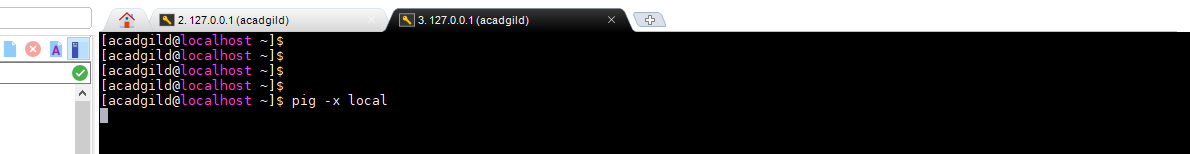
**ASSIGNMENT NUMBER 2.4**

**INTRODUCTION:** This is a acadgild dashboard session assignment 2.5 which demonstrate the basic insight of pig latin script. In this assignment we have two set of The U.S. Department of Transportation’s (DOT) Bureau of Transportation Statistics (BTS)

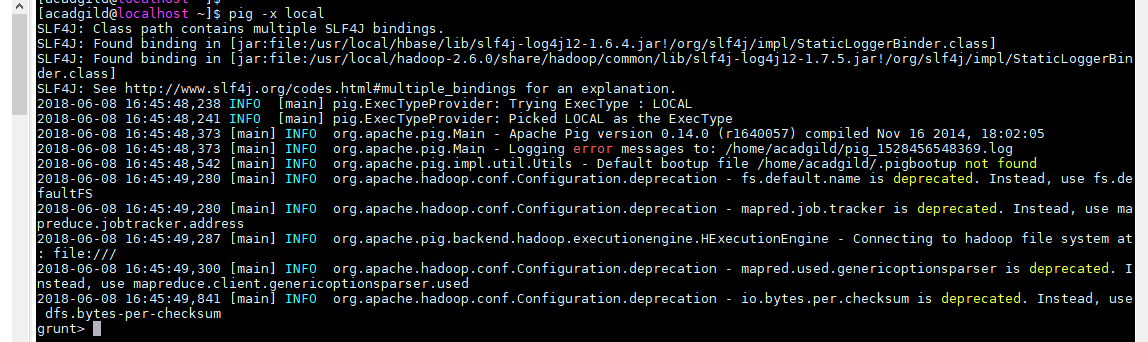
Data which is in csv format. To download this data we can follow link.

1. <https://drive.google.com/file/d/0B_Qjau8wv1KoWTVDUVFOdzlJNWM/view>
2. <https://drive.google.com/file/d/0B_Qjau8wv1KocDR3djk1Qm96Mmc/view>

Let give a name to this data for data 1 is Delayed\_flights.csv and for data 2 is Airports.csv. Now lets perform some pig latin operation to get following question output.Before going to question let me Register piggybank jar file from your local machine. Before start your grunt shell in local mode by using following screenshot command.



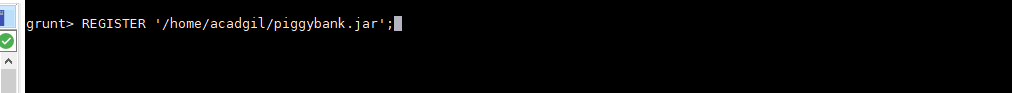
After executing this above command it will go to Grunt shell as follow.



Now Its time to do some pig latin operation on following question.

* Find out the top 5 most visited destinations.
* Which month has seen the most number of cancellations due to bad weather?
* Top ten origins with the highest AVG departure delay
* Which route (origin & destination) has seen the maximum diversion?

Now we will go to each and every problem .before doing that we should register our piggybank jar file from local machine by using REGESTER operation . If we donot register piggy jar file then we are unable to load our csv file because our data format is in csv so we have to register it first.



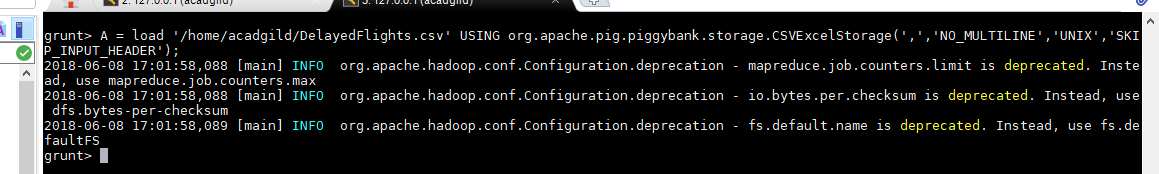
Now we can load our file in pig server.

**Problem No 1**

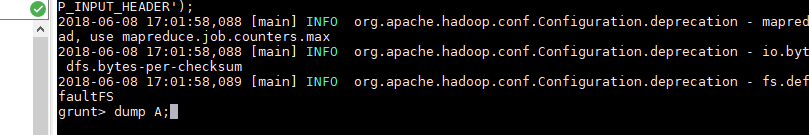
Find out the top 5 most visited destinations.

To find the solution to this problem by using pig latin we have to perform pig latin operation like order ,foreach generate, filter,group etc. Now for better understanding we will perform this operation step by step as follow.

**Step1:** Load the data using LOAD command as follow ..

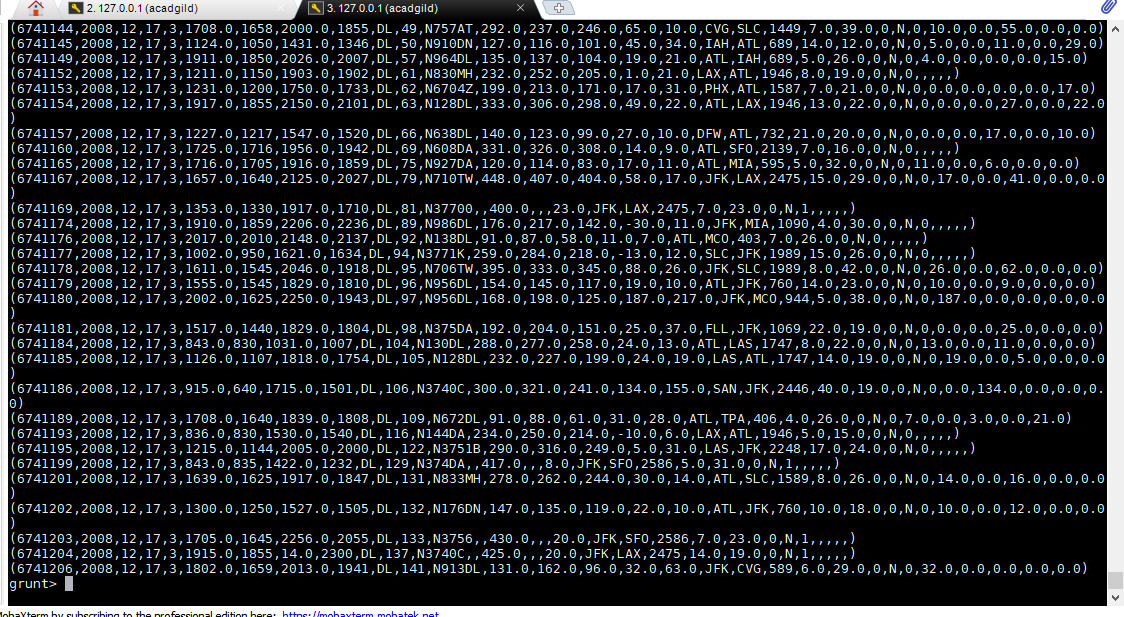


Load command is use a load a data from local machine or hadoop server .Here we have load a Delayed\_flights.csv file into relation name A. Now next step is to view a data. To view a data we use DUMP operation as follow by relation name .

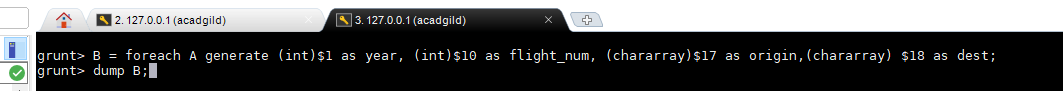


After applying dump operation we are able to see our data present in that file .

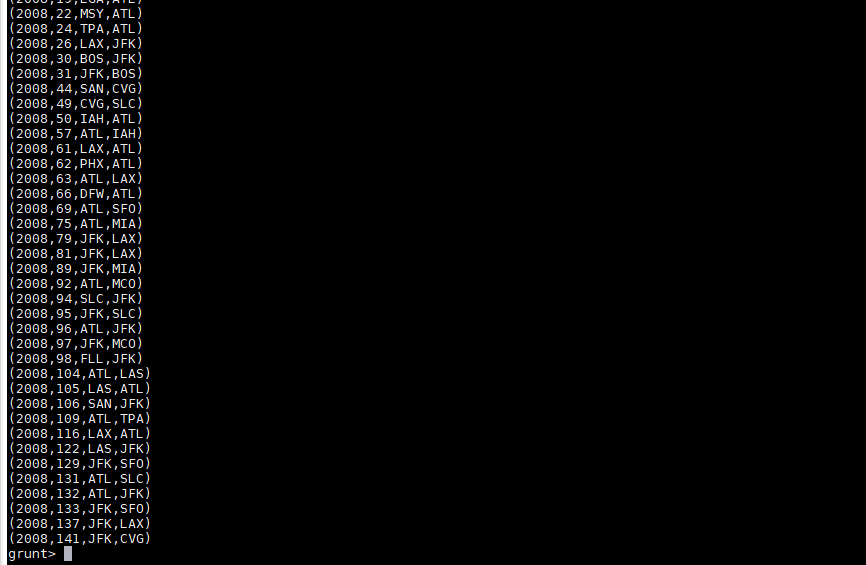
In our case data is as follow.



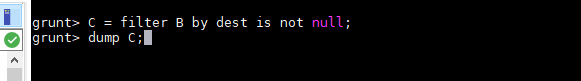
**Step 2** :Now it time generate require data from relation A . In this case we need only year,flight\_num ,orgin and dest.By using foreach generate operation we can find the relation which contain only related or valuable data.the query is as follow.



From the above relation A it will filter all unneccessay column and give you required column which is fit in our case. In our case we want only year , flight\_num ,orgin and dest data. So it will give us a following as output .

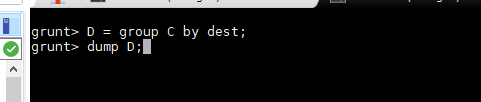


**Step 3** :Now it time time to filter generate data by using following command.

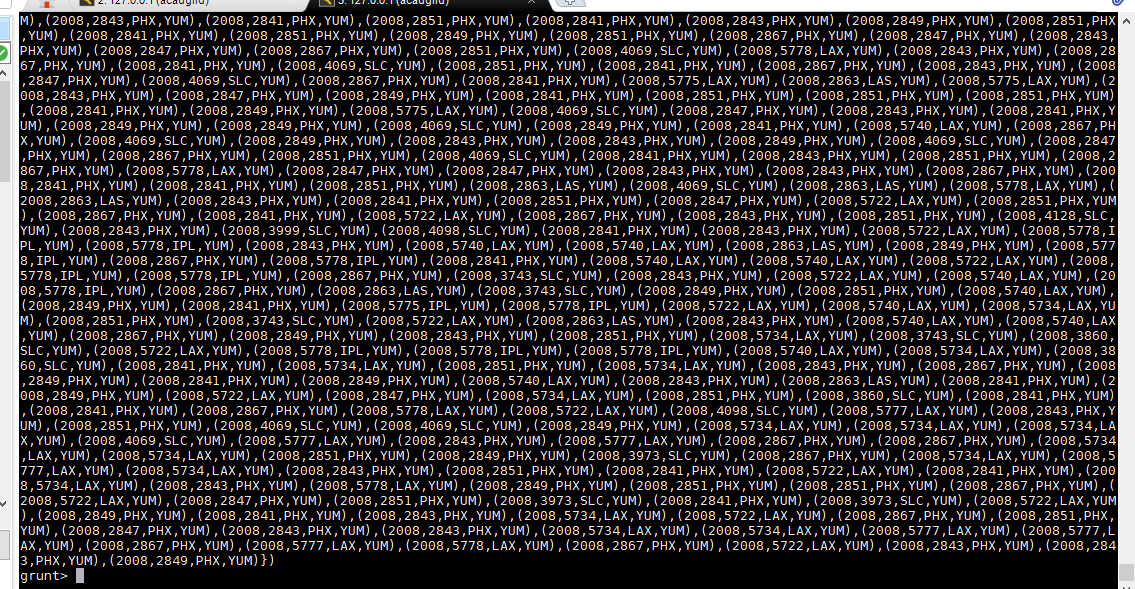


This command filter all row which does not contain any value that mean null value.

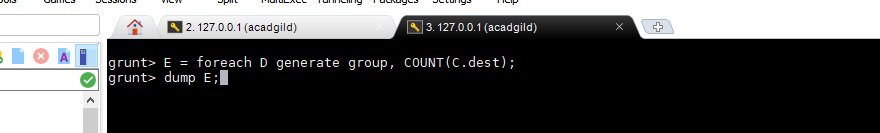
**Step 4:** In this step we have to group all the filter data which C in our case on the basis of dest because we have to find the top five destination place among these data.the query is as follow.



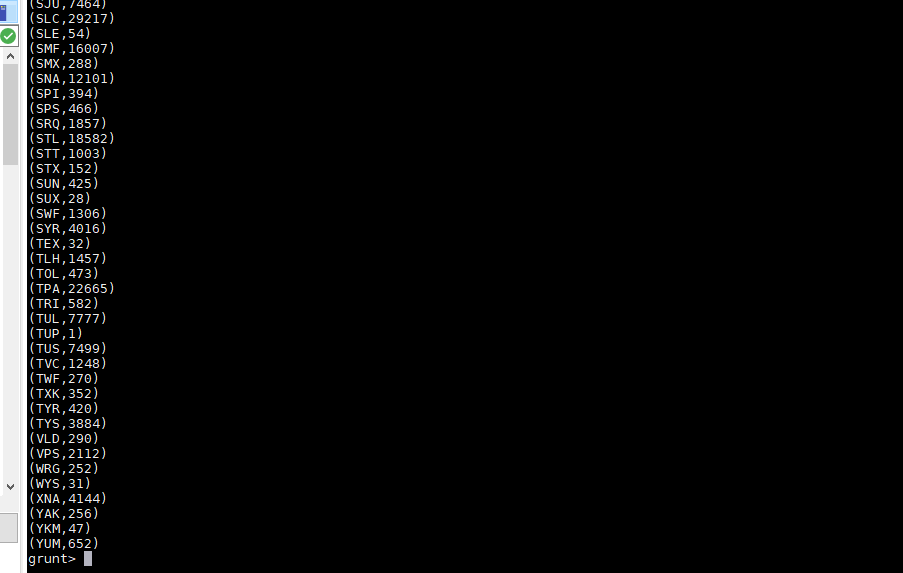
This data group filter data C on the basis of dest and give following as output if you dump a relation.



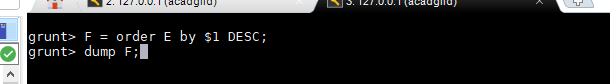
**Step 5** : now generate a count of dest by using COUNT function as follow.



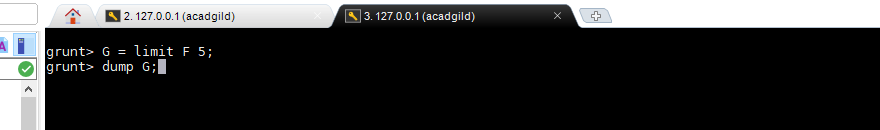
After executing this operation we will get following output.



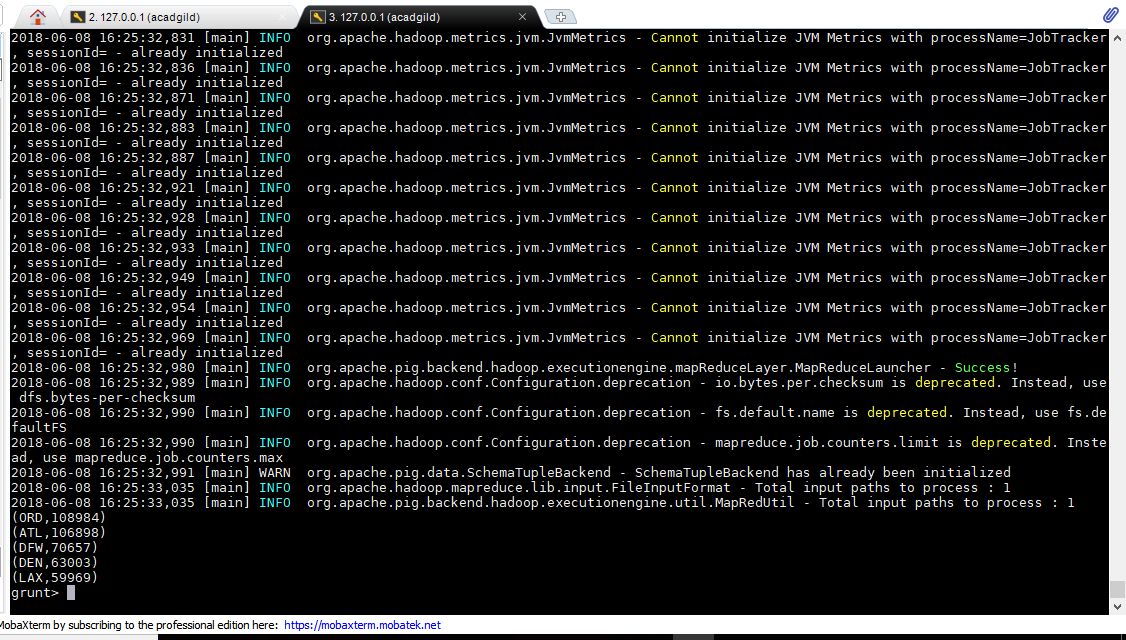
**Step 6** : order a above relation E on decreasing order by using order by operation as follow.



**Step 7:** Now our job is to find the top 5 most destination places so for this following command will work.



After running this operation we can get a required output as follow.

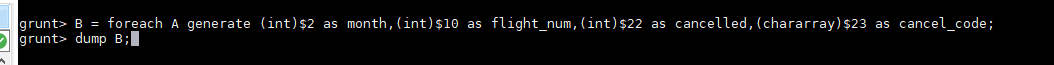


Now finally we get a output of solution 1.Now our time to jump into next problem 2.

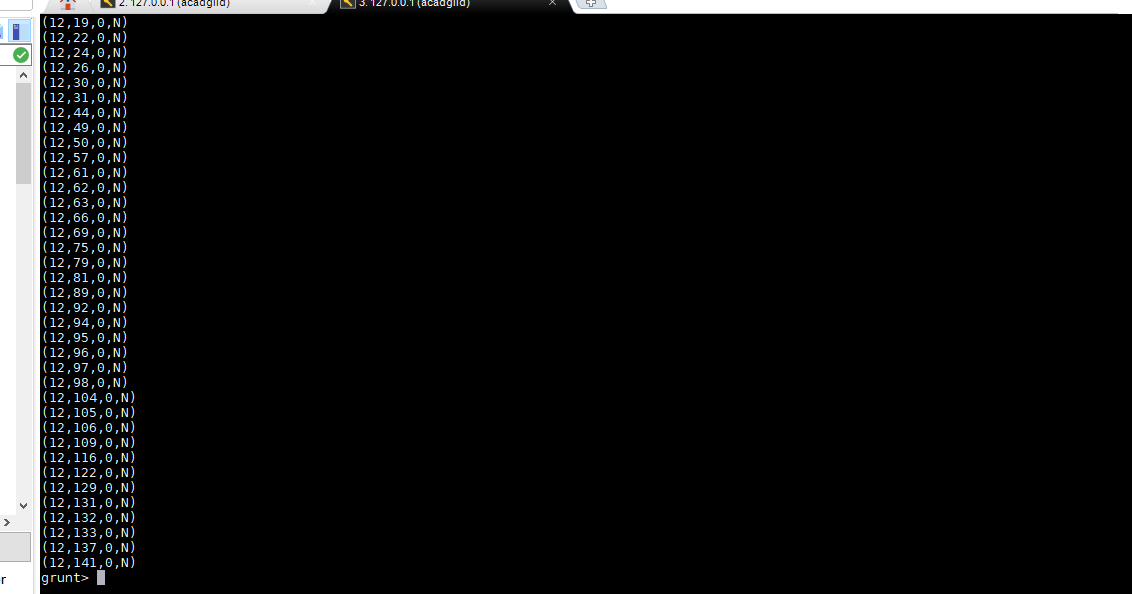
**Problem no 2**

**Step 1:** Follow same step as there in problem no 1 step 1.

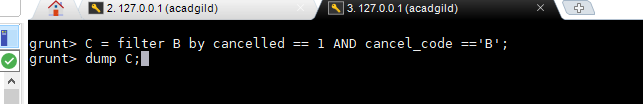
**Step 2**: In this step we generate a relation called B by taking only month ,flight\_num ,cancled,and cancle\_code as entites.operation is as follow.



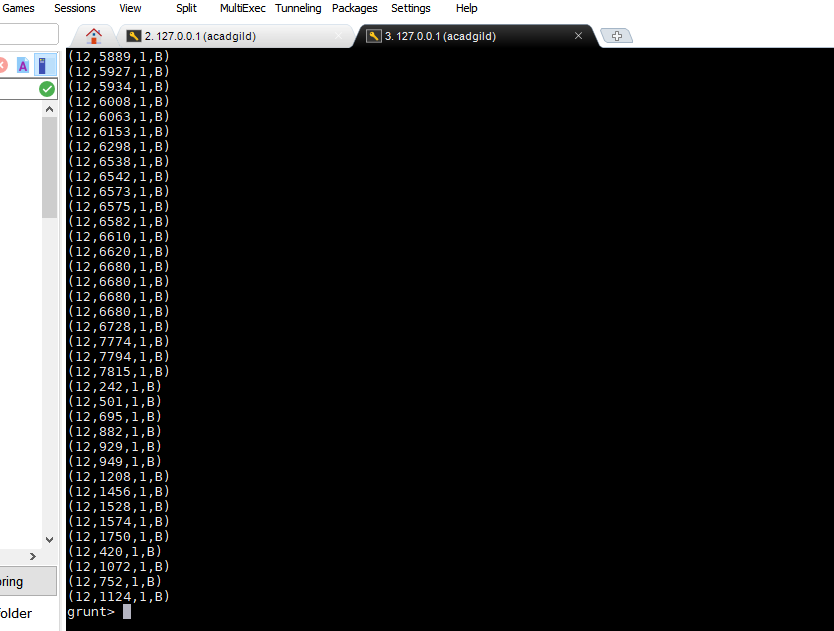
Output for this step 2 is as follow.



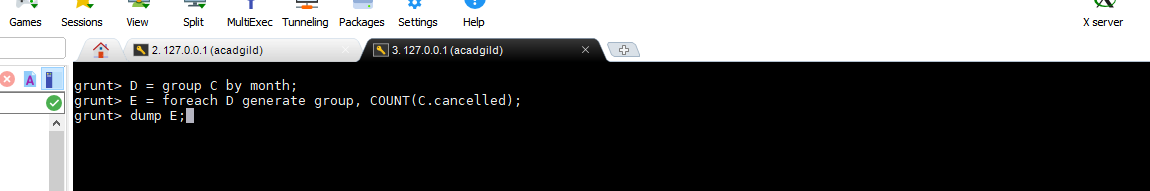
**Step 3**: In this step we have to perfrom filter operation to find the only cancelled code id weather or not and flight is cancelled or not by checking in data we came across following query.



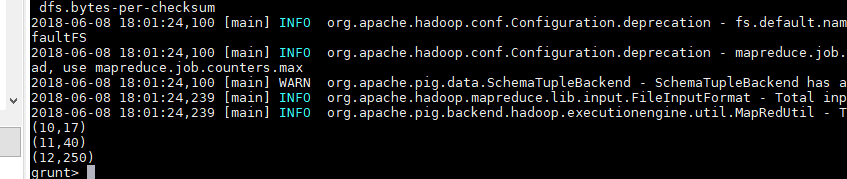
Output for this step is as follow.



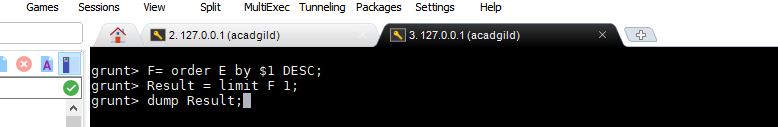
**Step 4:**Now group a data on the basis of month and generate a count of cancelled flight by using foreach generate operation. The query is as follow



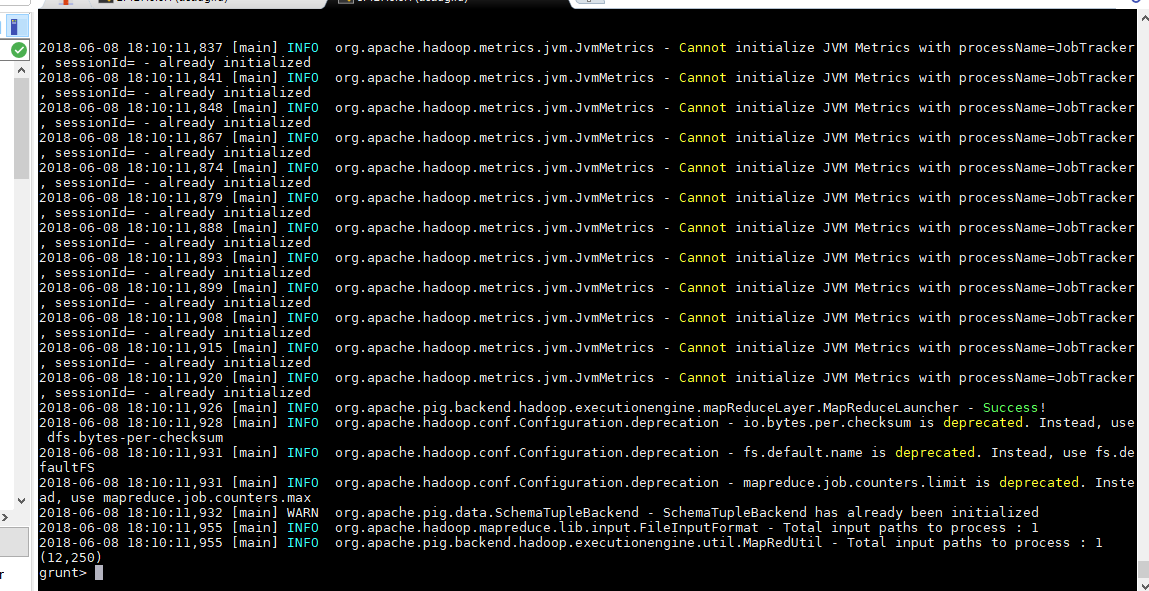
And the output of this format is as follow;



**Step 5**: This is final state where we can get result of all the input which is finding a month where number of flight cancelled due to bad weather is as follow



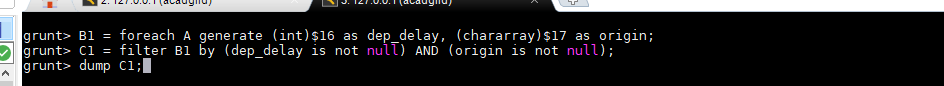
Now finally our problem no 2 is solved and our result is as follow.



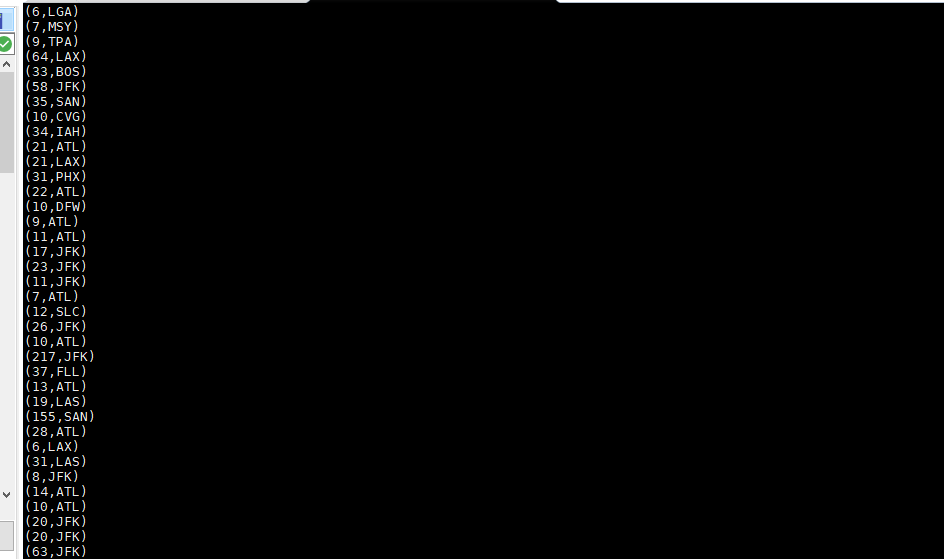
**Problem No 3**

**Step 1**: Follow the step 1 on problem no 1 to load the data into pig relation A.

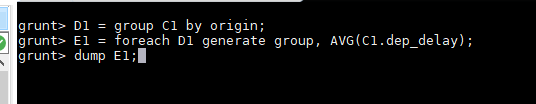
**Step 2**:Generate a data column by specifying foreach generate operation ,on which we should give a $16 as dep\_delayed and $17 as a origin and in this case only I have filter a both entities on the basis of not null . The command is as follow.



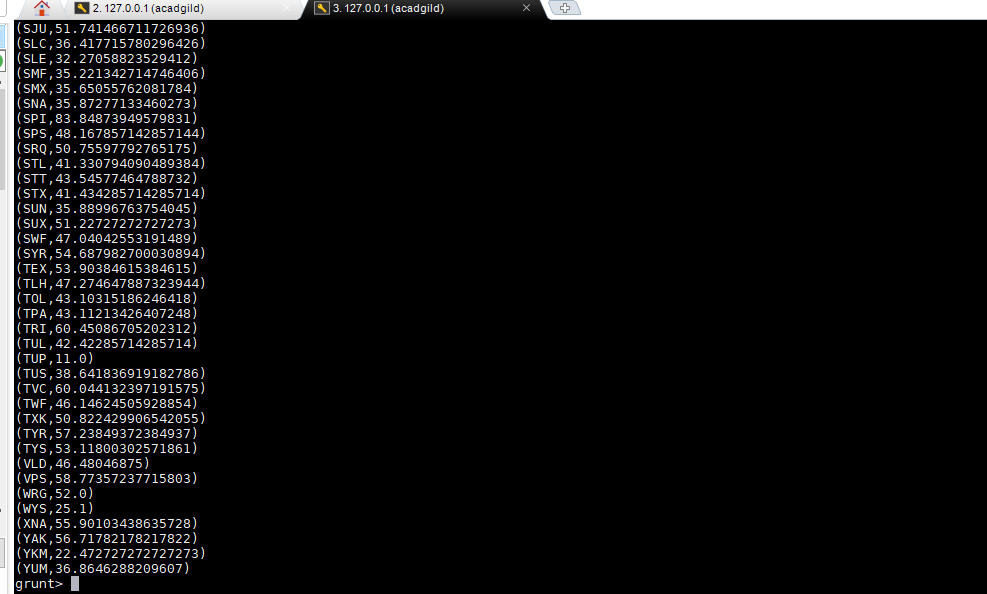
Output for this operation is as follow;



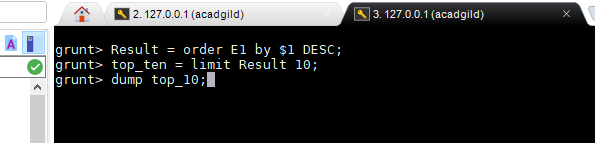
**Step 3**:Now we group a C1 relation by origin and generate a average of dep\_delayed by using following command as follow .



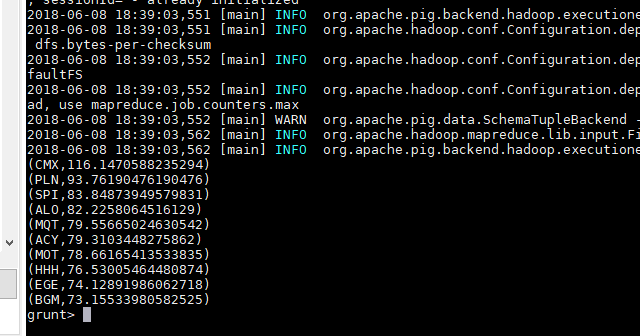
Output is as follow;



**Step 4**: Now this is a final step where we find the top 10 origin with higest avg dep\_delayed and limit this with 10 the command is as follow;



Output for this program is as follow;

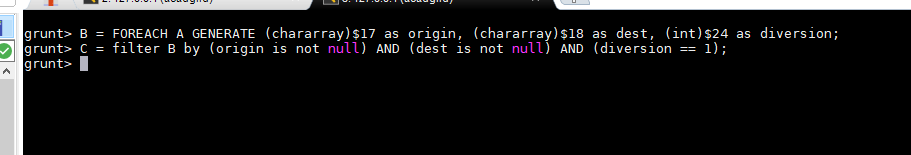


**Problem no 4**

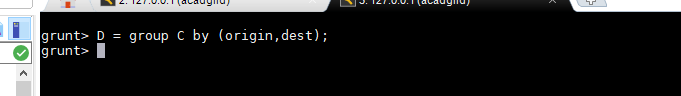
**Sept1:**Follow the sept 1 on problem no 1 to load the in relation A.

**Step 2**: Now this step, generate a column data which is required for our case. In our cases the required column data is destination, origin and derivert data.we give it name as relation B.

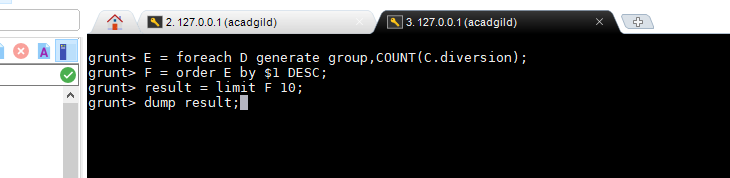
After generating a required data we have to filter a data to remove null value column from our generate data. To implement this we have to use filter operator as follow. Let it be a relation C.



**Step 3**: In this step we have group a data on basis of origin and dest and give a relation name as D as follow.



**Step 4:** To generate a maximum diversion we have again generate a count of diversion taken per unique origin and dest we give this relation name as E . now we have to order this generate a relation E on the basis of group .and produce a relation F . and finally produce a result by finding a only 10 cloumn among a order data as follow.



This will give a following output.

