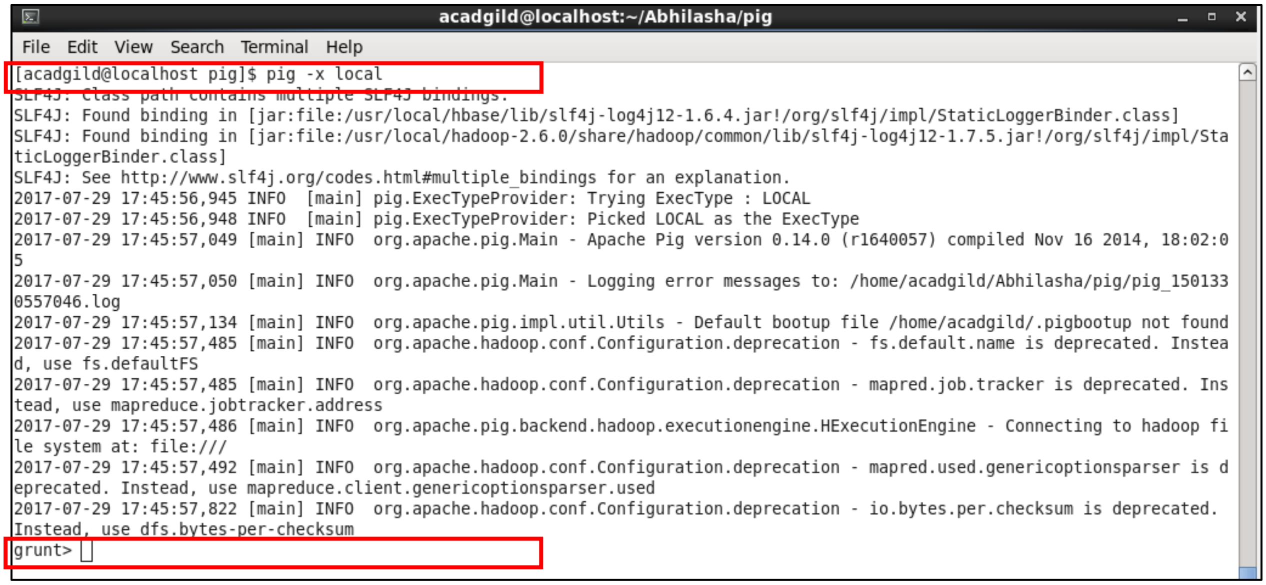
b. Content of input files are as follows :



acadgild@localhost:~/Abhilasha/pig/assignment7.2
File Edit View Search Terminal Help

```
[acadgild@localhost assignment7.2]$ cat employee_details.txt
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[acadgild@localhost assignment7.2]$
[acadgild@localhost assignment7.2]$
[acadgild@localhost assignment7.2]$
[acadgild@localhost assignment7.2]$ cat employee_expenses.txt
101     200
102     100
110     400
114     200
119     200
105     100
101     100
104     300
102     400[acadgild@localhost assignment7.2]$
```

c. Starting pig in local mode. This was needed to test the script, in step by step execution.


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

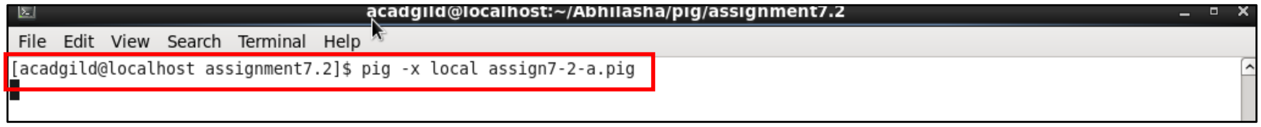
```
[acadgild@localhost pig]$ pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Sta
ticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2017-07-29 17:45:56,945 INFO  [main] pig.ExecTypeProvider: Trying ExecType : LOCAL
2017-07-29 17:45:56,948 INFO  [main] pig.ExecTypeProvider: Picked LOCAL as the ExecType
2017-07-29 17:45:57,049 [main] INFO  org.apache.pig.Main - Apache Pig version 0.14.0 (r1640057) compiled Nov 16 2014, 18:02:0
5
2017-07-29 17:45:57,050 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/Abhilasha/pig/pig_150133
0557046.log
2017-07-29 17:45:57,134 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2017-07-29 17:45:57,485 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-07-29 17:45:57,485 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2017-07-29 17:45:57,486 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
le system at: file:///
2017-07-29 17:45:57,492 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.used.genericoptionsparser is d
eprecated. Instead, use mapreduce.client.genericoptionsparser.used
2017-07-29 17:45:57,822 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
grunt>
```

This shows that the grunt shell is launched.

--Problem Statement 1--

(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)

Solution: The script execution was as follows:


acadgild@localhost:~/Abhilasha/pig/assignment7.2
File Edit View Search Terminal Help

```
[acadgild@localhost assignment7.2]$ pig -x local assign7-2-a.pig
```

Name of the script executed was assign7-2-a.pig.
The details of script are as follows :
The script was executed in local mode. Hence, **-x local** was used while executing the script.

Step 1: *`empDetails = LOAD 'employee_details.txt' USING PigStorage(',') AS (empId:int, empName:chararray,empSalary:int,empRating:int);`*

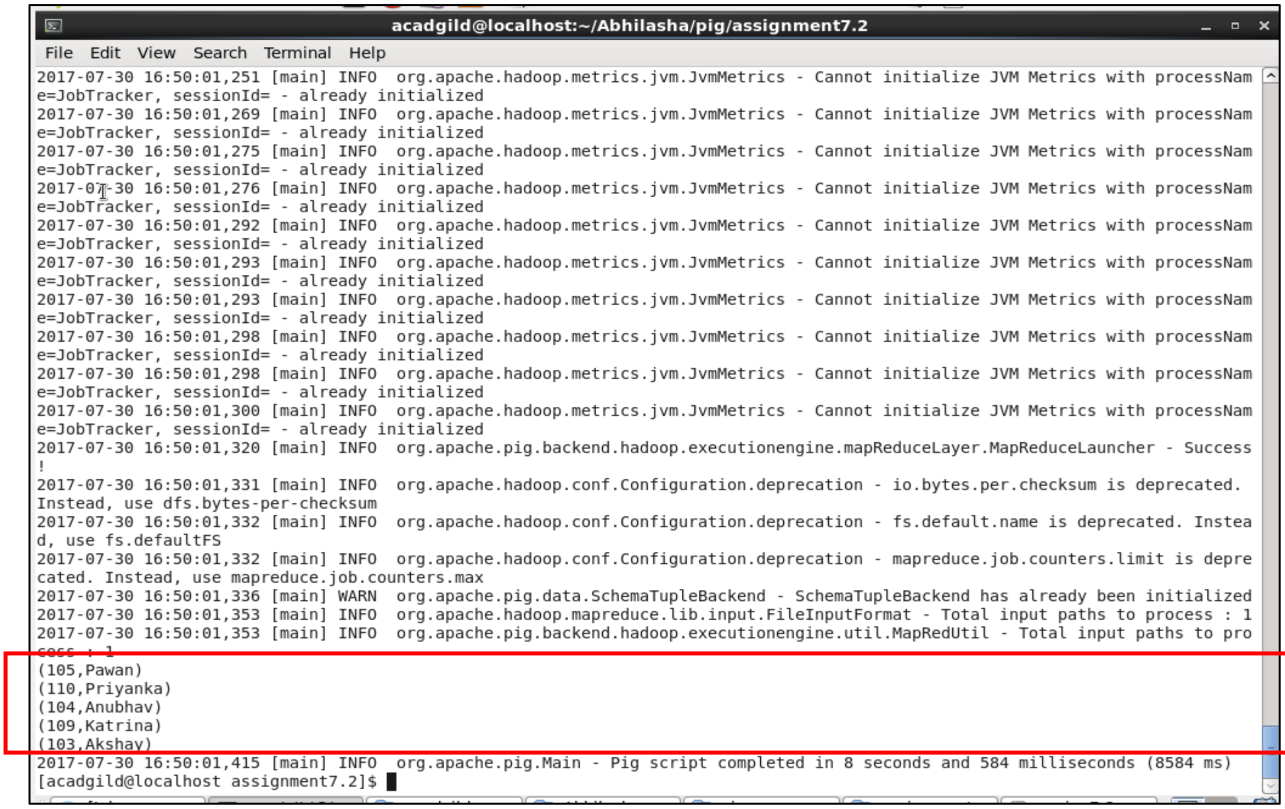
Load **employee_details.tx**t in a variable **empDetails**. Using PigStorage operator, we have specified the delimiter for records, i.e., ‘**,**’. Also, we have specified the schema of the data and named the columns as empId, empName, empSalary, empRating that have data-types integer, chararray, integer and integer respectively.

Step 2: `sortByRating = ORDER empDetails by empRating DESC, empName;`
This is to sort the records based on **empRating** in descending order. Also the records are arranged in dictionary order of names.

Step 3: `limitedRecords = LIMIT sortByRating 5;`
This is to limit the number of records in output to 5. Hence, used **LIMIT** command.

Step 4: `requiredEmps = FOREACH limitedRecords generate empId,empName;`
This is to fetch only **empId** and **empName** from the resultset.

Step 5: `dump requiredEmps;`
This is to dump the result on console. The Output is as follows
The output is as follows:

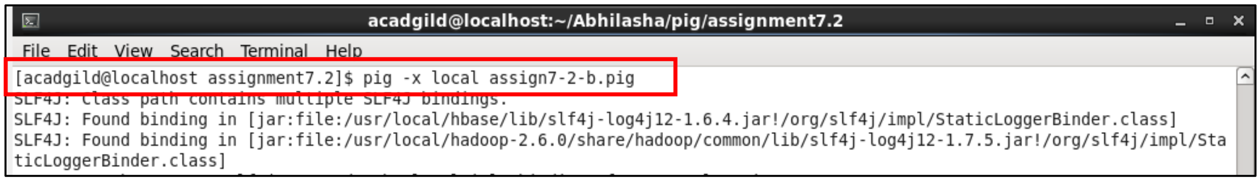

acadgild@localhost:~/Abhilasha/pig/assignment7.2
File Edit View Search Terminal Help

```
2017-07-30 16:50:01,251 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 16:50:01,269 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 16:50:01,275 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 16:50:01,276 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 16:50:01,292 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 16:50:01,293 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 16:50:01,293 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 16:50:01,298 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 16:50:01,298 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 16:50:01,300 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 16:50:01,320 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-07-30 16:50:01,331 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-07-30 16:50:01,332 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-07-30 16:50:01,332 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-07-30 16:50:01,336 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-07-30 16:50:01,353 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-07-30 16:50:01,353 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(105,Pawan)
(110,Priyanka)
(104,Anubhav)
(109,Katrina)
(103,Akshay)
2017-07-30 16:50:01,415 [main] INFO  org.apache.pig.Main - Pig script completed in 8 seconds and 584 milliseconds (8584 ms)
[acadgild@localhost assignment7.2]$ 
```

--Problem Statement 2---

(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)

Solution: The script execution was as follows:


acadgild@localhost:~/Abhilasha/pig/assignment7.2
File Edit View Search Terminal Help

```
[acadgild@localhost assignment7.2]$ pig -x local assign7-2-b.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Sta
ticLoggerBinder.class]
```

Name of the script executed was assign7-2-b.pig.

The details of script are as follows:
The script was executed in local mode. Hence, **-x local** was used while executing the script.

Step 1:
```
empDetails = LOAD 'employee_details.txt' USING PigStorage(',') AS
(empId:int, empName:chararray,empSalary:int);
```

Load **employee_details**.txt in a variable **empDetails**. Using PigStorage operator, we have specified the delimiter for records, i.e., ‘**,**’. Also, we have specified the schema of the data and named the columns as empId, empName, empSalary that have data-types integer, chararray and integer respectively.

Step 2:
```
filteredEmps = FILTER empDetails BY empId % 2 != 0;
```
Apply filter to get only those records that have odd **empId**. This is checked by performing a modulo operation, where, if output of modulo is 0 -> even else **empId** is odd.

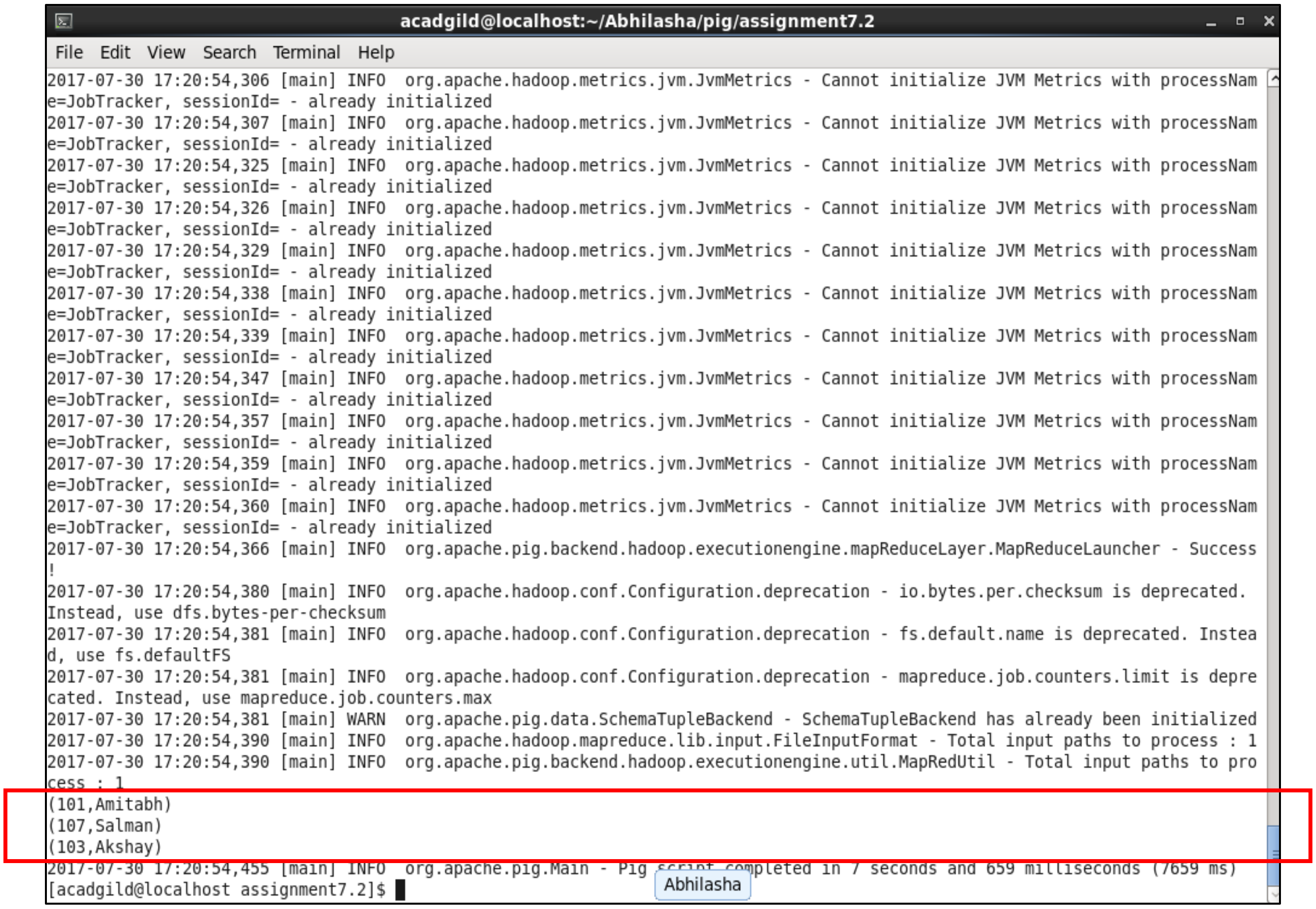
Step 3:
```
sortBySalary = ORDER filteredEmps by empSalary DESC,empName ;
```
Sort the records on the basis of **empSalary** in descending order and **empName**.

Step 4:
```
limitedRecords = LIMIT sortBySalary 3;
```
To fetch only 3 records, we are using **LIMIT** operation.

Step 5:
```
requiredEmps = FOREACH limitedRecords generate empId,empName;
```
To get only empId and empName columns from resultset.

Step 6:
```
dump requiredEmps;
```
Dump the output on the console.

The output is as follows:

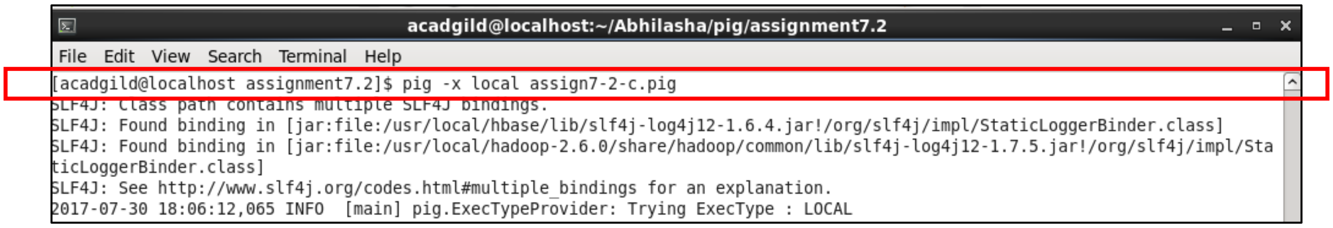

acadgild@localhost:~/Abhilasha/pig/assignment7.2
File Edit View Search Terminal Help
Abhilasha

```
2017-07-30 17:20:54,306 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 17:20:54,307 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 17:20:54,325 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 17:20:54,326 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 17:20:54,329 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 17:20:54,338 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 17:20:54,339 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 17:20:54,347 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 17:20:54,357 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 17:20:54,359 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 17:20:54,360 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 17:20:54,366 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-07-30 17:20:54,380 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-07-30 17:20:54,381 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-07-30 17:20:54,381 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-07-30 17:20:54,381 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-07-30 17:20:54,390 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-07-30 17:20:54,390 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(101,Amitabh)
(107,Salman)
(103,Akshay)
2017-07-30 17:20:54,455 [main] INFO  org.apache.pig.Main - Pig script completed in 7 seconds and 659 milliseconds (7659 ms)
[acadgild@localhost assignment7.2]$
```

--Problem Statement 3---

(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)

Solution: The script execution was as follows:


acadgild@localhost:~/Abhilasha/pig/assignment7.2
File Edit View Search Terminal Help

```
[acadgild@localhost assignment7.2]$ pig -x local assign7-2-c.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Sta
ticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2017-07-30 18:06:12,065 INFO  [main] pig.ExecTypeProvider: Trying ExecType : LOCAL
```

Name of the script executed was assign7-2-c.pig.

The details of script are as follows:

The script was executed in local mode. Hence, **-x local** was used while executing the script.

Step 1: `empDetails = LOAD 'employee_details.txt' USING PigStorage(',') AS (empId:int, empName:chararray);`

Load **employee_details**.txt in a variable **empDetails**. Using PigStorage operator, we have specified the delimiter for records, i.e., ‘,’. Also, we have specified the schema of the data and named the columns as empId, empName that have data-types integer and chararray respectively.

Step 2: `empExpenses = LOAD 'employee_expenses.txt' USING PigStorage('	') AS (empId:int, expenses:int);`
Load **employee_expenses**.txt in a variable **empExpenses**. Using PigStorage operator, we have specified the delimiter for records, i.e., ‘	’. Also, we have specified the schema of the data and named the columns as empId, expenses that have data-types integer and integer respectively.

Step 3: `joinData = JOIN empDetails BY empId, empExpenses by empId;`
Perform join of empDetails and empExpenses on employee id.

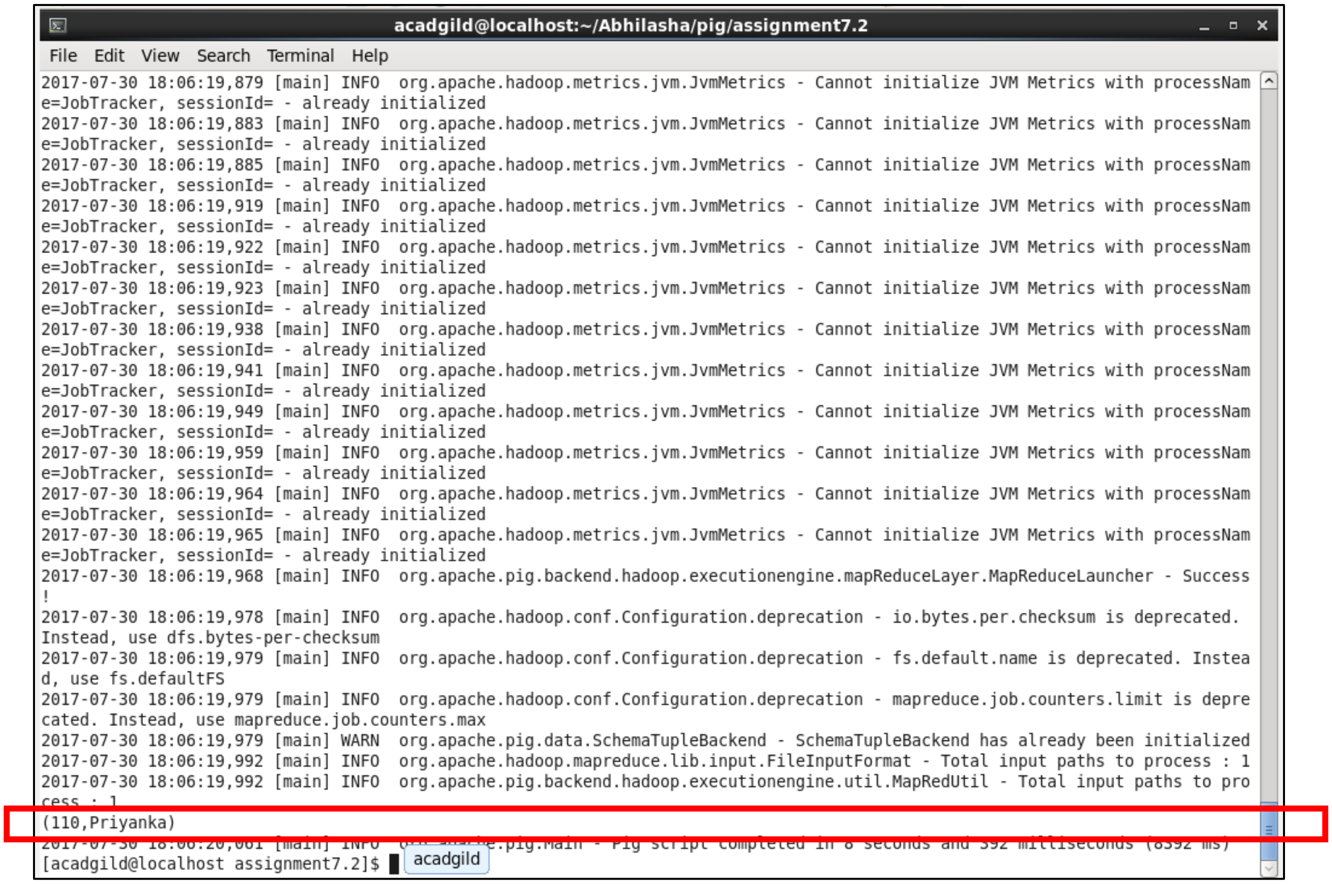
Step 4: `sortedData = ORDER joinData by empExpenses::expenses DESC,empDetails::empName ;`
Sort the resultset by expenses in descending order and by employee name.

Step 5: `firstRecord = LIMIT sortedData 1;`
Get first record only. Hence, use **LIMIT** operator.

Step 6: `requiredEmp = FOREACH firstRecord generate empDetails::empId,empDetails::empName;`
To get only empId and empName columns from resultset.

Step 7: `dump requiredEmp;`
Dump the output on the console.

The output is as follows :

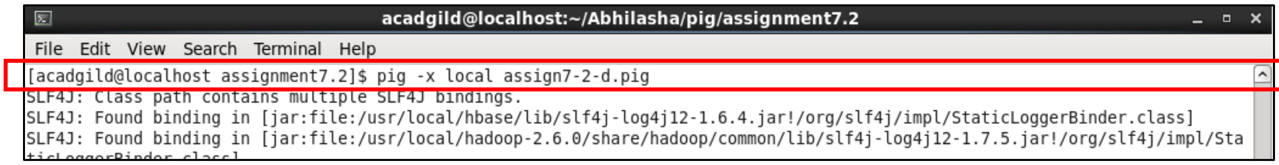

acadgild@localhost:~/Abhilasha/pig/assignment7.2
File Edit View Search Terminal Help

```
2017-07-30 18:06:19,879 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:06:19,883 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:06:19,885 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:06:19,919 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:06:19,922 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:06:19,923 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:06:19,938 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:06:19,941 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:06:19,949 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:06:19,959 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:06:19,964 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:06:19,965 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:06:19,968 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-07-30 18:06:19,978 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-07-30 18:06:19,979 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-07-30 18:06:19,979 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-07-30 18:06:19,979 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-07-30 18:06:19,992 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-07-30 18:06:19,992 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(110,Priyanka)
2017-07-30 18:06:20,061 [main] INFO  org.apache.pig.Main - Pig script completed in 8 seconds and 392 milliseconds (8392 ms)
[acadgild@localhost assignment7.2]$ 
```

---Problem Statement 4---

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

Solution: The script execution was as follows:


acadgild@localhost:~/Abhilasha/pig/assignment7.2
File Edit View Search Terminal Help

```
[acadgild@localhost assignment7.2]$ pig -x local assign7-2-d.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Sta
ticLoggerBinder.class]
```

Name of the script executed was assign7-2-d.pig.

The details of script are as follows:

The script was executed in local mode. Hence, **-x local** was used while executing the script.

`Step 1: empDetails = LOAD 'employee_details.txt' USING PigStorage(',') AS (empId:int, empName:chararray);`

Load **employee_details**.txt in a variable **empDetails**. Using PigStorage operator, we have specified the delimiter for records, i.e., '**,**'. Also, we have specified the schema of the data and named the columns as empId, empName that have data-types integer and chararray respectively.

Step 2:

```
empExpenses = LOAD 'employee_expenses.txt' USING PigStorage('') AS
(empId:int);
```

Load **employee_expenses**.txt in a variable **empExpenses**. Using PigStorage operator, we have specified the delimiter for records, i.e., ‘ ’. Also, we have specified the schema of the data and named the column as empId that has data-type integer.

Step 3:

```
joinData = JOIN empDetails BY empId, empExpenses by empId;
```

Perform join of empDetails and empExpenses on employee id.

Step 4:

```
distinctRecords = DISTINCT joinData;
```

Data in 'employee_expenses.txt' has multiple entries for a few empIds. Hence, to remove duplicates, we have used **DISTINCT**.

Step 5:

```
requiredEmps = FOREACH distinctRecords generate
empDetails::empId,empDetails::empName;
```

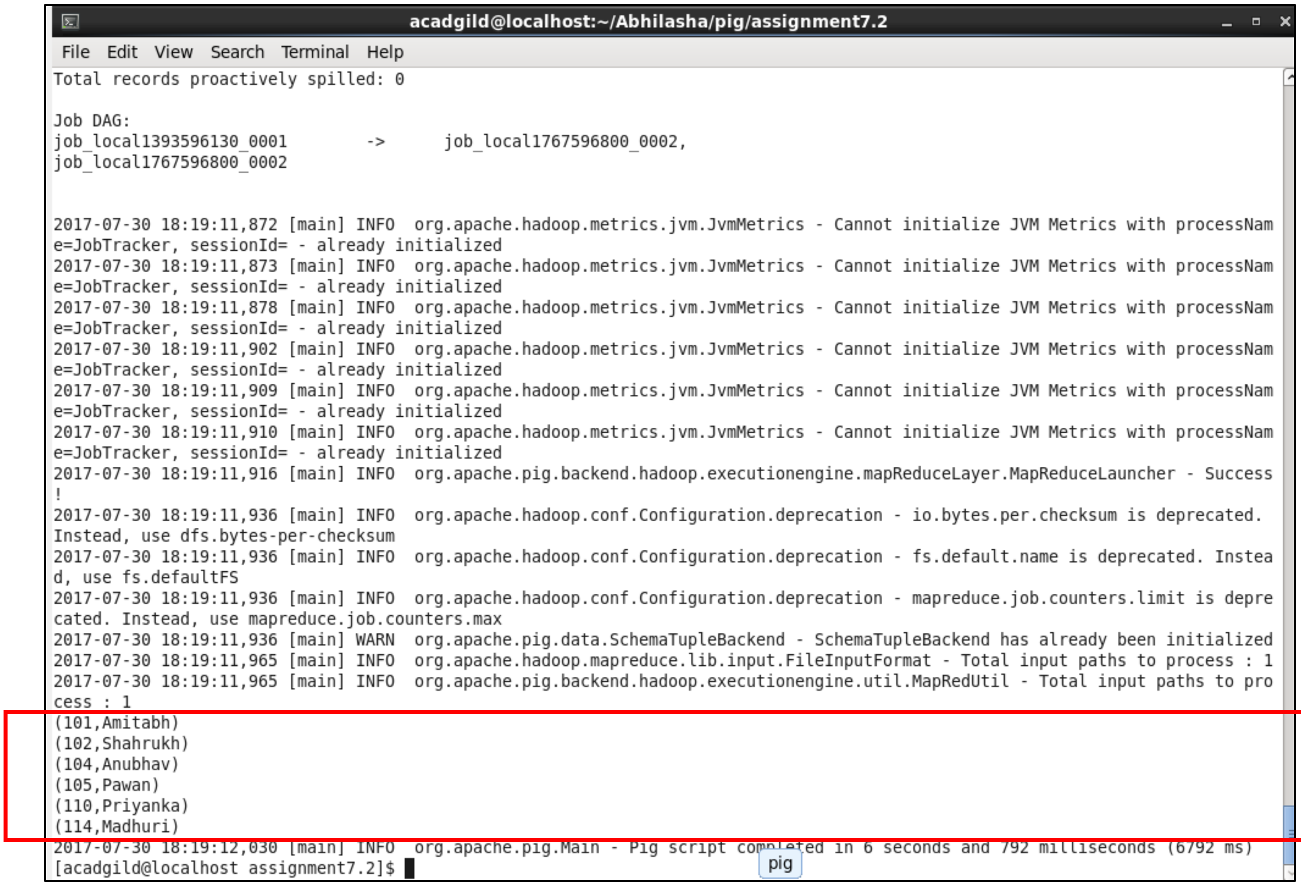
To get only empId and empName columns from resultset.

Step 7:

```
dump requiredEmp;
```

Dump the output on the console.

The output is as follows :

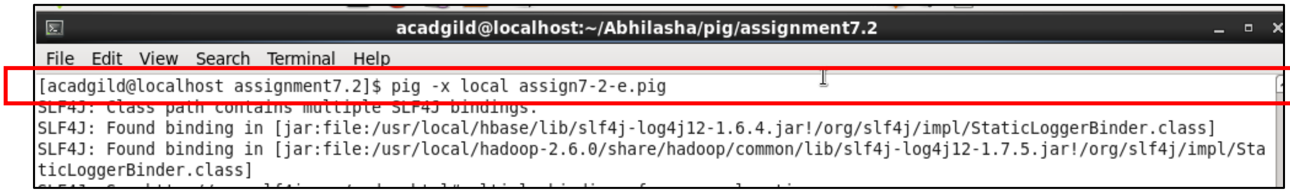

acadgild@localhost:~/Abhilasha/pig/assignment7.2
File Edit View Search Terminal Help
pig

```
Total records proactively spilled: 0

Job DAG:
job_local1393596130_0001        ->      job_local1767596800_0002,
job_local1767596800_0002


2017-07-30 18:19:11,872 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:19:11,873 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:19:11,878 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:19:11,902 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:19:11,909 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:19:11,910 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:19:11,916 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-07-30 18:19:11,936 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-07-30 18:19:11,936 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-07-30 18:19:11,936 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-07-30 18:19:11,936 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-07-30 18:19:11,965 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-07-30 18:19:11,965 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(101,Amitabh)
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
2017-07-30 18:19:12,030 [main] INFO  org.apache.pig.Main - Pig script completed in 6 seconds and 792 milliseconds (6792 ms)
[acadgild@localhost assignment7.2]$ 
```

--Problem Statement 5---

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

Solution: The script execution was as follows:


acadgild@localhost:~/Abhilasha/pig/assignment7.2
File Edit View Search Terminal Help

```
[acadgild@localhost assignment7.2]$ pig -x local assign7-2-e.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Sta
ticLoggerBinder.class]
```

Name of the script executed was assign7-2-e.pig.

The details of script are as follows :

The script was executed in local mode. Hence, **-x local** was used while executing the script.

Step 1: `empDetails = LOAD 'employee_details.txt' USING PigStorage(',') AS (empId:int, empName:chararray);`

Load **employee_details**.txt in a variable **empDetails**. Using PigStorage operator, we have specified the delimiter for records, i.e., **','**. Also, we have specified the schema of the data and named the columns as empId, empName that have data-types integer and chararray respectively.

Step 2: `empExpenses = LOAD 'employee_expenses.txt' USING PigStorage(' ') AS (empId:int);`

Load **employee_expenses**.txt in a variable **empExpenses**. Using PigStorage operator, we have specified the delimiter for records, i.e., ‘ ’. Also, we have specified the schema of the data and named the column as empId that has data-type integer.

Step 3: `coGroupData = COGROUP empDetails BY empId, empExpenses by empId;`

COGROUP is used to achieve cross-product join as well as group by. Here, the purpose is to get columns of both the relations in a record.

Step 4: `filteredData = FILTER coGroupData BY IsEmpty(empExpenses);`

Filter the records to get only those that have no data for columns from employee expenses data. Used IsEmpty()for the same.

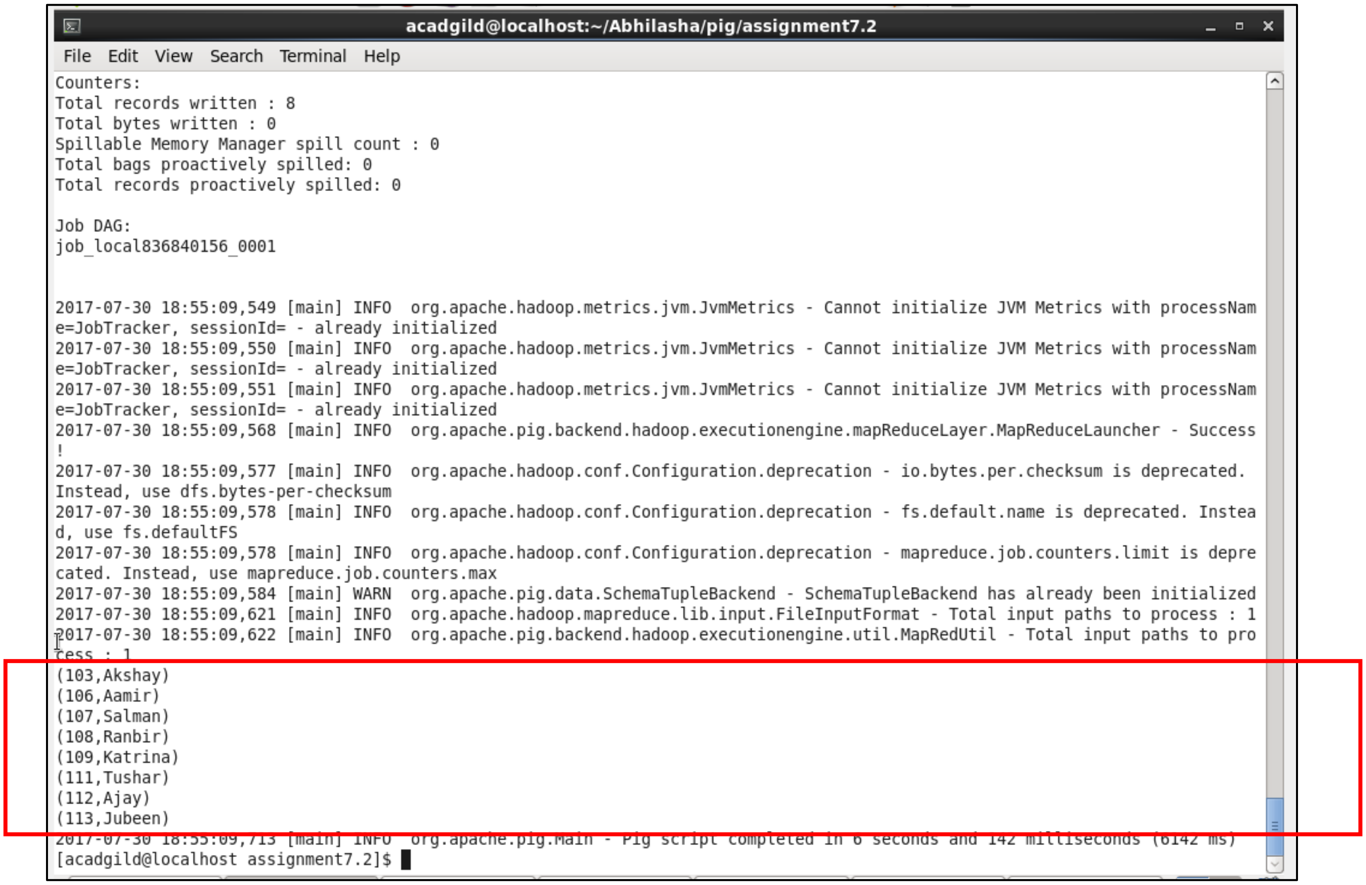
Step 5: `flattenedData = FOREACH filteredData generate FLATTEN(empDetails);`

To flatten the data, i.e., convert bag of tuples into distinct tuples.

Step 7: `dump flattenedData;`

Dump the output on the console.

The output is as follows :


acadgild@localhost:~/Abhilasha/pig/assignment7.2
File Edit View Search Terminal Help

```
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_local836840156_0001


2017-07-30 18:55:09,549 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:55:09,550 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:55:09,551 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-07-30 18:55:09,568 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-07-30 18:55:09,577 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-07-30 18:55:09,578 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-07-30 18:55:09,578 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-07-30 18:55:09,584 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-07-30 18:55:09,621 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-07-30 18:55:09,622 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(103,Akshay)
(106,Aamir)
(107,Salman)
(108,Ranbir)
(109,Katrina)
(111,Tushar)
(112,Ajay)
(113,Jubeen)
2017-07-30 18:55:09,713 [main] INFO  org.apache.pig.Main - Pig script completed in 6 seconds and 142 milliseconds (6142 ms)
[acadgild@localhost assignment7.2]$
```