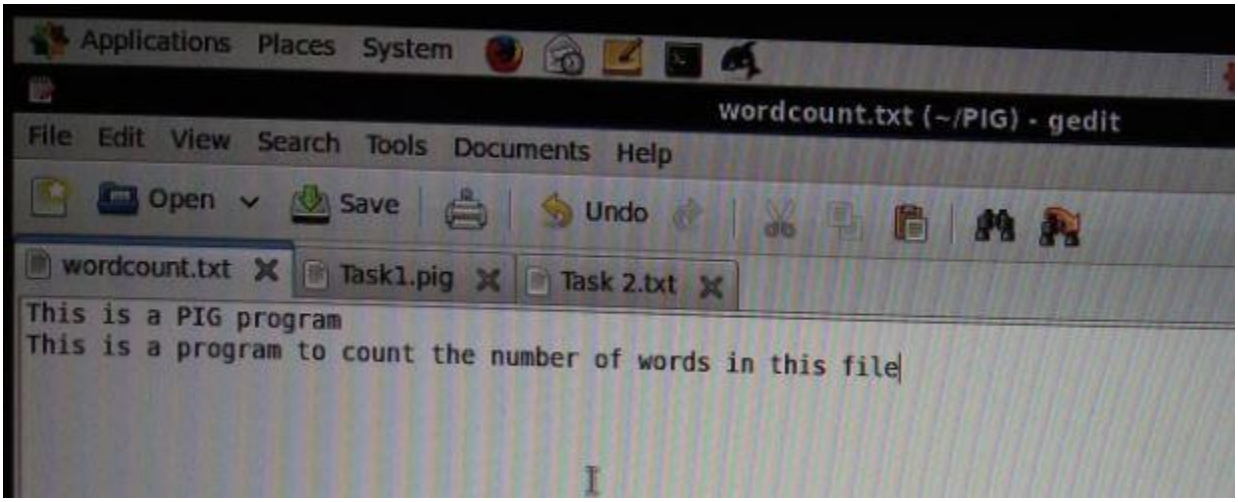
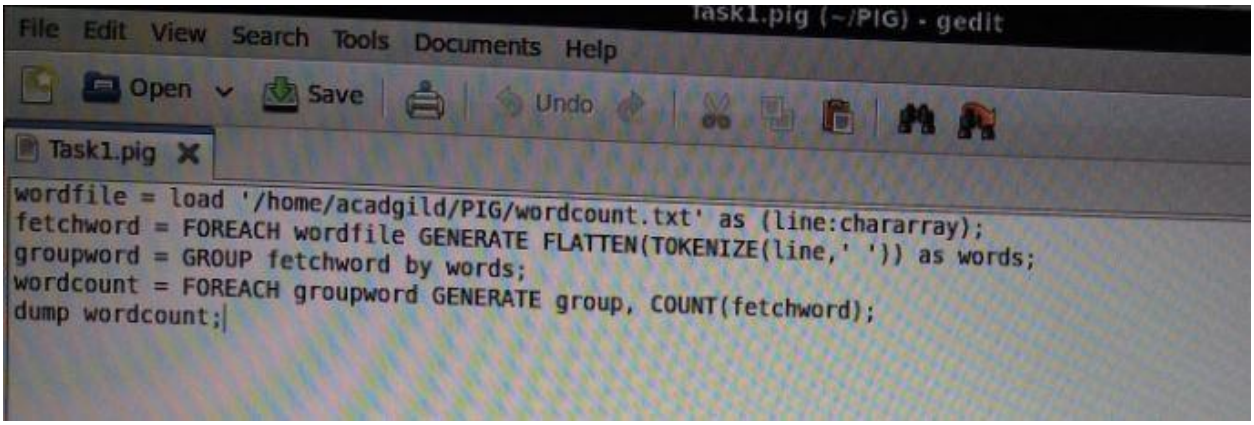
Assignment_7.1

Task 1:

1. File used for word count script.



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



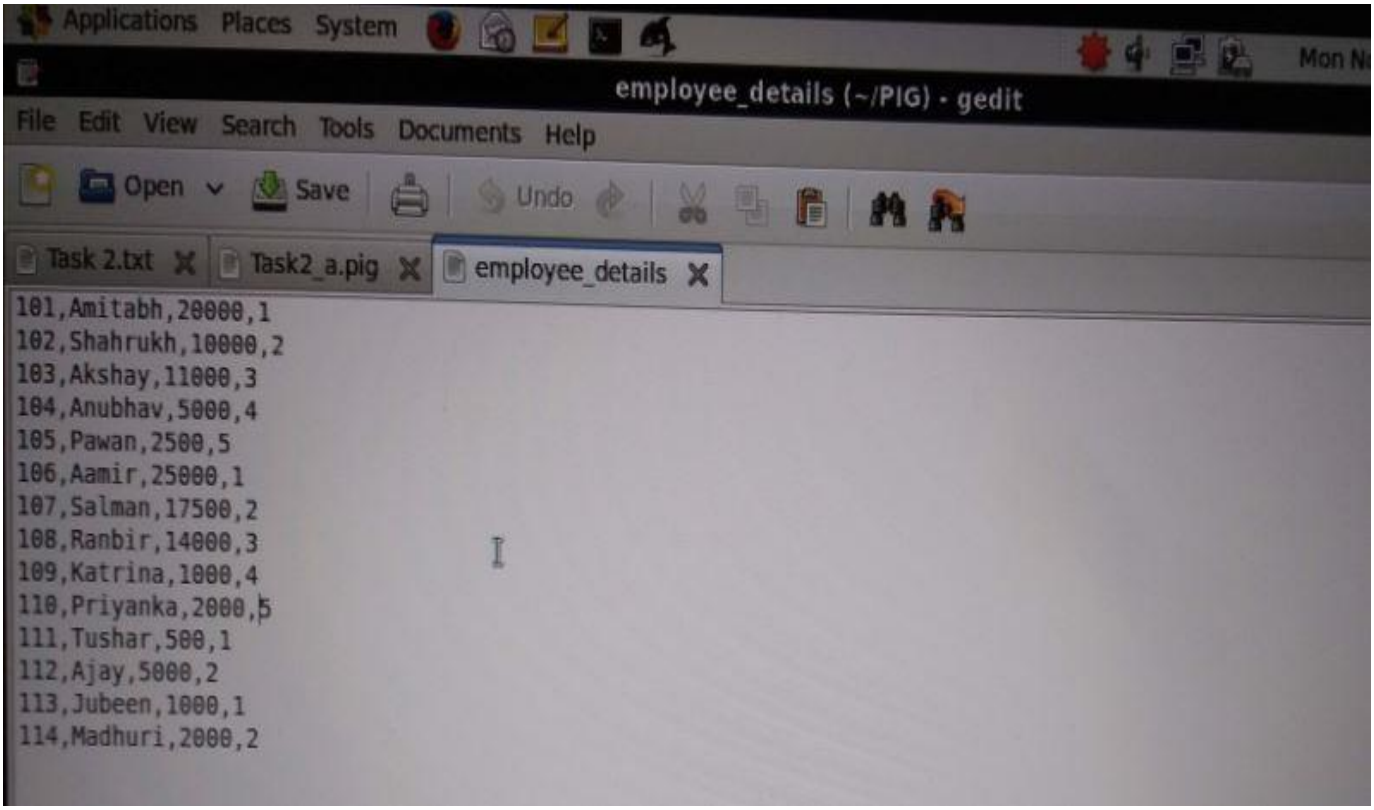
3. Output.

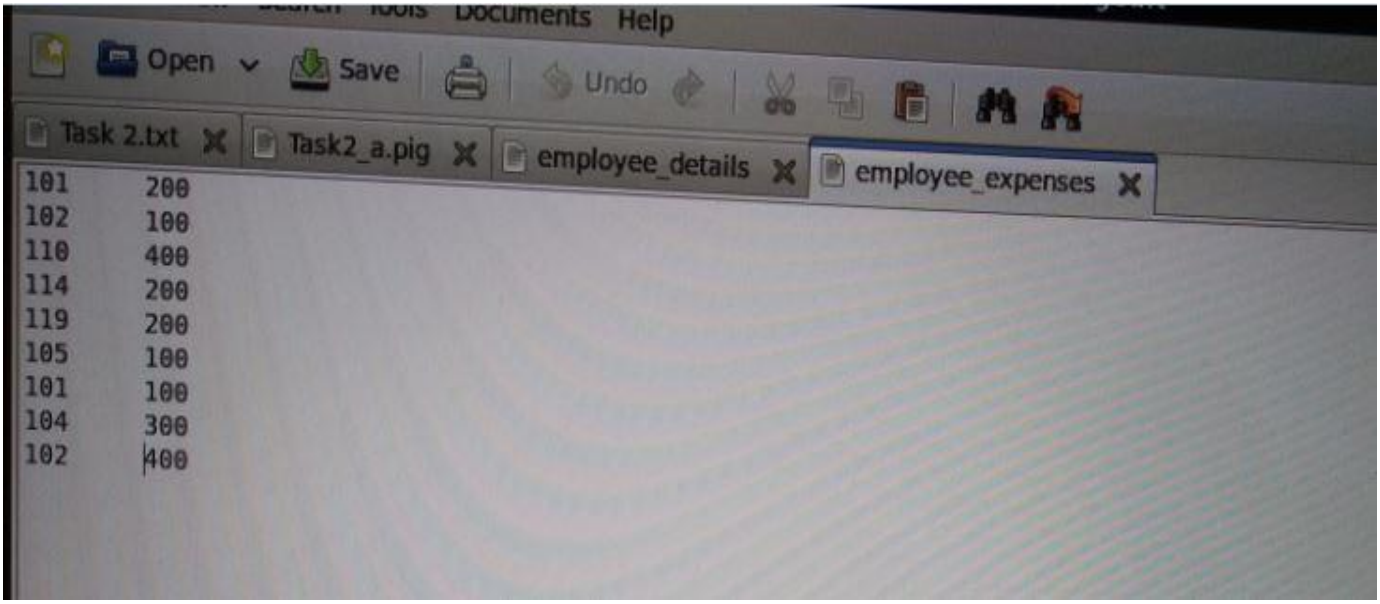
```
Applications  Places  System
acadgild@localhost:~/PIG
File  Edit  View  Search  Terminal  Help
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost PIG]$ pig -x local Task1.pig
```

```
2018-11-13 04:36:36,046 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, 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 pro
cess : 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.**

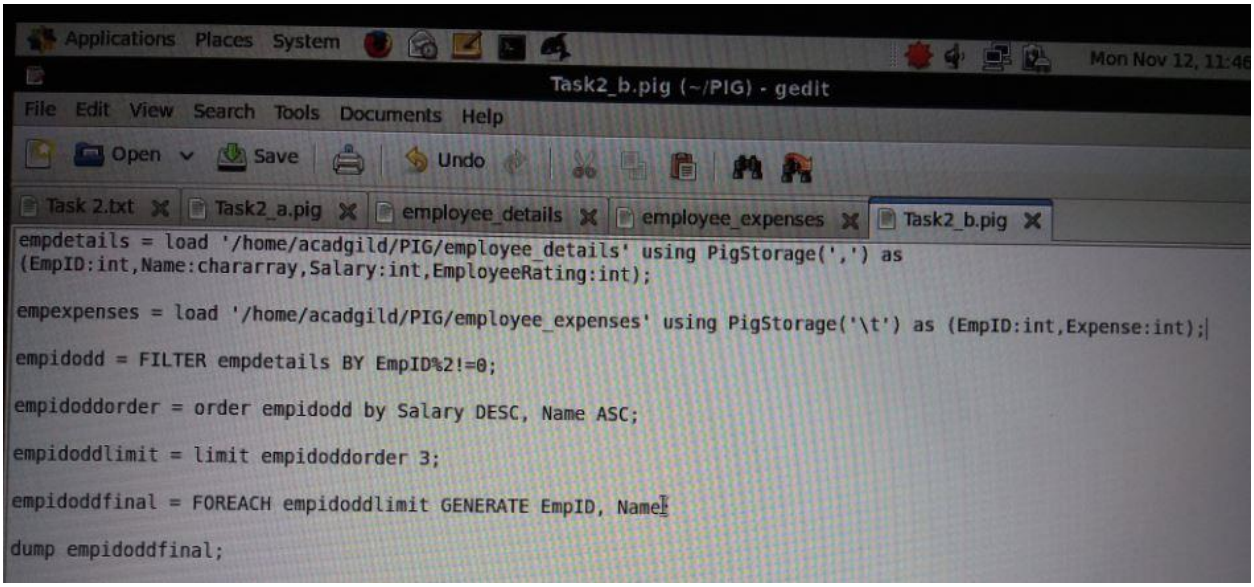
```
empdetails = load '/home/acadgild/PIG/employee_details' using PigStorage(',') as
(EmpID:int,Name:chararray,Salary:int,EmployeeRating:int);

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.**

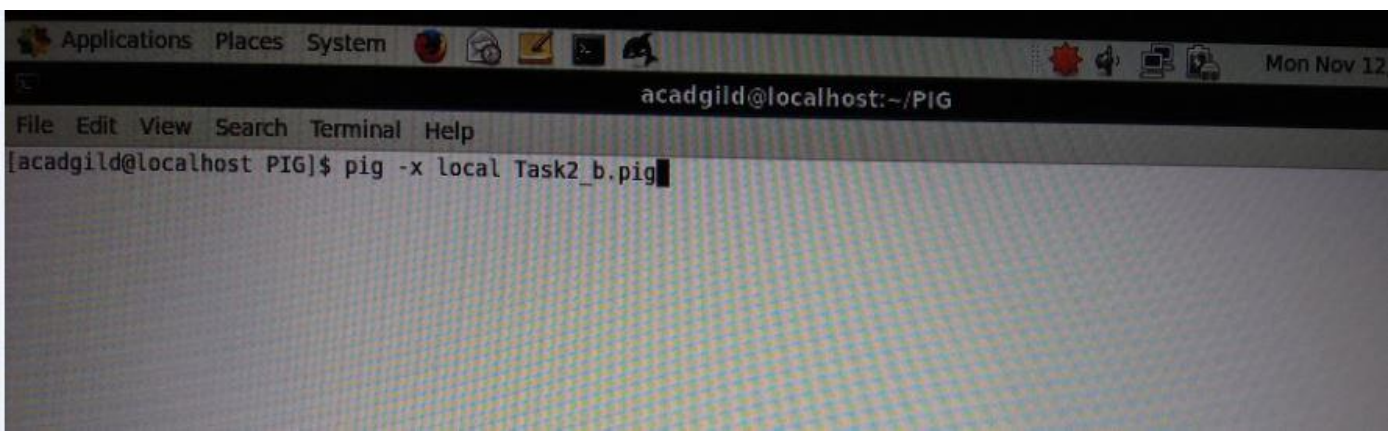
```
[acadgild@localhost PIG]$ pig -x local Task2_a.pig
```

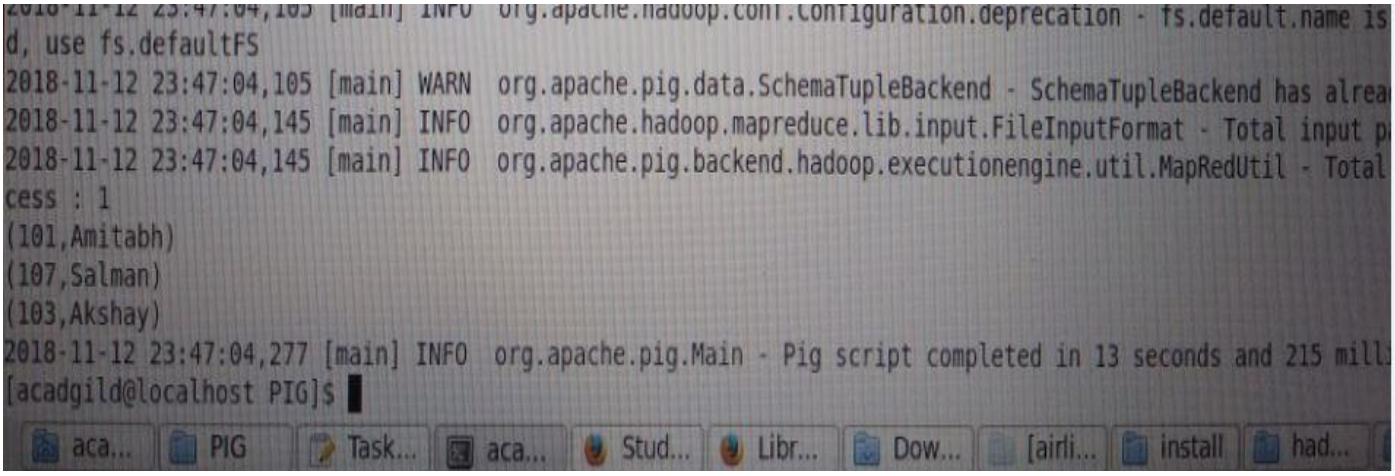
```
2018-11-12 23:44:22,735 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-11-12 23:44:22,750 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-11-12 23:44:22,755 [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 pro
cess : 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.**



5. **Output.**





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);

empdetailexpence = join empdetails by EmpID, empexpenses by EmpID;

groupByname = group empdetailexpence by Name;

maxexpense = FOREACH groupByname GENERATE  group, MAX(empdetailexpence.Expense);

maxempexpense = order maxexpense by group;

dump maxempexpense;
```

7. **Output.**

```
[acadgild@localhost PIG]$ pig -x local Task2_c.pig
```

```
d, use fs.defaultFS
2018-11-13 00:00:33,329 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already bee
2018-11-13 00:00:33,358 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths t
2018-11-13 00:00:33,359 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
cess : 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 millisecond
[acadgild@localhost PIG]$
```

8. **Script used for Task2_d.**

```
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.**

```
[acadgild@localhost PIG]$ pig -x local Task2_d.pig
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. Instea
d, 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 pro
cess : 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 | Task2_d.pig | Task2_e.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);

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

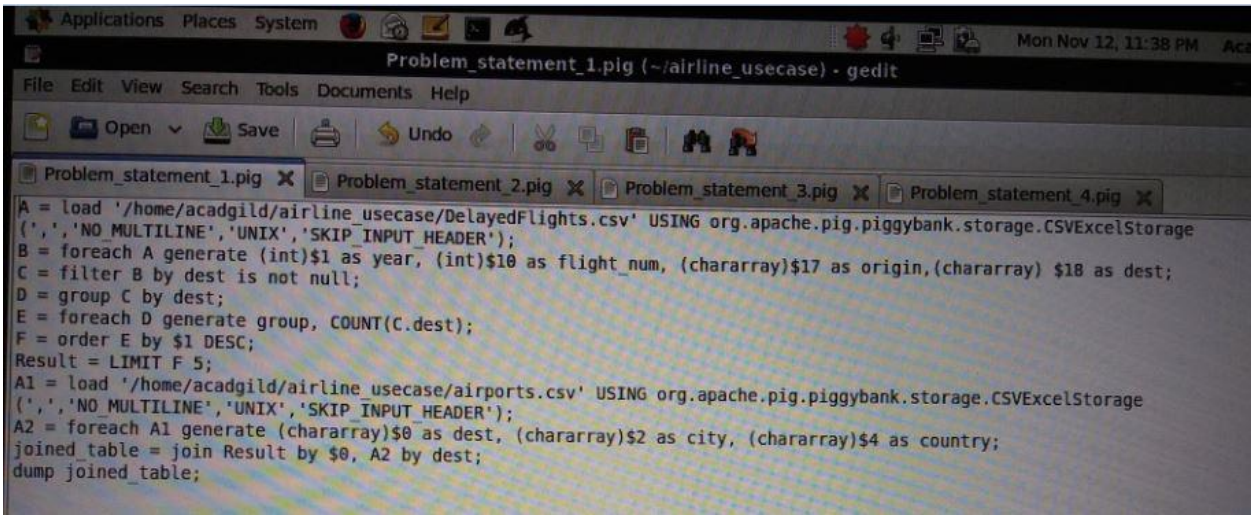
acadgild@localhost:~/PIG

File Edit View Search Terminal Help

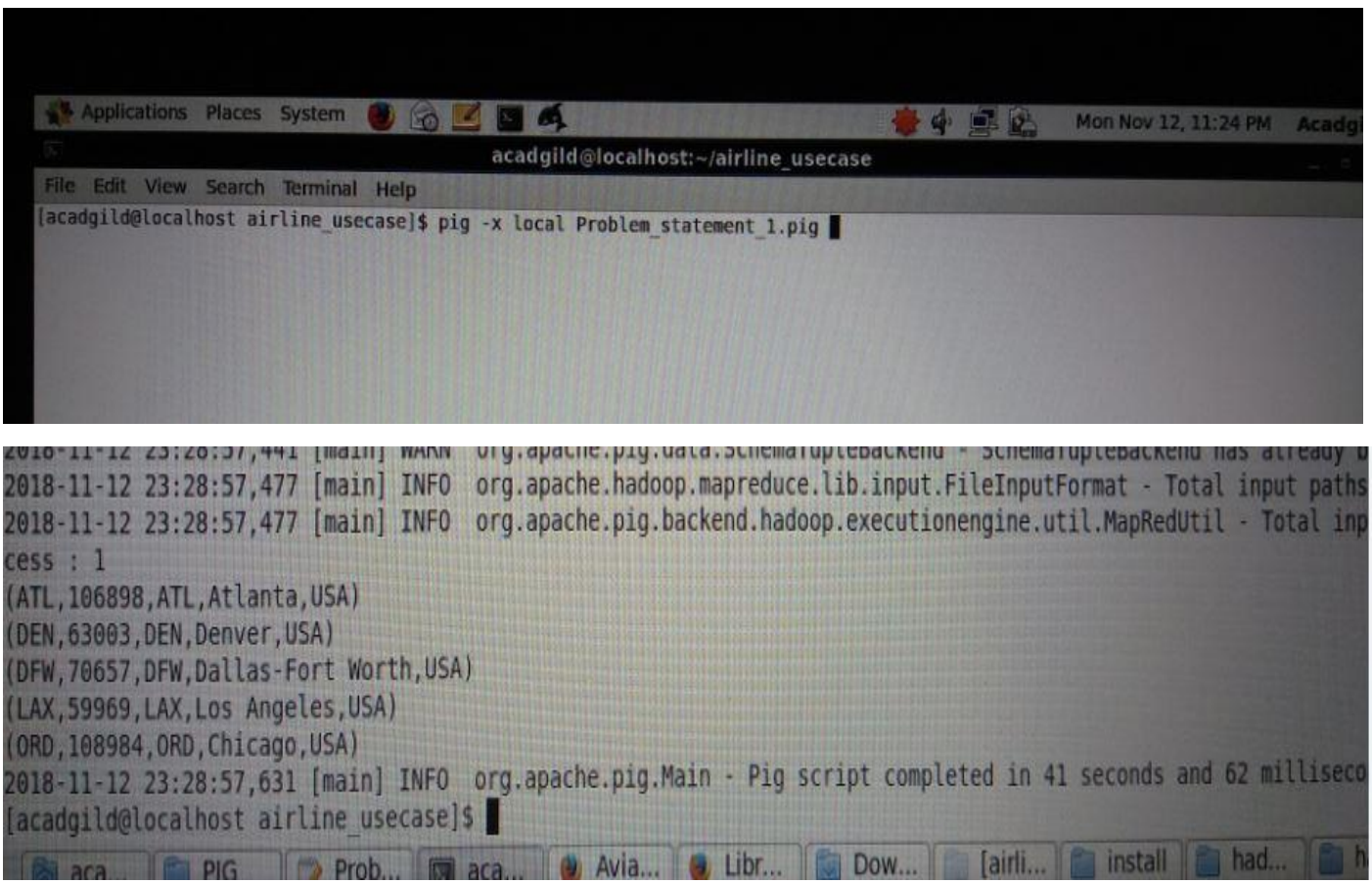
```
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost PIG]$ pig -x local Task2_e.pig
```

Task 3:

1. **Script used for Problem statement 1.**



2. **Output.**



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

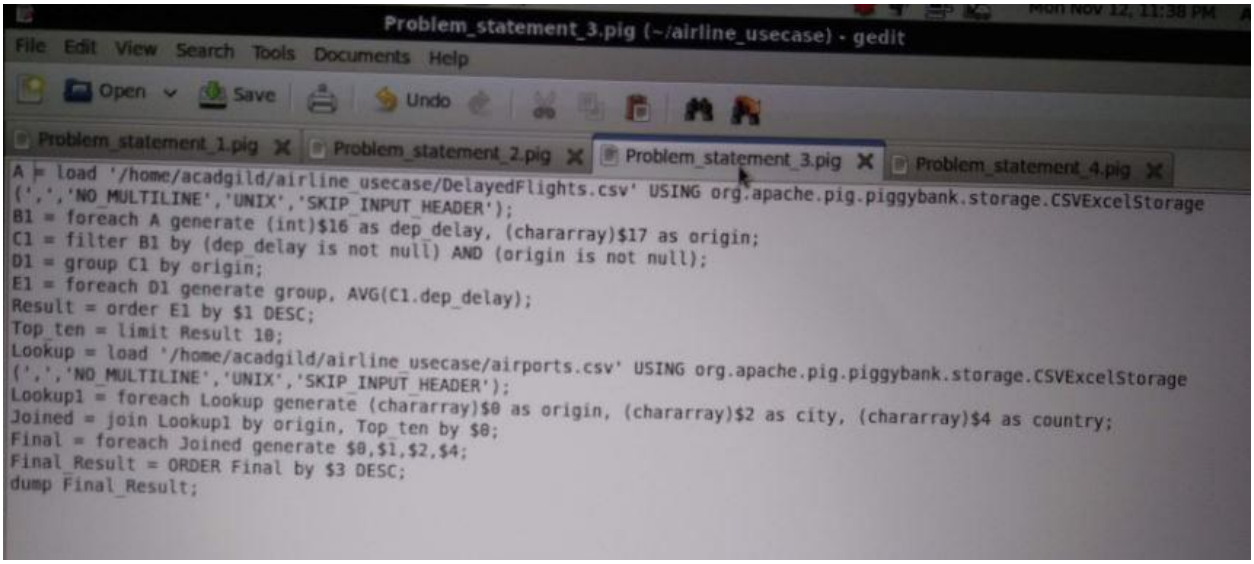
```
A = load '/home/acadgild/airline_usecase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
C = filter B by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

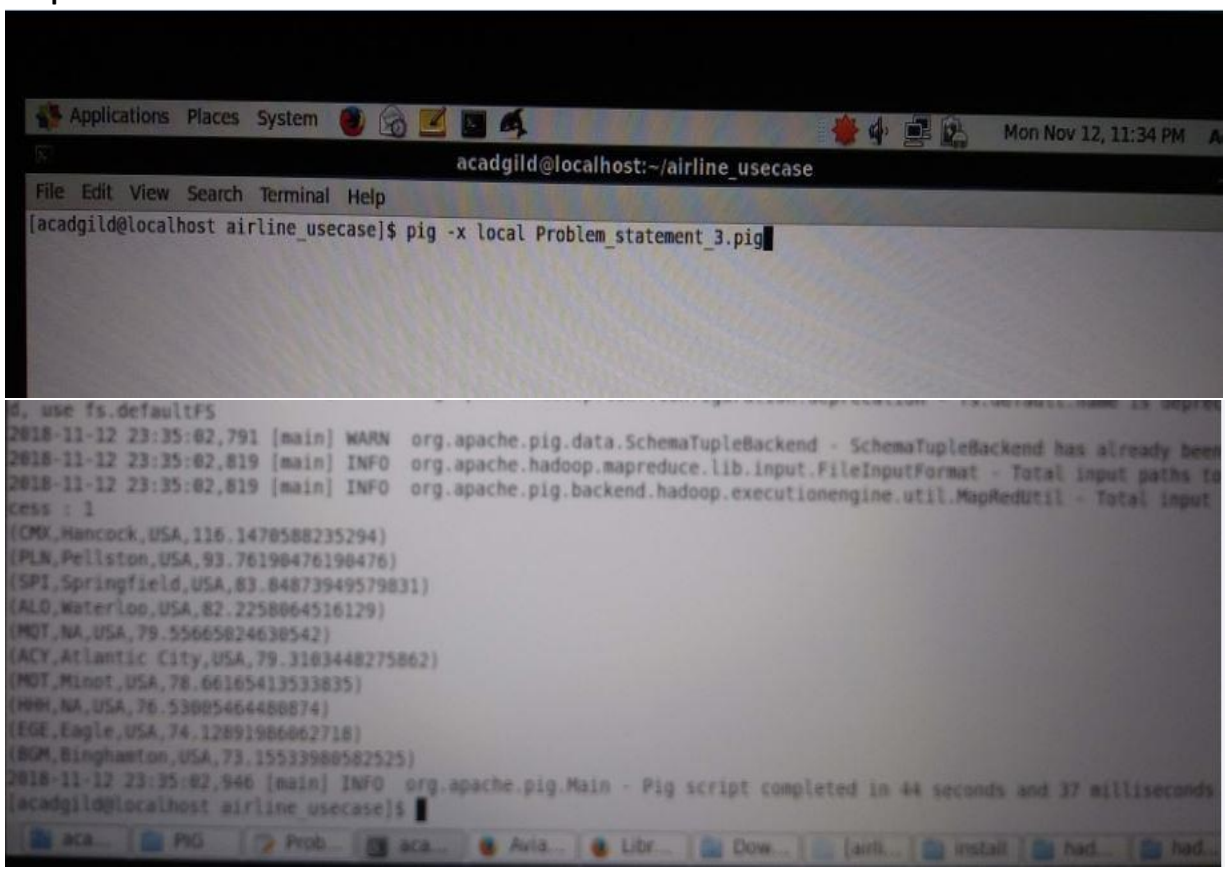
```
[acadgild@localhost airline_usecase]$ pig -x local Problem_statement_2.pig
```

```
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 alread
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 milli
[acadgild@localhost airline_usecase]$
```

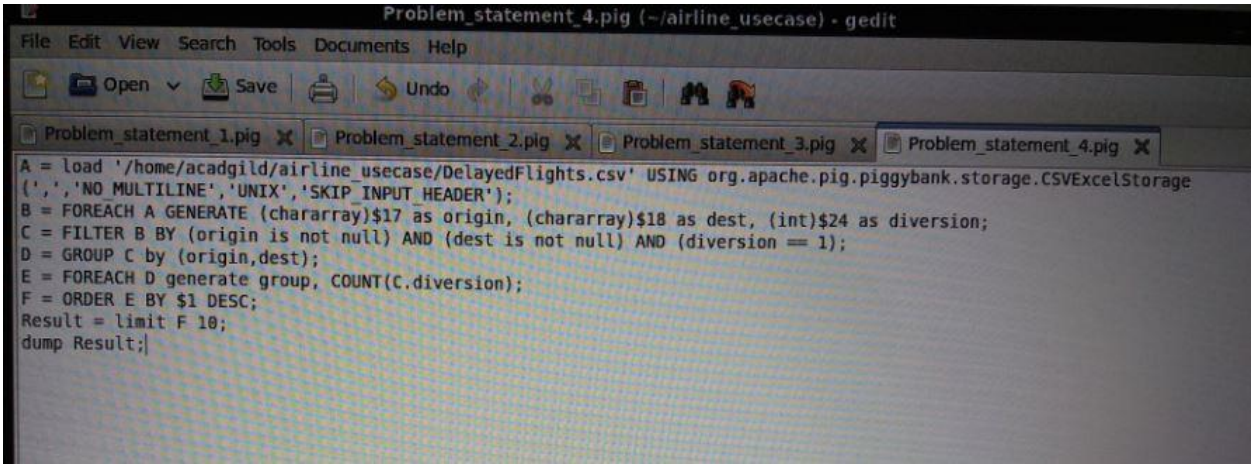
5. **Script used for Problem Statement 3.**



6. **Output.**



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



8. **Output.**