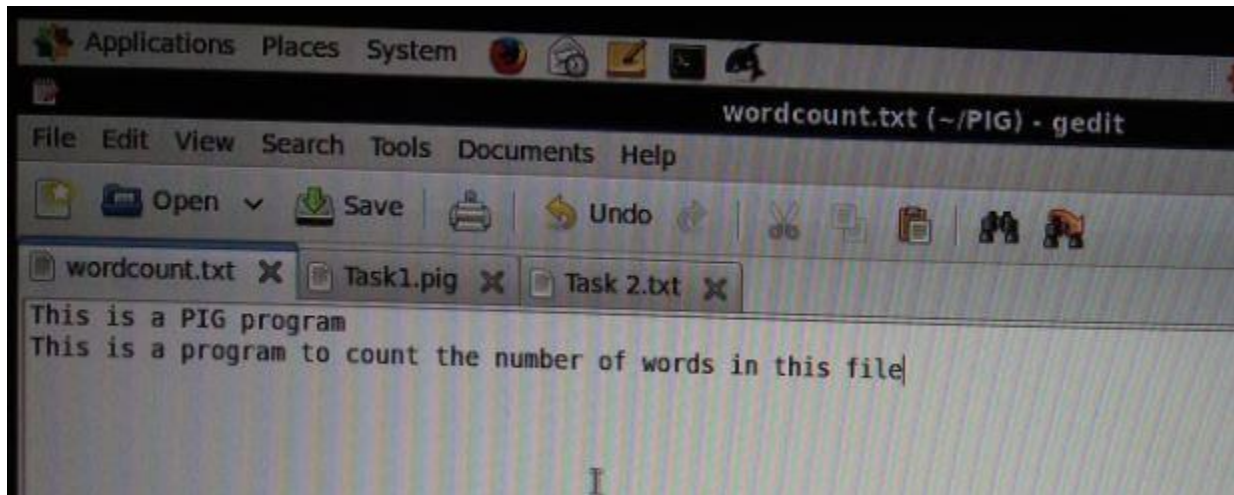


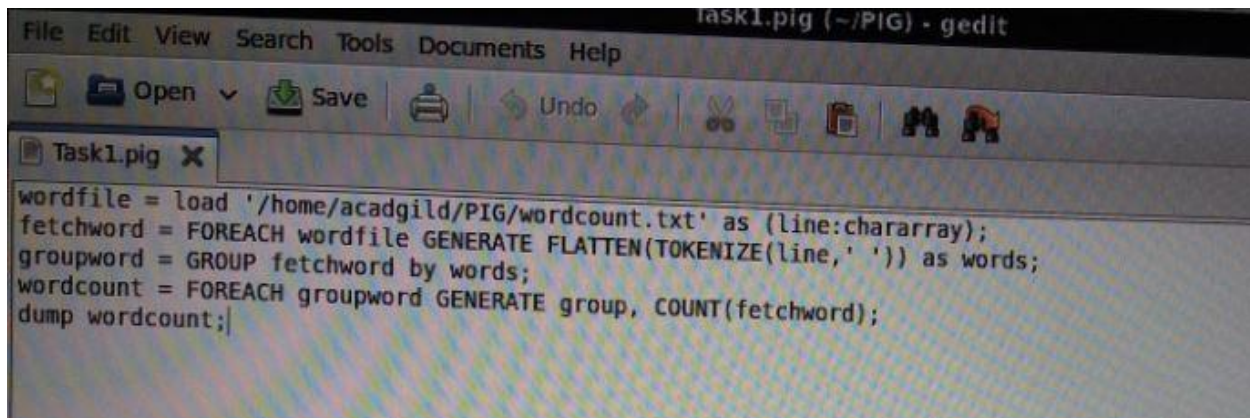
## Assignment 7.1

### Task 1:

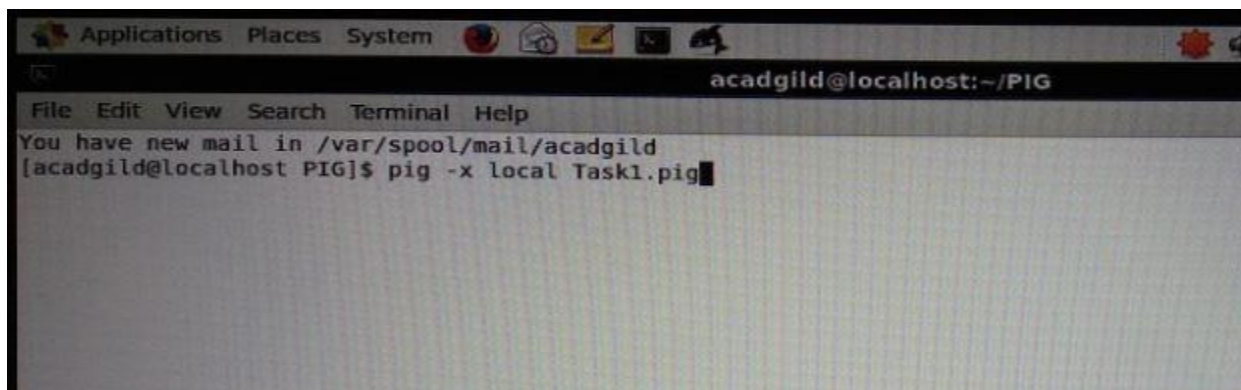
1. File used for word count script.



2. Wrote the following script to count the number of words in the above file.



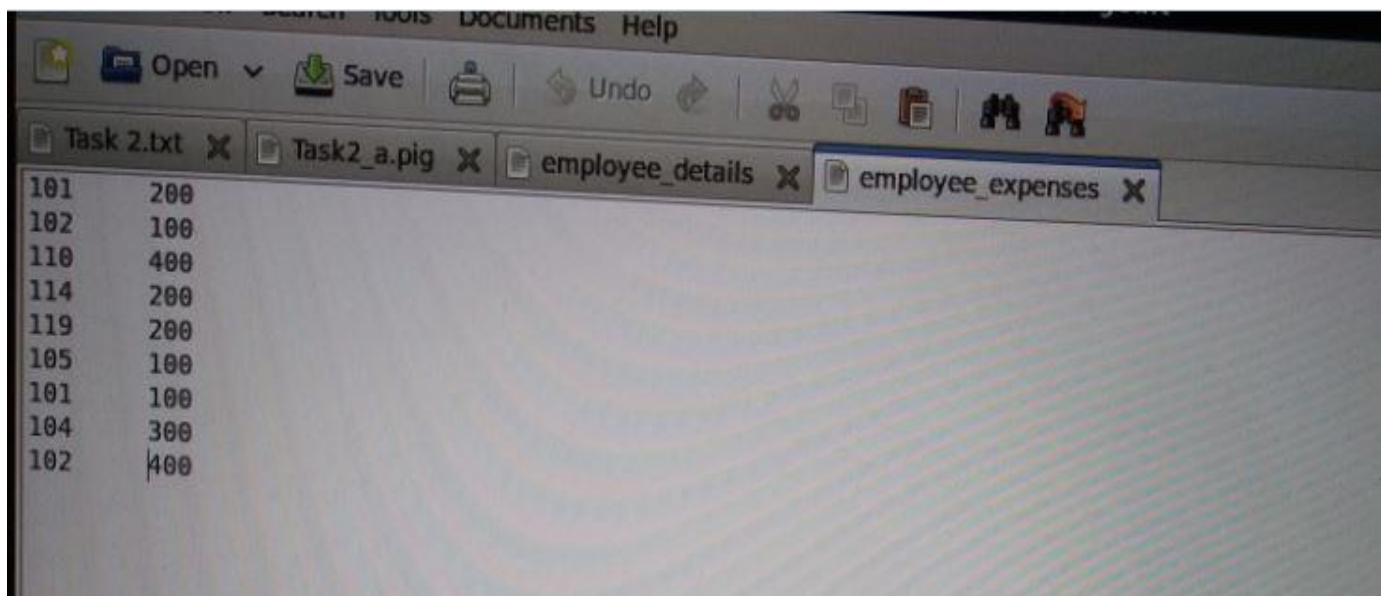
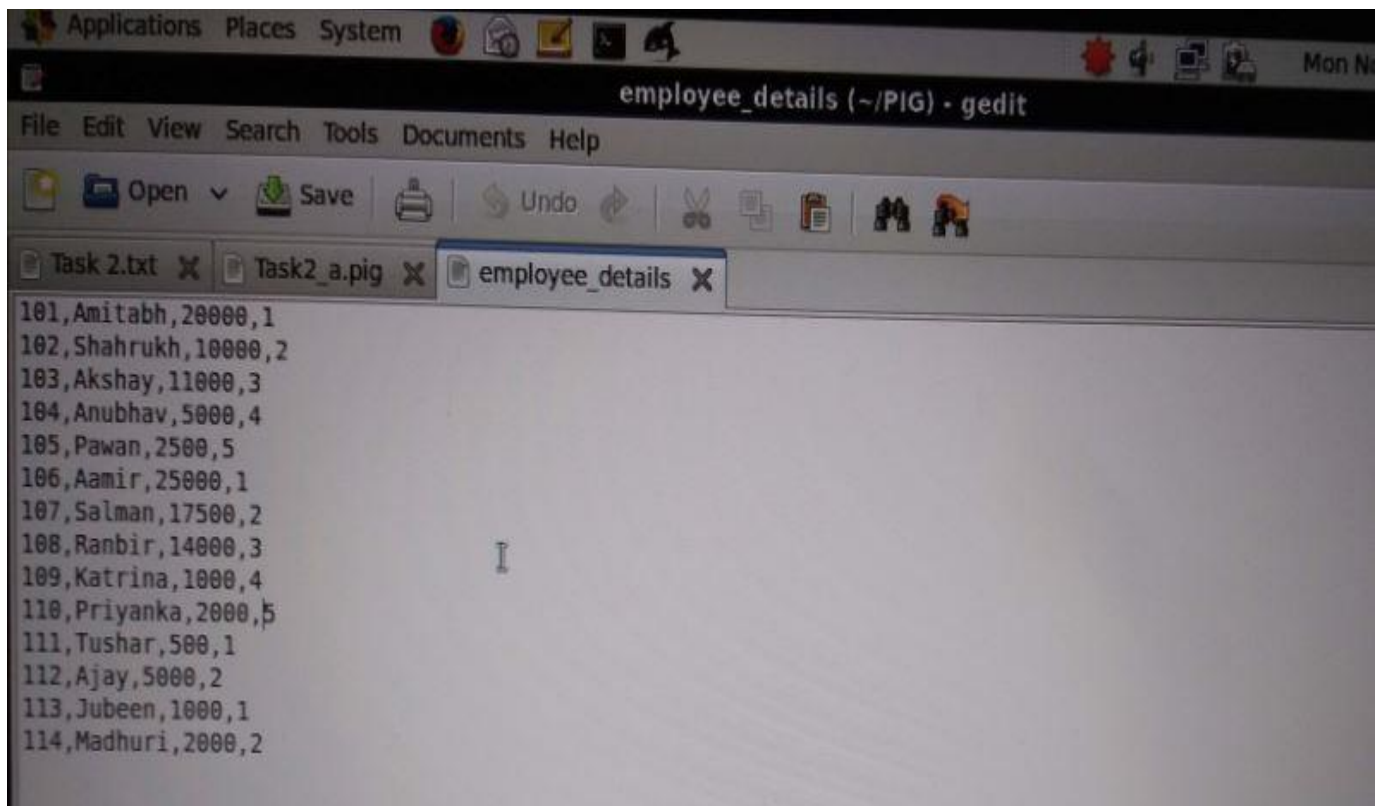
3. Output.



```
2018-11-13 04:36:36,046 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-13 04:36:36,046 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-11-13 04:36:36,080 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-11-13 04:36:36,080 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(a,2)
(in,1)
(is,2)
(of,1)
(to,1)
(PIG,1)
(the,1)
(This,2)
(file,1)
(this,1)
(count,1)
(words,1)
(number,1)
(program,2)
2018-11-13 04:36:36,215 [main] INFO org.apache.pig.Main - Pig script completed in 8 seconds and 723 milliseconds (8723 ms)
[acadgild@localhost PIG]$
```

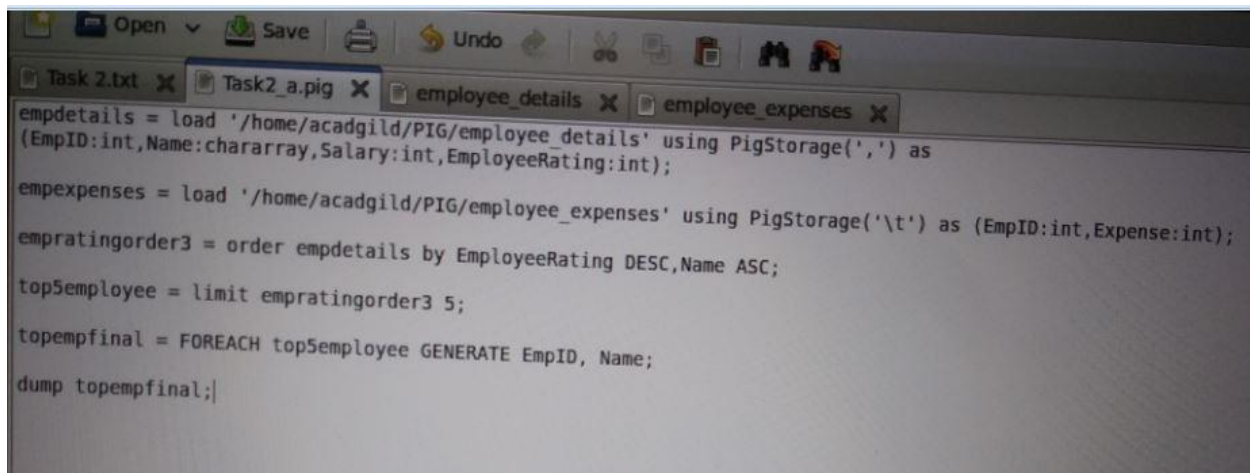
## Task 2:

1. Files used for task 2.





## 2. Script used for Task2\_a.

A screenshot of a text editor window showing a Pig script. The script is named 'Task2\_a.pig' and is located at '/home/acadgild/PIG/employee\_details'. The script contains several lines of Pig Latin code: 'empdetails = load '/home/acadgild/PIG/employee\_details' using PigStorage(',') as (EmpID:int,Name:chararray,Salary:int,EmployeeRating:int);', 'empexpenses = load '/home/acadgild/PIG/employee\_expenses' using PigStorage('\t') as (EmpID:int,Expense:int);', 'empratingorder3 = order empdetails by EmployeeRating DESC,Name ASC;', 'top5employee = limit empratingorder3 5;', 'topempfinal = FOREACH top5employee GENERATE EmpID, Name;', and 'dump topempfinal;'. The editor has tabs for 'Task2.txt', 'Task2\_a.pig', 'employee\_details', and 'employee\_expenses'.

```
empdetails = load '/home/acadgild/PIG/employee_details' using PigStorage(',') as
(EmpID:int,Name:chararray,Salary:int,EmployeeRating:int);

empexpenses = load '/home/acadgild/PIG/employee_expenses' using PigStorage('\t') as (EmpID:int,Expense:int);

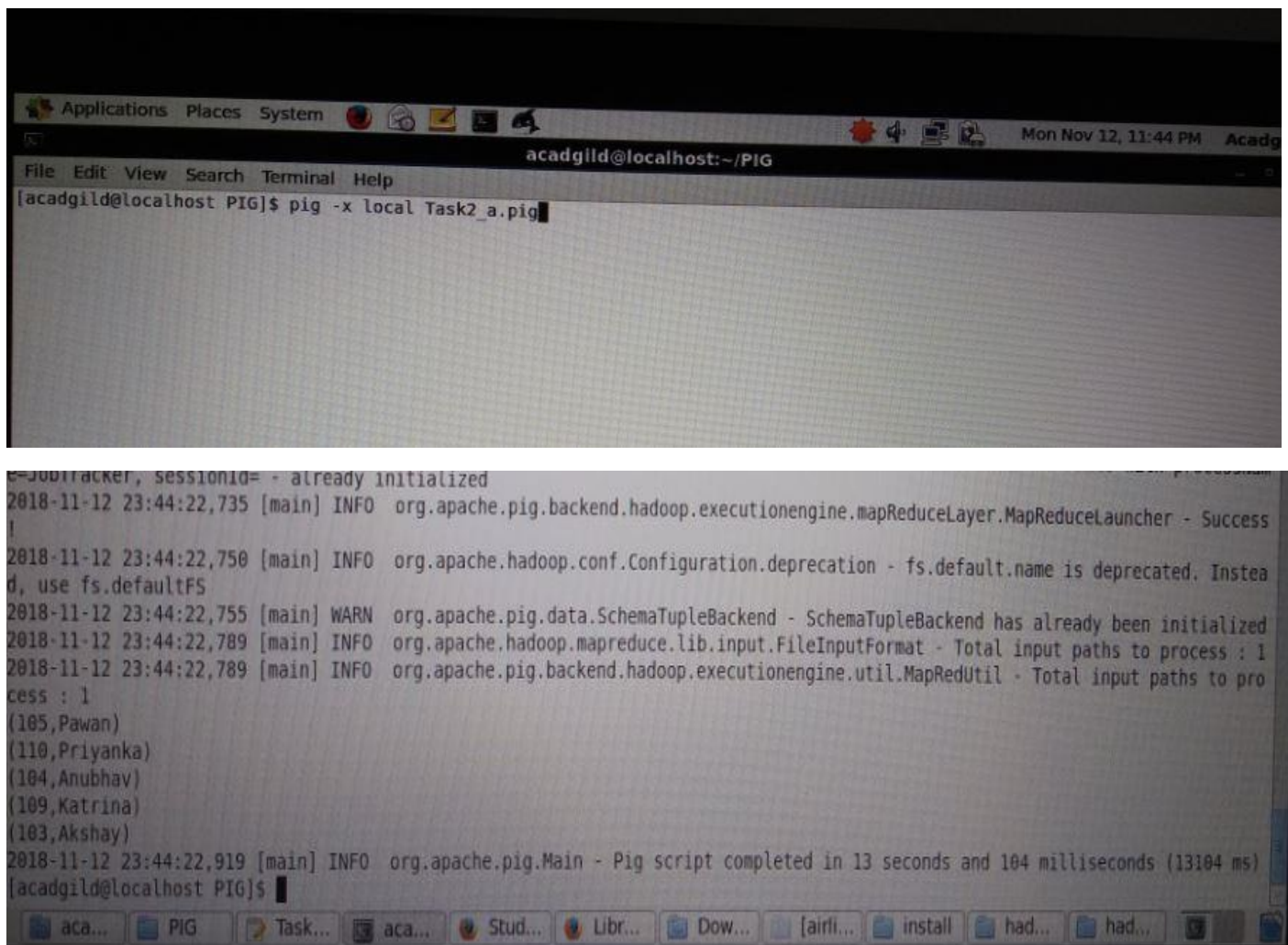
empratingorder3 = order empdetails by EmployeeRating DESC,Name ASC;

top5employee = limit empratingorder3 5;

topempfinal = FOREACH top5employee GENERATE EmpID, Name;

dump topempfinal;
```

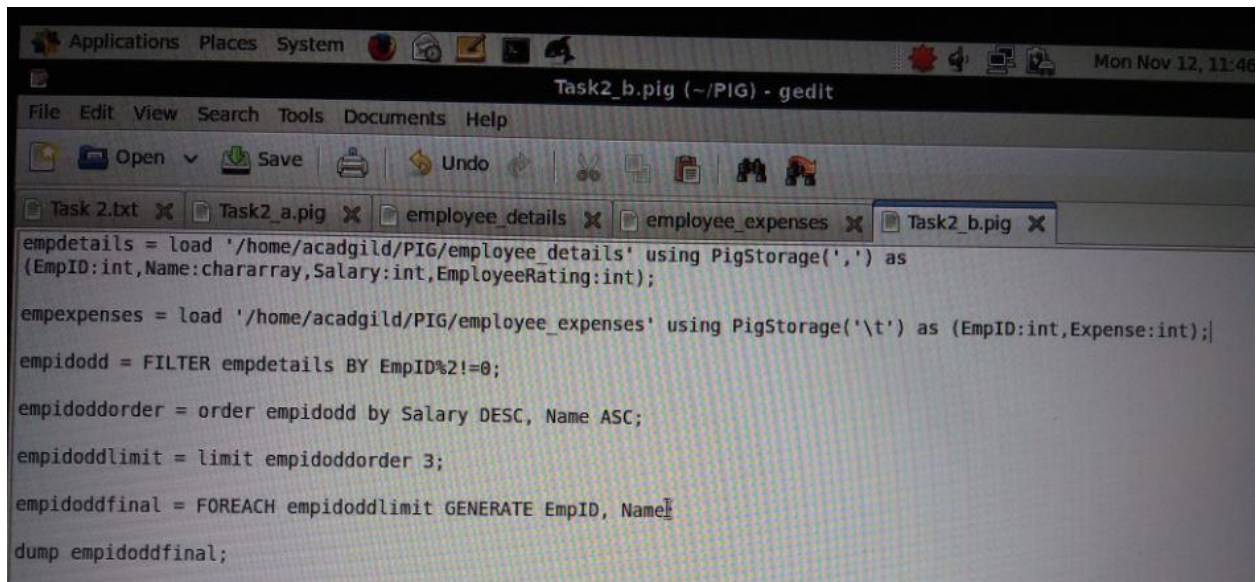
## 3. Output.

A screenshot of a terminal window showing the execution of the Pig script. The terminal is titled 'acadgild@localhost: ~/PIG'. The command 'pig -x local Task2\_a.pig' has been executed. The output shows various log messages from the Pig execution engine, including 'Success!', 'fs.default.name is deprecated', and 'Pig script completed in 13 seconds and 104 milliseconds (13104 ms)'. The final output of the script is a list of employee details: (105,Pawan), (110,Priyanka), (104,Anubhav), (109,Katrina), and (103,Akshay).

```
File Edit View Search Terminal Help
[acadgild@localhost PIG]$ pig -x local Task2_a.pig

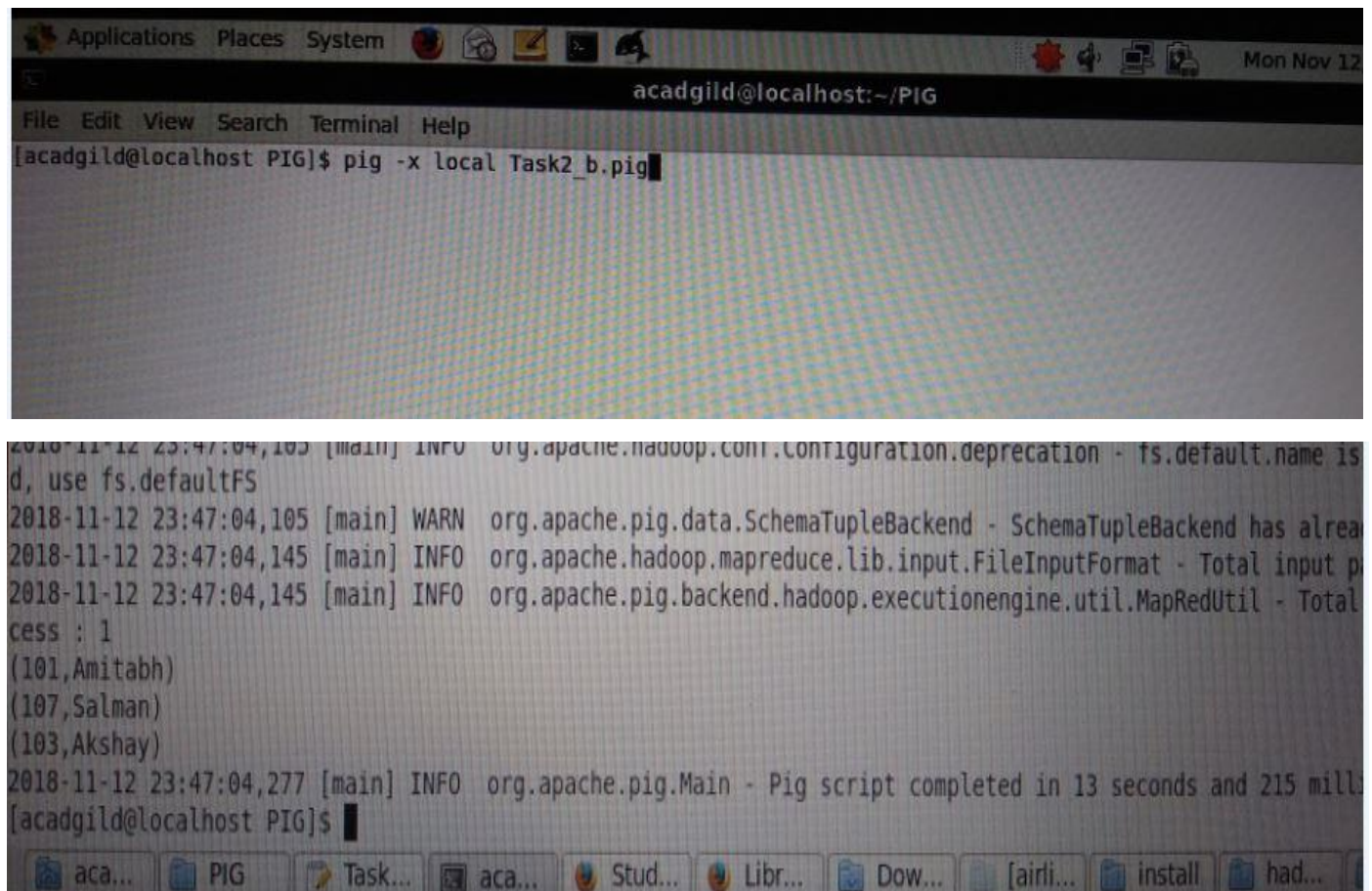
e-JobTracker, sessionId= - already initialized
2018-11-12 23:44:22,735 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-11-12 23:44:22,750 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-12 23:44:22,755 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-11-12 23:44:22,789 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-11-12 23:44:22,789 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(105,Pawan)
(110,Priyanka)
(104,Anubhav)
(109,Katrina)
(103,Akshay)
2018-11-12 23:44:22,919 [main] INFO org.apache.pig.Main - Pig script completed in 13 seconds and 104 milliseconds (13104 ms)
[acadgild@localhost PIG]$
```

#### 4. Script used for Task2\_b.



```
Task2_b.pig (~/PIG) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
Task 2.txt Task2_a.pig employee_details employee_expenses Task2_b.pig
empdetails = load '/home/acadgild/PIG/employee_details' using PigStorage(',') as
(EmpID:int,Name:chararray,Salary:int,EmployeeRating:int);
empexpenses = load '/home/acadgild/PIG/employee_expenses' using PigStorage('\t') as (EmpID:int,Expense:int);
empidodd = FILTER empdetails BY EmpID%2!=0;
empidoddorder = order empidodd by Salary DESC, Name ASC;
empidoddlimit = limit empidoddorder 3;
empidoddfinal = FOREACH empidoddlimit GENERATE EmpID, Name;
dump empidoddfinal;
```

#### 5. Output.

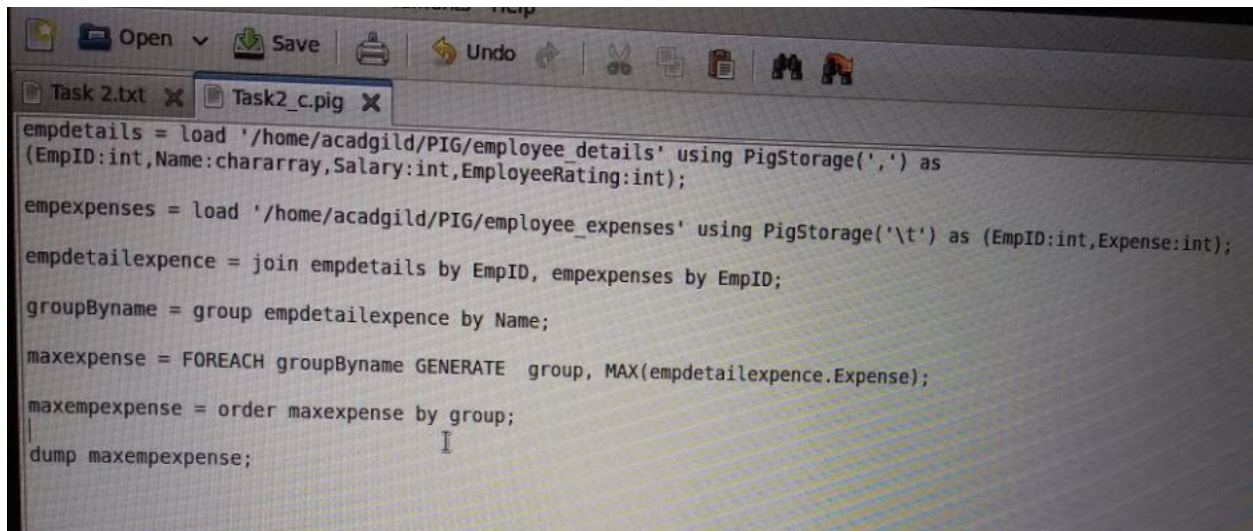


```
acadgild@localhost:~/PIG
File Edit View Search Terminal Help
[acadgild@localhost PIG]$ pig -x local Task2_b.pig

2018-11-12 23:47:04,105 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
d, use fs.defaultFS
2018-11-12 23:47:04,105 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has alrea
2018-11-12 23:47:04,145 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input p
2018-11-12 23:47:04,145 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total
cess : 1
(101,Amitabh)
(107,Salman)
(103,Akshay)
2018-11-12 23:47:04,277 [main] INFO org.apache.pig.Main - Pig script completed in 13 seconds and 215 mill
[acadgild@localhost PIG]$
```

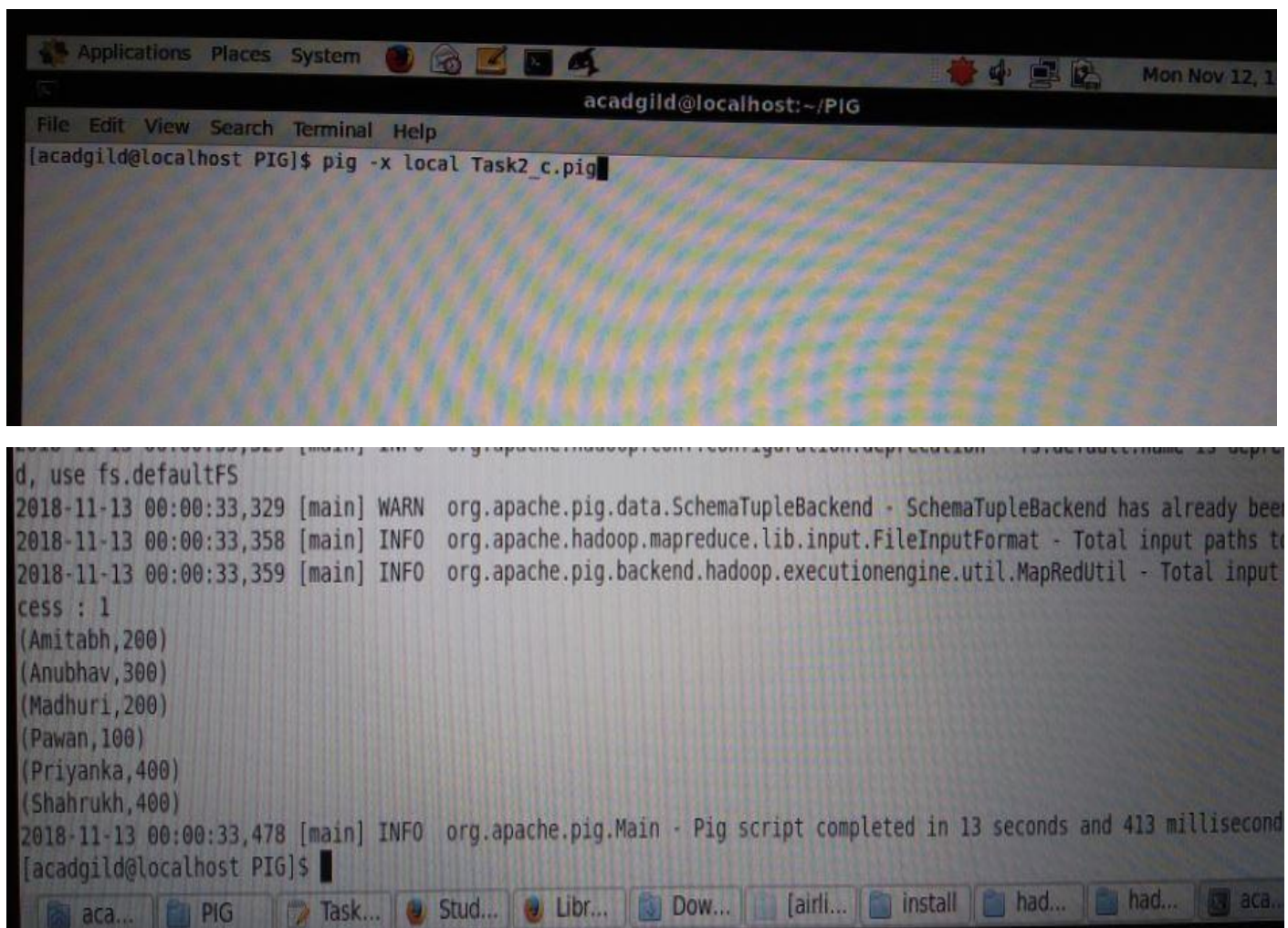


## 6. Script used for Task2\_c.



```
empdetails = load '/home/acadgild/PIG/employee_details' using PigStorage(',') as
(EmpID:int,Name:chararray,Salary:int,EmployeeRating:int);
empexpenses = load '/home/acadgild/PIG/employee_expenses' using PigStorage('\t') as (EmpID:int,Expense:int);
empdetailexpense = join empdetails by EmpID, empexpenses by EmpID;
groupByname = group empdetailexpense by Name;
maxexpense = FOREACH groupByname GENERATE group, MAX(empdetailexpense.Expense);
maxempexpense = order maxexpense by group;
dump maxempexpense;
```

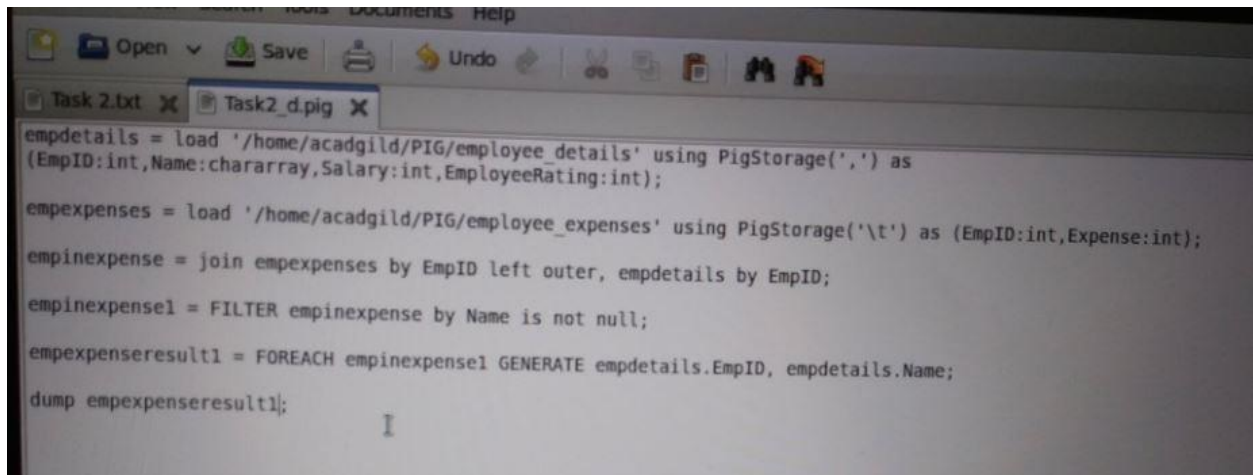
## 7. Output.



```
acadmild@localhost:~/PIG
[acadmild@localhost PIG]$ pig -x local Task2_c.pig

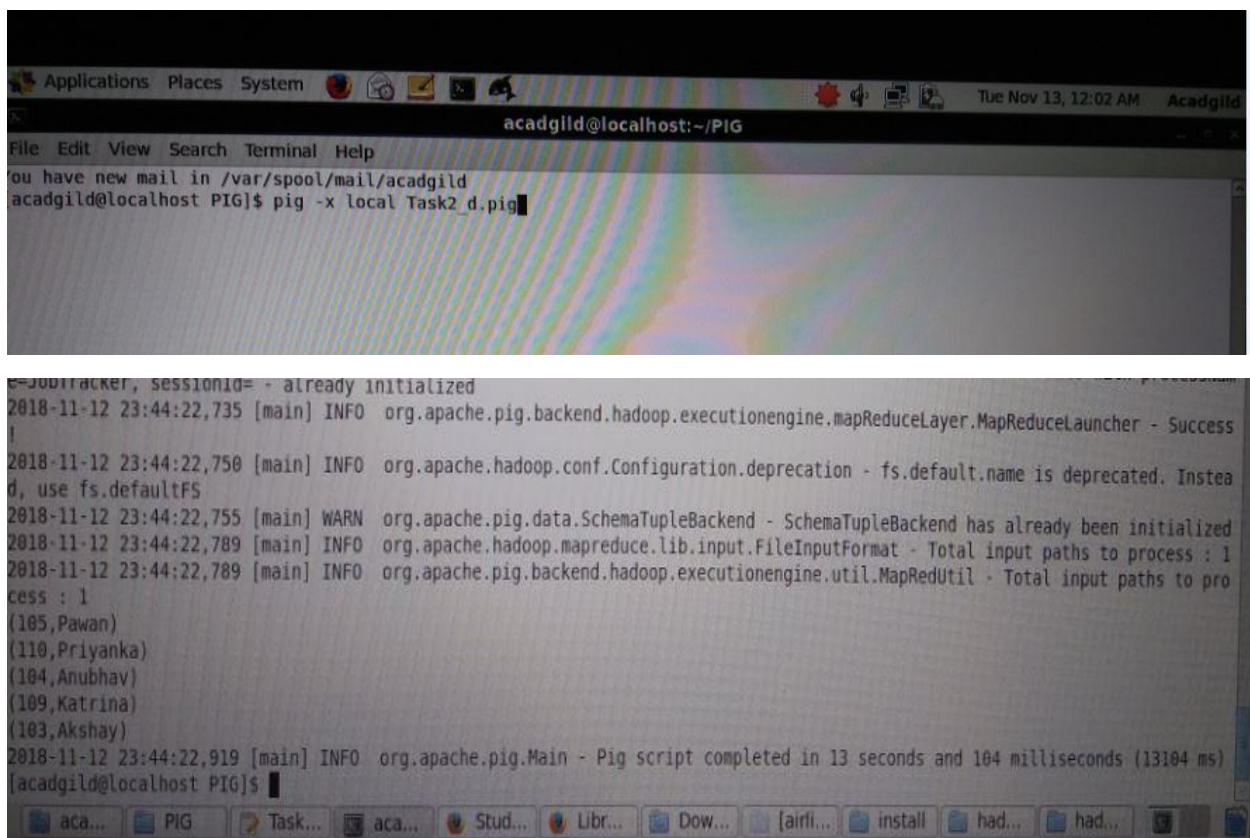
2018-11-13 00:00:33,329 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-11-13 00:00:33,358 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
(Amitabh,200)
(Anubhav,300)
(Madhuri,200)
(Pawan,100)
(Priyanka,400)
(Shahrukh,400)
2018-11-13 00:00:33,478 [main] INFO org.apache.pig.Main - Pig script completed in 13 seconds and 413 milliseconds
[acadmild@localhost PIG]$
```

## 8. Script used for Task2\_d.

A screenshot of a text editor window with two tabs: 'Task2.txt' and 'Task2\_d.pig'. The 'Task2\_d.pig' tab is active, displaying a Pig script. The script defines two tables, 'empdetails' and 'empexpenses', and performs a join operation. The script is as follows:

```
empdetails = load '/home/acadgild/PIG/employee_details' using PigStorage(',') as
(EmpID:int,Name:chararray,Salary:int,EmployeeRating:int);
empexpenses = load '/home/acadgild/PIG/employee_expenses' using PigStorage('\t') as (EmpID:int,Expense:int);
empinexpense = join empexpenses by EmpID left outer, empdetails by EmpID;
empinexpense1 = FILTER empinexpense by Name is not null;
empexpenseresult1 = FOREACH empinexpense1 GENERATE empdetails.EmpID, empdetails.Name;
dump empexpenseresult1;
```

## 9. Output.

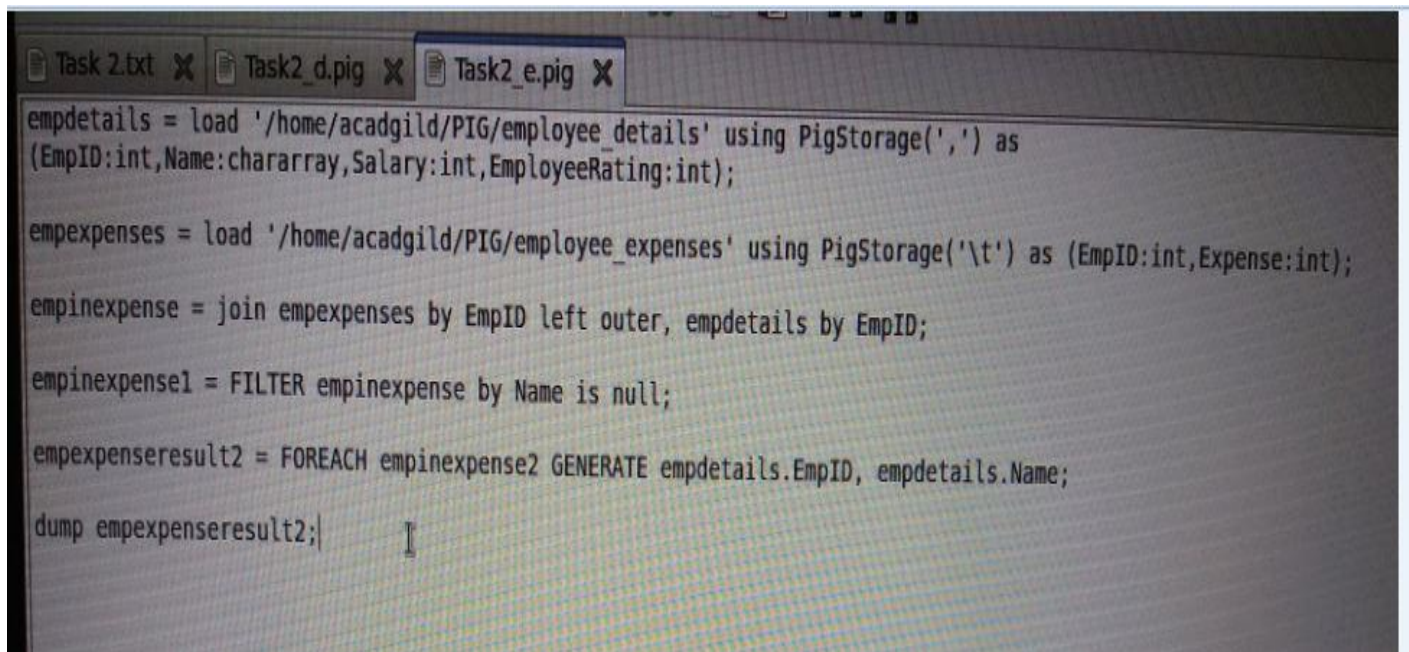
A screenshot of a terminal window showing the execution of a Pig script. The terminal title is 'acadgild@localhost:~/PIG'. The command executed is 'pig -x local Task2\_d.pig'. The output shows the script execution details, including log messages from the Pig backend and the final output of the script. The output is as follows:

```
acadgild@localhost:~/PIG
File Edit View Search Terminal Help
You have new mail in /var/spool/mail/acadgild
acadgild@localhost PIG$ pig -x local Task2_d.pig

e=jobtracker, sessionId= - already initialized
2018-11-12 23:44:22,735 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-11-12 23:44:22,750 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-11-12 23:44:22,755 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-11-12 23:44:22,789 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-11-12 23:44:22,789 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(105,Pawan)
(110,Priyanka)
(104,Anubhav)
(109,Katrina)
(103,Akshay)
2018-11-12 23:44:22,919 [main] INFO org.apache.pig.Main - Pig script completed in 13 seconds and 104 milliseconds (13104 ms)
acadgild@localhost PIG$
```



10. Script used for Task2\_e.



```
Task 2.txt X Task2_d.pig X Task2_e.pig X
empdetails = load '/home/acadgild/PIG/employee_details' using PigStorage(',') as
(EmpID:int,Name:chararray,Salary:int,EmployeeRating:int);

empexpenses = load '/home/acadgild/PIG/employee_expenses' using PigStorage('\t') as (EmpID:int,Expense:int);

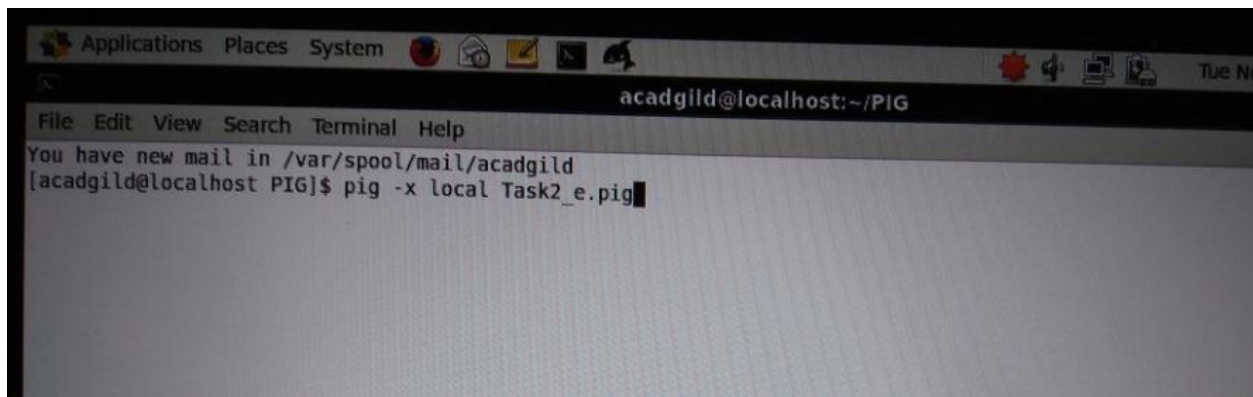
empinexpense = join empexpenses by EmpID left outer, empdetails by EmpID;

empinexpense1 = FILTER empinexpense by Name is null;

empexpenseresult2 = FOREACH empinexpense2 GENERATE empdetails.EmpID, empdetails.Name;

dump empexpenseresult2;
```

11. Output.

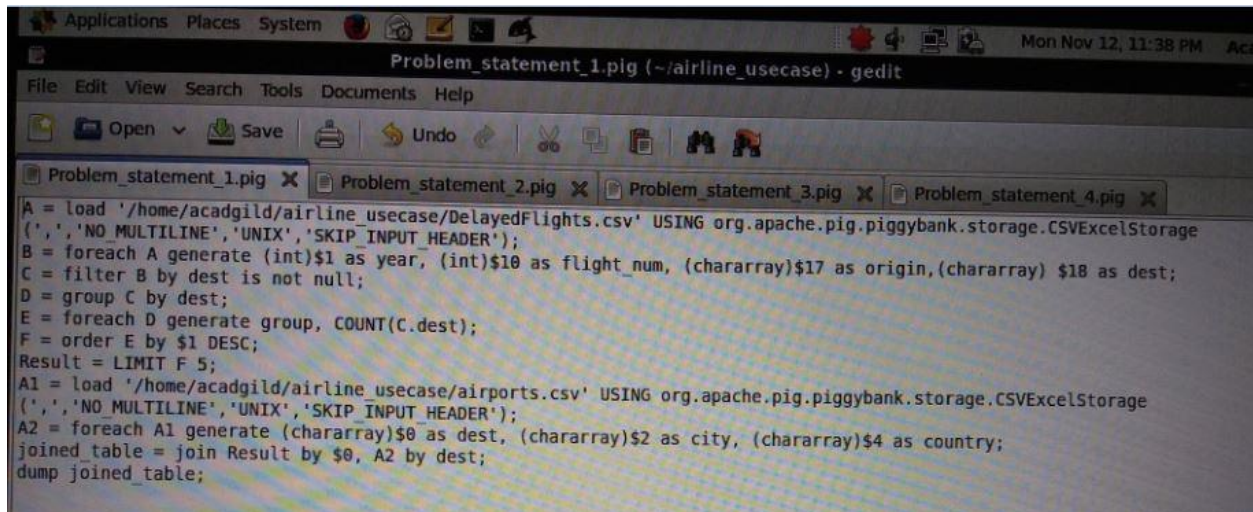


```
Applications Places System acadgild@localhost:~/PIG
File Edit View Search Terminal Help
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost PIG]$ pig -x local Task2_e.pig
```



### Task 3:

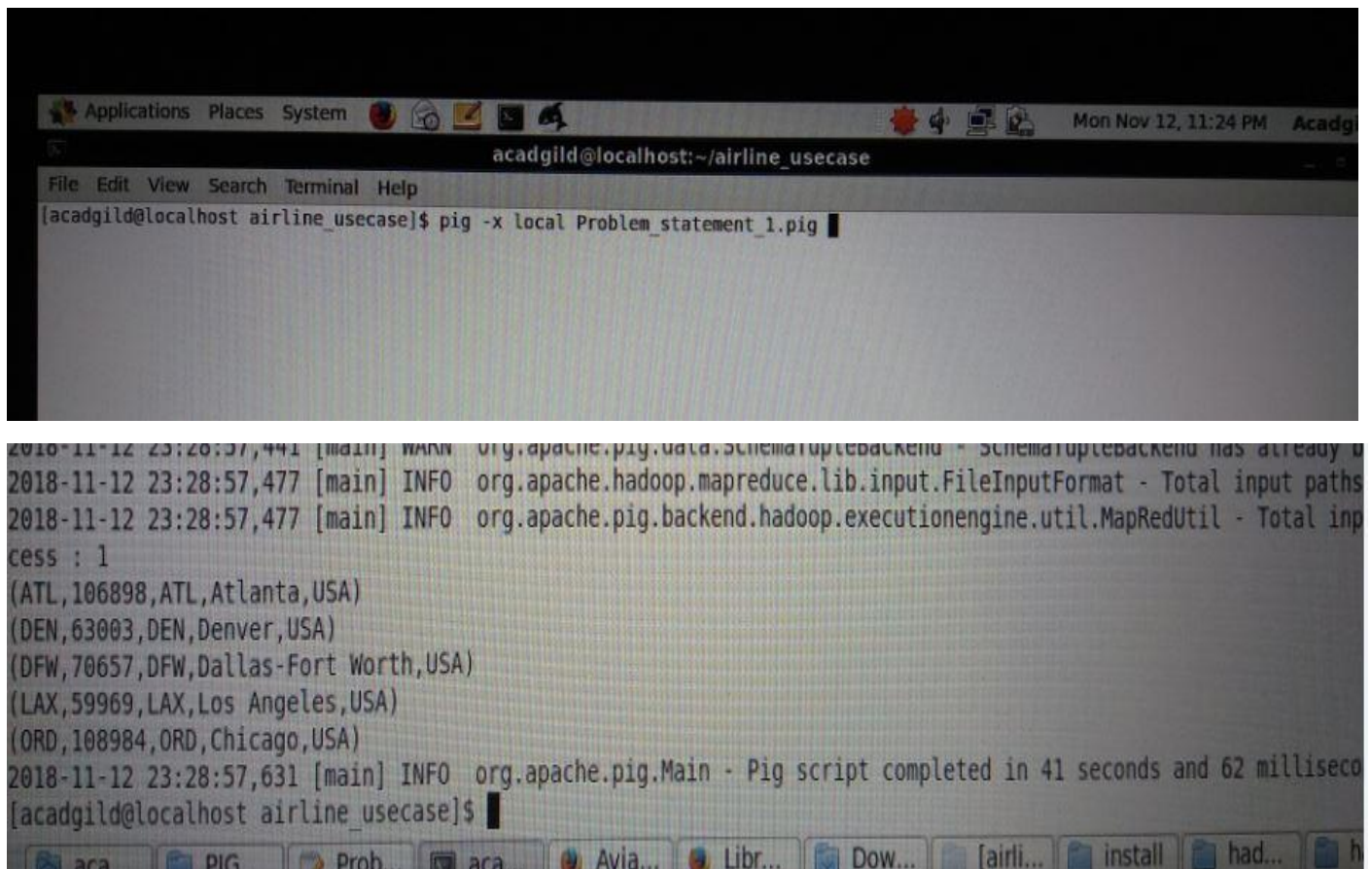
#### 1. Script used for Problem statement 1.



The screenshot shows a text editor window titled "Problem\_statement\_1.pig (~/airline\_usecase) - gedit". The editor contains a Pig script that processes flight data. The script loads a CSV file of delayed flights, generates rows for each flight, filters for non-null destinations, groups by destination, and orders by the count of flights. It then loads a CSV file of airport information, generates rows for each airport, and joins the two datasets by destination. Finally, it dumps the joined table.

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

#### 2. Output.

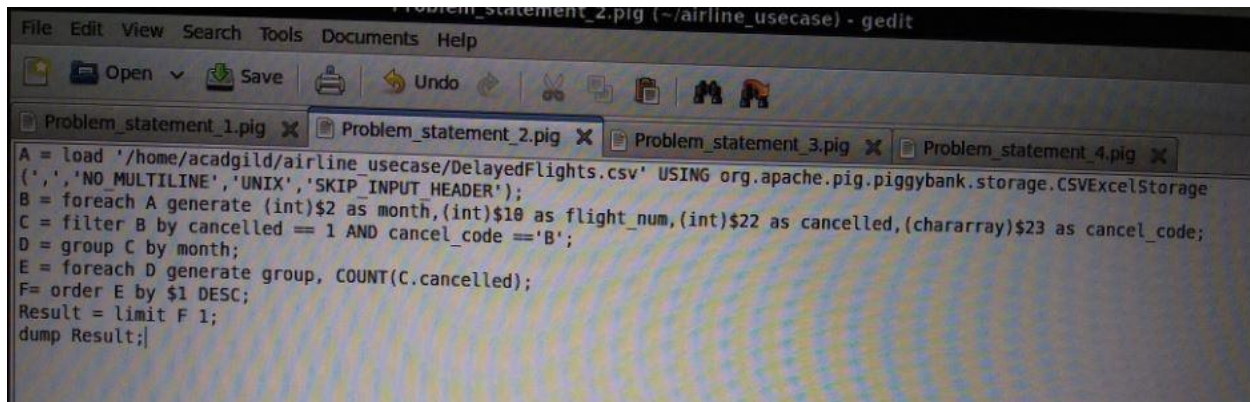


The screenshot shows a terminal window titled "acadgild@localhost:~/airline\_usecase". The user has executed the command "pig -x local Problem\_statement\_1.pig". The output shows the execution of the Pig script, including log messages from the Hadoop MapReduce framework and the final output of the joined table.

```
acadgild@localhost:~/airline_usecase
[acadgild@localhost airline_usecase]$ pig -x local Problem_statement_1.pig

2018-11-12 23:28:57,441 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already b
2018-11-12 23:28:57,477 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths
2018-11-12 23:28:57,477 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total inp
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)
2018-11-12 23:28:57,631 [main] INFO org.apache.pig.Main - Pig script completed in 41 seconds and 62 milliseco
[acadgild@localhost airline_usecase]$
```

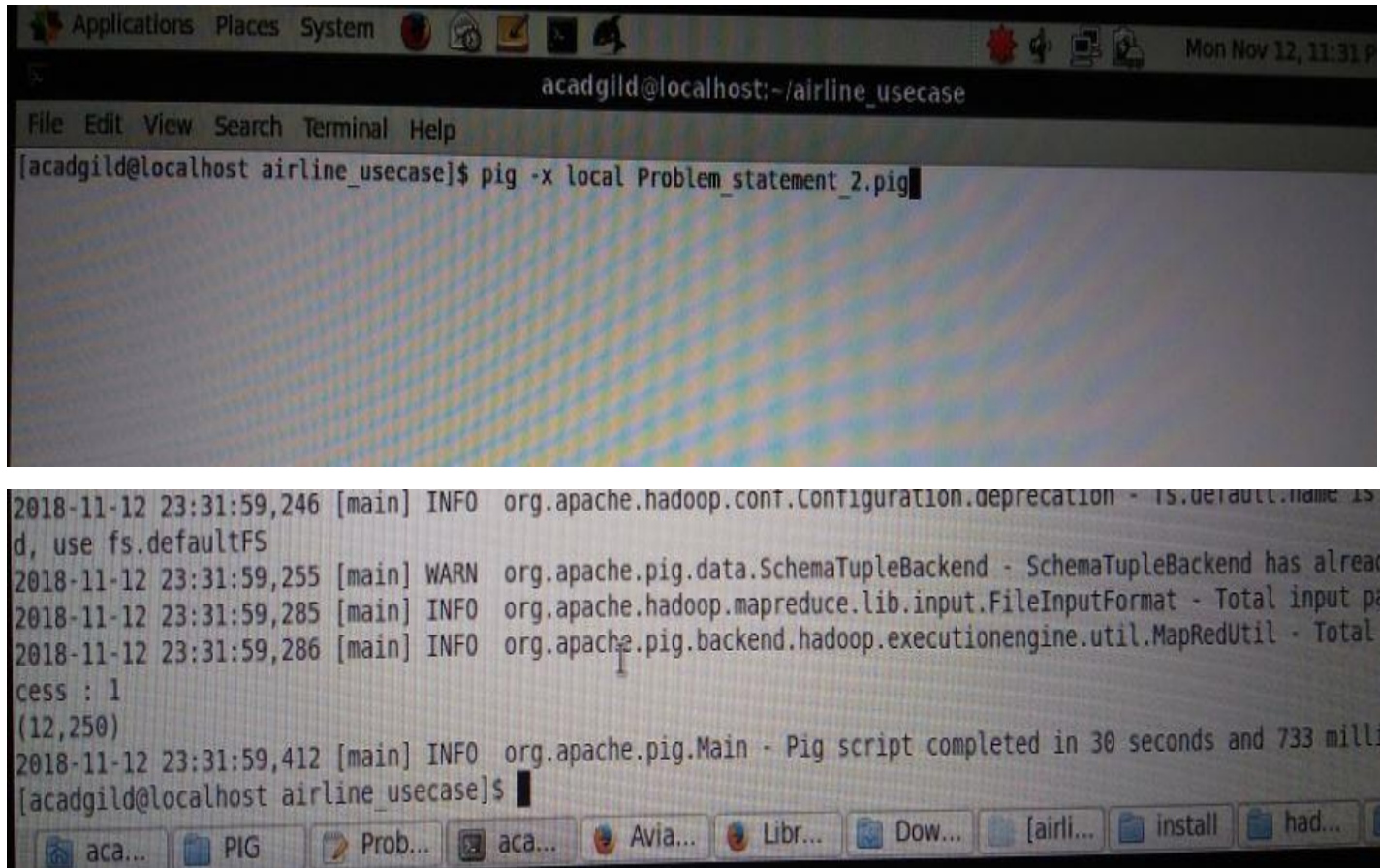
### 3. Script used for Problem Statement 2.



The screenshot shows a gedit editor window titled "Problem\_statement\_2.pig (~/airline\_usecase) - gedit". The editor has a menu bar (File, Edit, View, Search, Tools, Documents, Help) and a toolbar with icons for Open, Save, Print, Undo, and others. The script content is as follows:

```
Problem_statement_1.pig x Problem_statement_2.pig x Problem_statement_3.pig x Problem_statement_4.pig x
A = load '/home/acadgild/airline_usecase/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;
```

### 4. Output.



The screenshot shows a terminal window titled "acadgild@localhost:~/airline\_usecase". The terminal has a menu bar (File, Edit, View, Search, Terminal, Help) and a toolbar. The command executed is:

```
[acadgild@localhost airline_usecase]$ pig -x local Problem_statement_2.pig
```

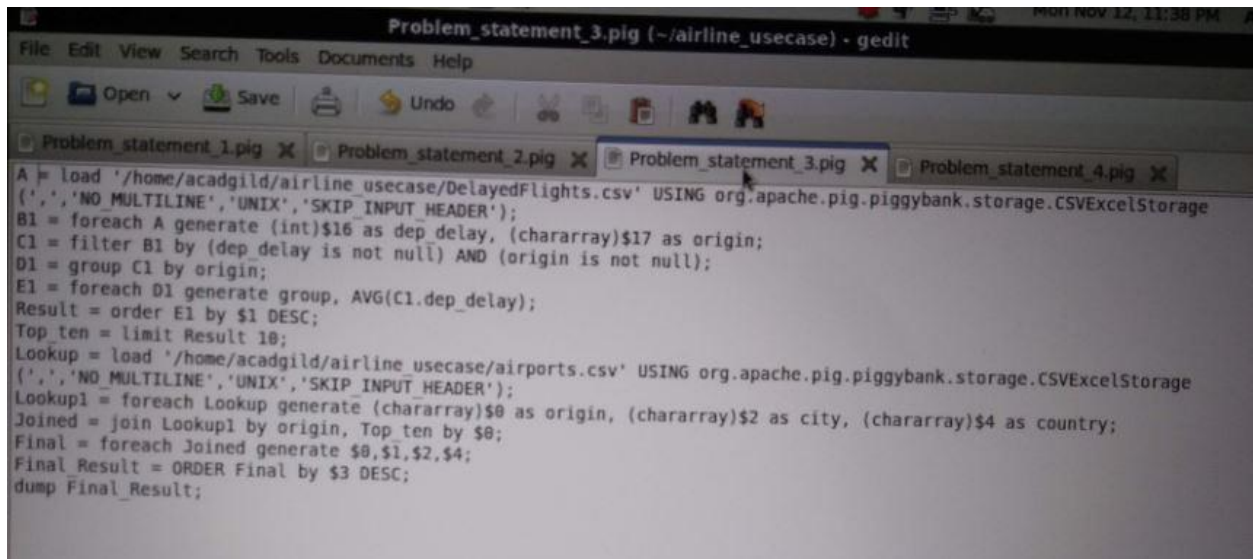
The output of the command is as follows:

```
2018-11-12 23:31:59,246 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
d, use fs.defaultFS
2018-11-12 23:31:59,255 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already
2018-11-12 23:31:59,285 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input pa
2018-11-12 23:31:59,286 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total
cess : 1
(12,250)
2018-11-12 23:31:59,412 [main] INFO org.apache.pig.Main - Pig script completed in 30 seconds and 733 mill
[acadgild@localhost airline_usecase]$
```

The terminal window also shows a taskbar at the bottom with icons for "aca...", "PIG", "Prob...", "aca...", "Avia...", "Libr...", "Dow...", "[airli...", "install", and "had...".



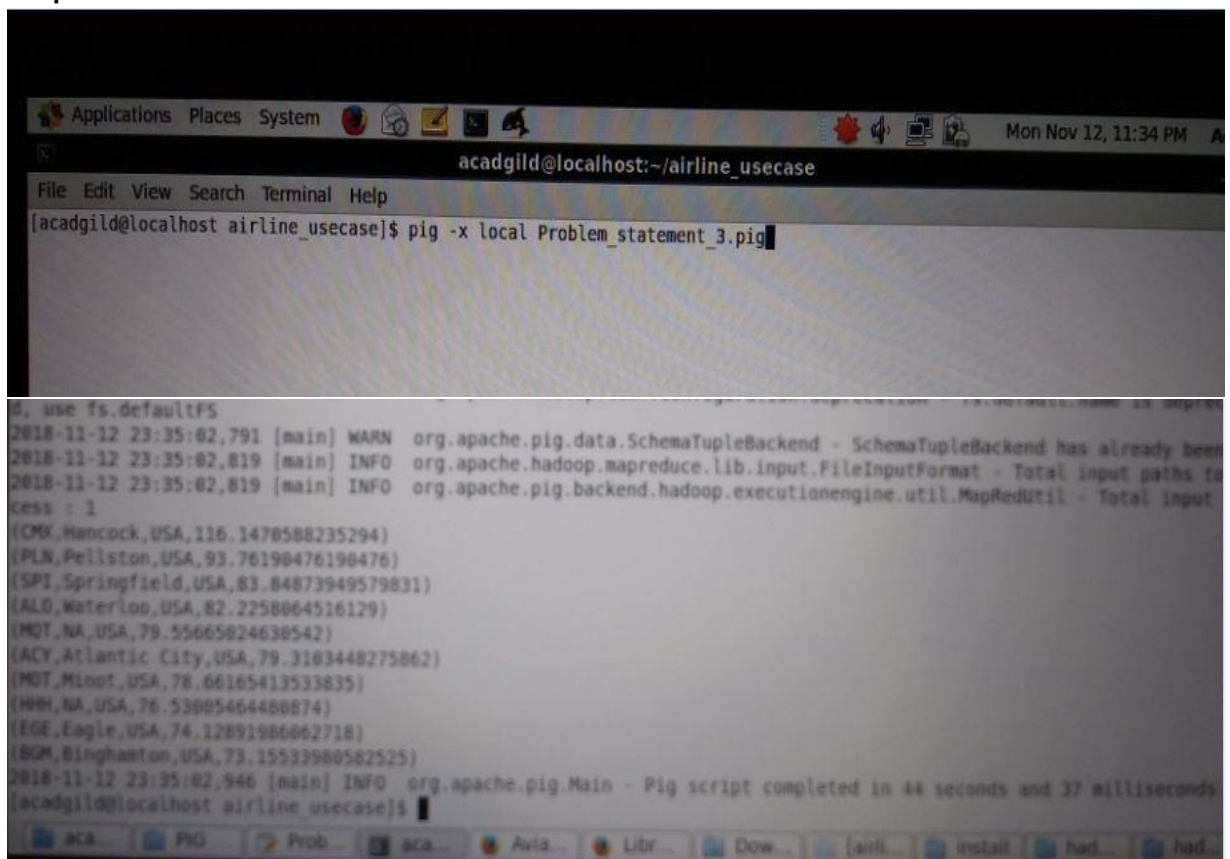
## 5. Script used for Problem Statement 3.



The screenshot shows a gedit editor window titled "Problem\_statement\_3.pig (~/airline\_usecase) - gedit". The editor contains a Pig script that processes flight data. The script loads a CSV file of delayed flights, filters for non-null delay and origin, groups by origin, and calculates the average delay. It then loads a CSV file of airport information, joins it with the flight data, and orders the results by the average delay in descending order, limiting the output to the top 10 airports. The script is as follows:

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

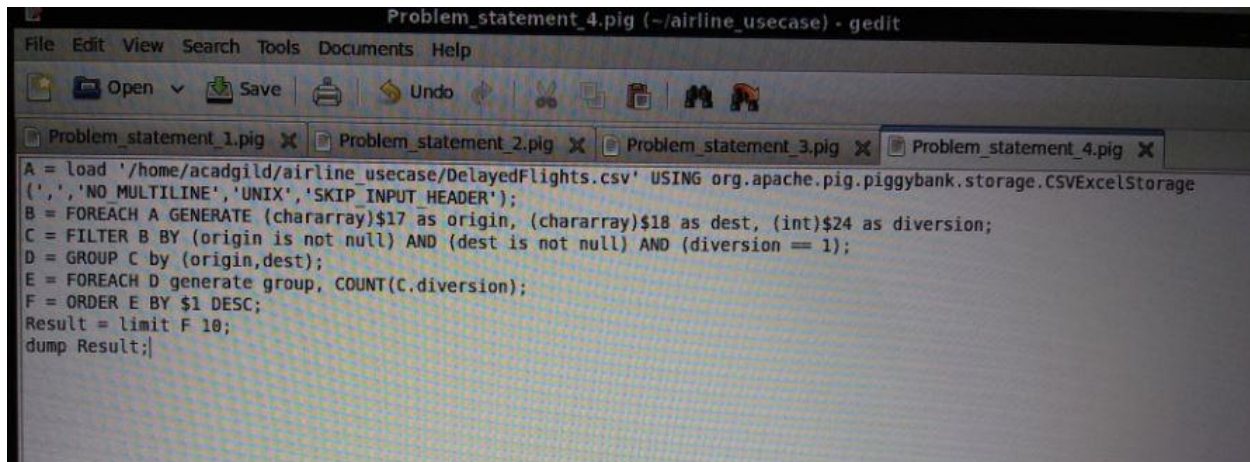
## 6. Output.



The screenshot shows a terminal window titled "acadgild@localhost:~/airline\_usecase". The user has executed the command `pig -x local Problem_statement_3.pig`. The output shows the execution of the Pig script, including warnings and information messages. The final output is a list of the top 10 airports by average delay, with columns for origin, city, country, and average delay.

```
2018-11-12 23:35:02,791 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been
2018-11-12 23:35:02,819 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to
2018-11-12 23:35:02,819 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
cess : 1
(CMX,Hancock,USA,116.1470588235294)
(PLN,Pellston,USA,93.76190476190476)
(SPI,Springfield,USA,83.84873949579831)
(ALD,Waterloo,USA,82.2258064516129)
(MOT,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)
2018-11-12 23:35:02,946 [main] INFO org.apache.pig.Main - Pig script completed in 44 seconds and 37 milliseconds
acadgild@localhost airline_usecase]$
```

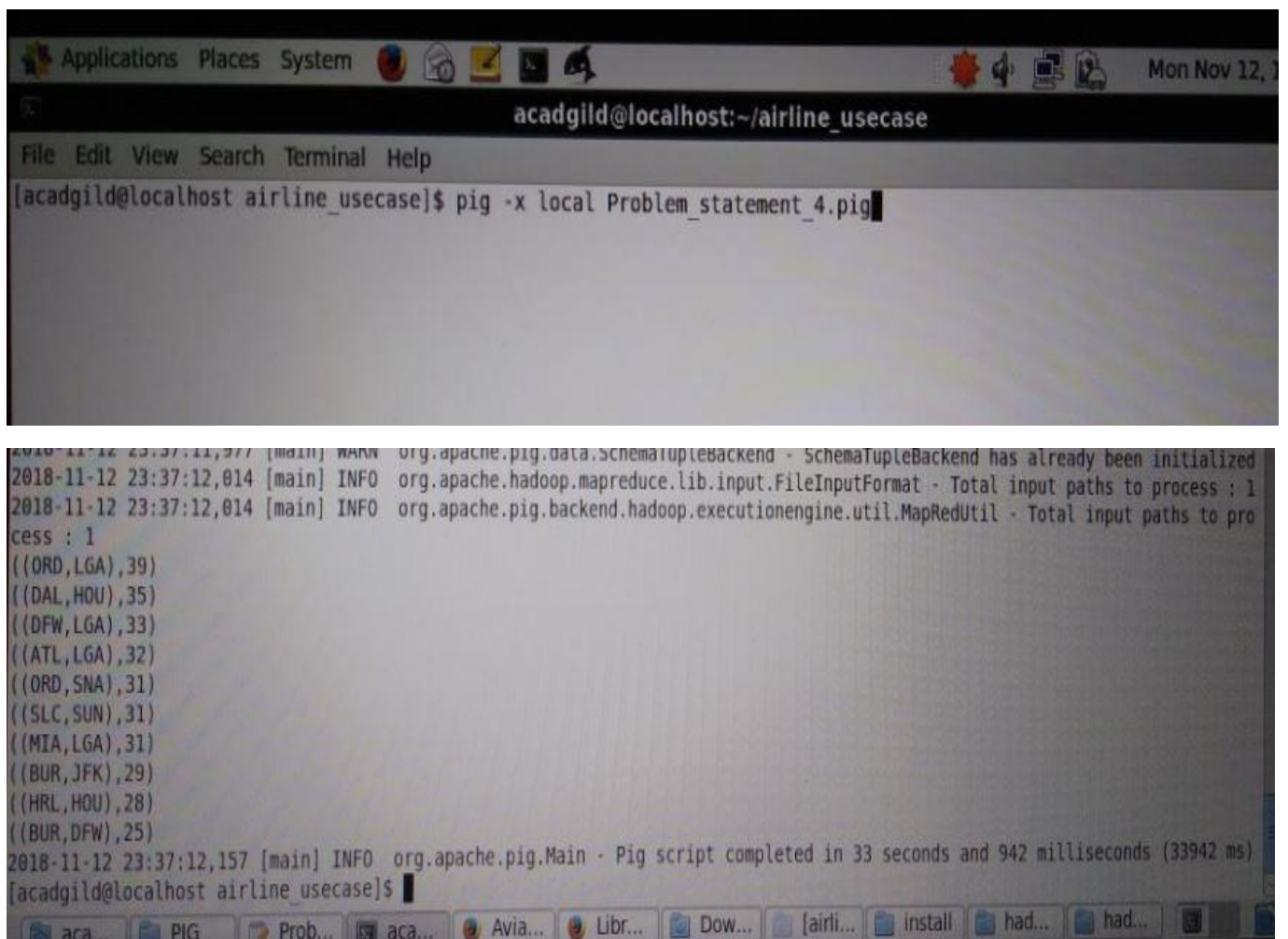
#### 7. Script used for Problem Statement 4.



The screenshot shows a gedit window titled "Problem\_statement\_4.pig (~/airline\_usecase) - gedit". The window contains a Pig script with the following code:

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

#### 8. Output.



The screenshot shows a terminal window titled "acadgild@localhost:~/airline\_usecase". The terminal displays the command to execute the Pig script and its output:

```
[acadgild@localhost airline_usecase]$ pig -x local Problem_statement_4.pig
```

The output of the script is as follows:

```
2018-11-12 23:37:11,377 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-11-12 23:37:12,014 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-11-12 23:37:12,014 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 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)
2018-11-12 23:37:12,157 [main] INFO org.apache.pig.Main - Pig script completed in 33 seconds and 942 milliseconds (33942 ms)
[acadgild@localhost airline_usecase]$
```