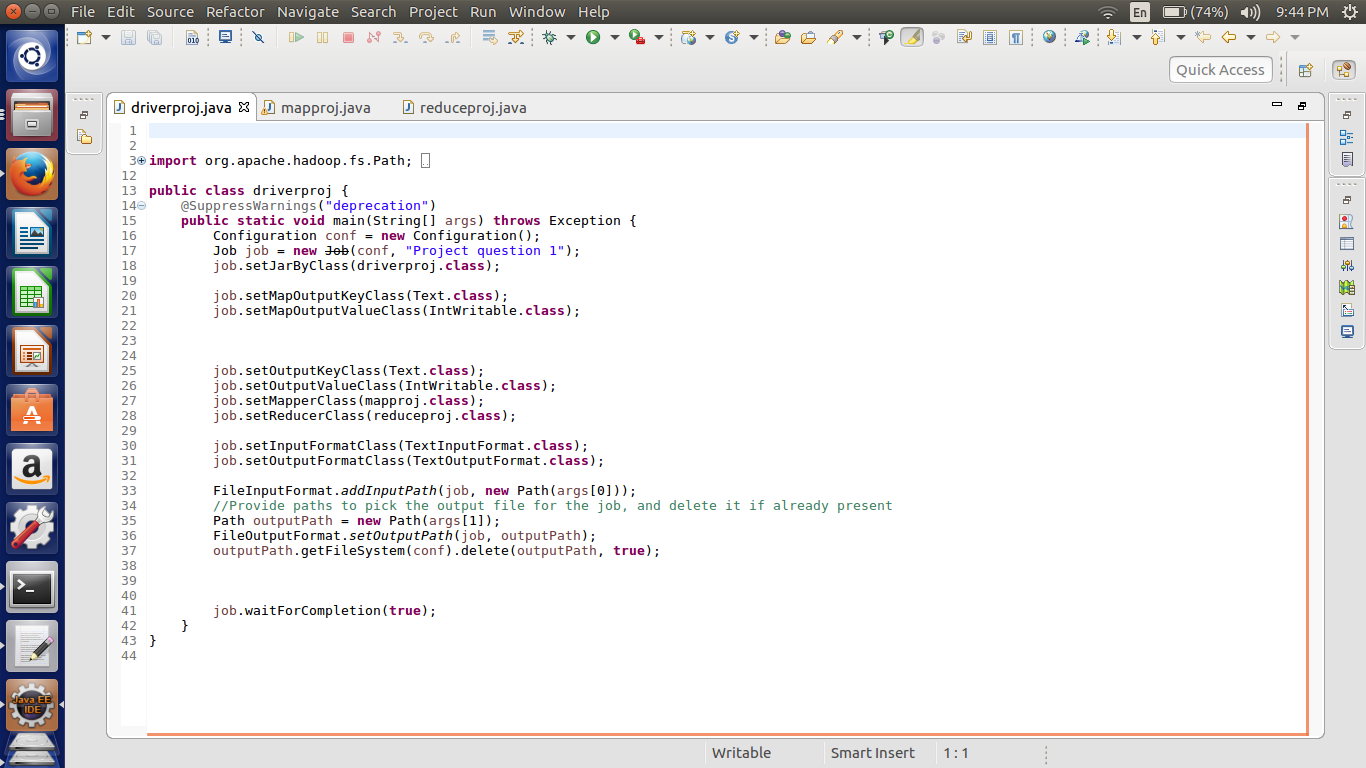
Project

1. Write a mapreduce and pig program to calculate the number of cases investigated under each FBI code

//Driver Class

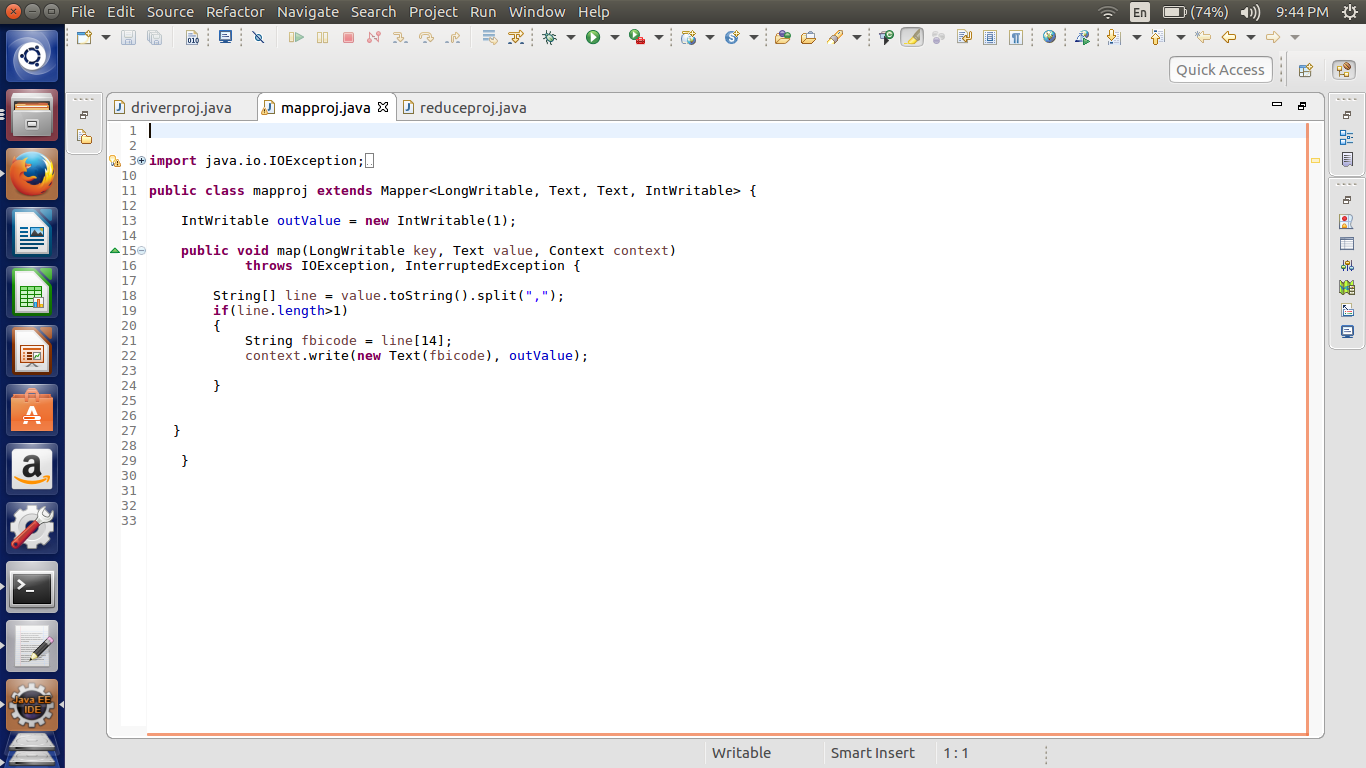


Here we are setting the job configurations and in the first question we can achieve the results by doing the map and reduce job hence setting the in the job configurations about the map class as well as the reducer class.

Here I have taken the input and output format as the “Text Input Format class”.

Third last line of the code is written for the deleting the output file if it already exists in the HDFS.

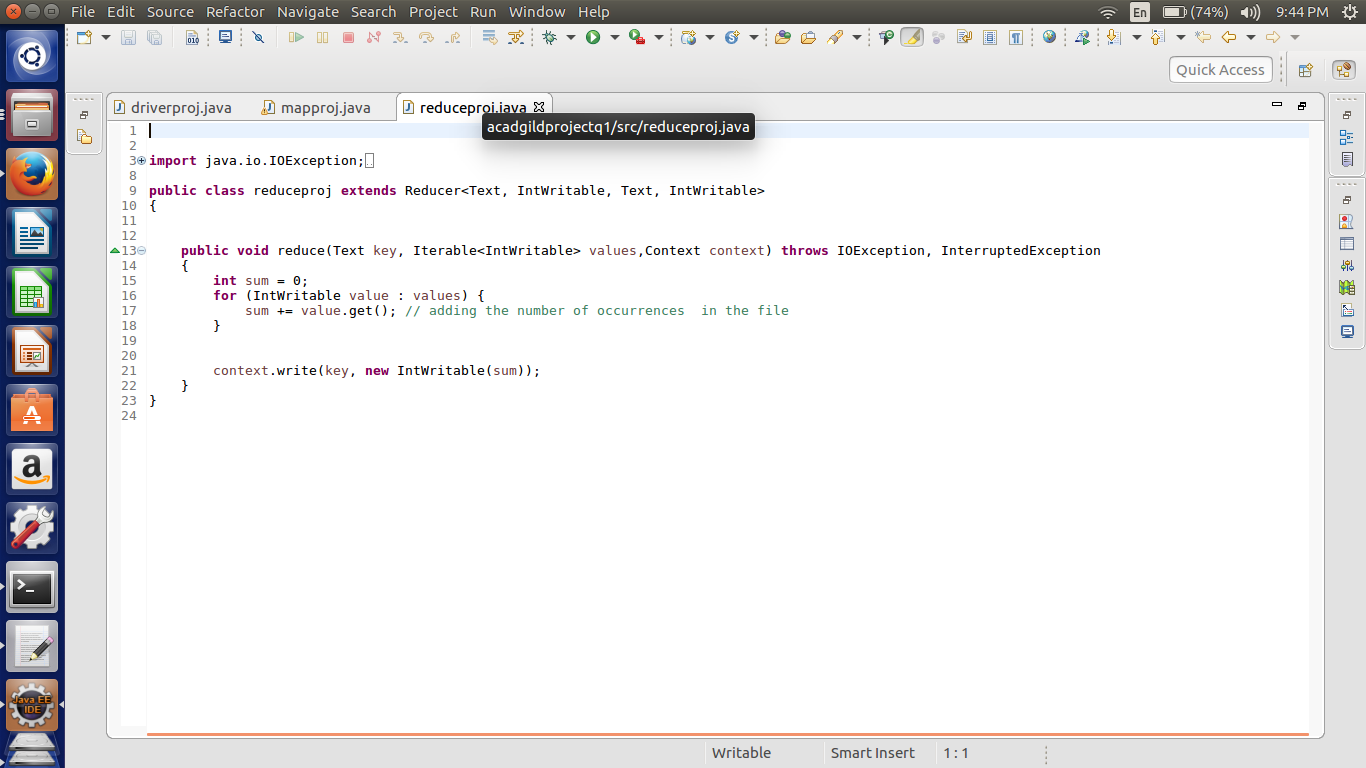
//Mapper class



In the mapper class first converted the input line as the string then split using the delimiter as “ , ”.

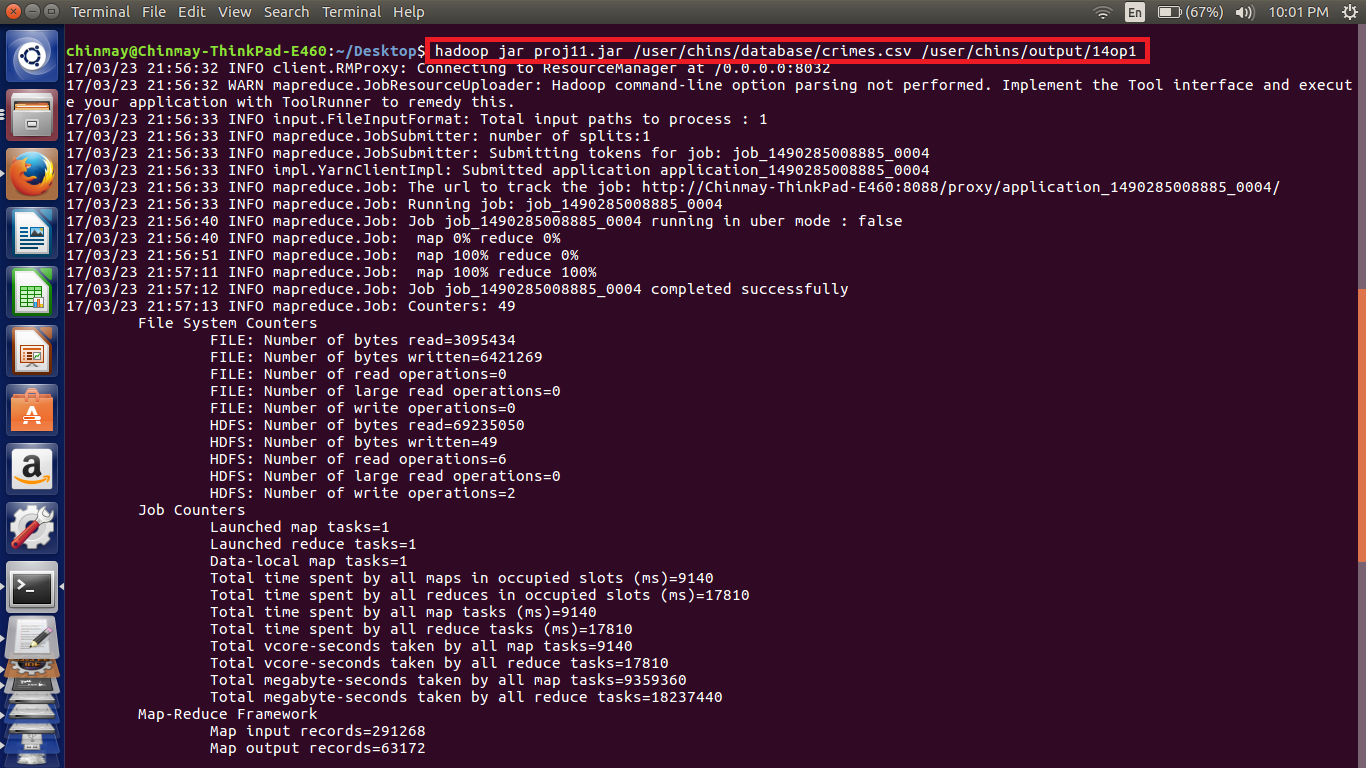
No consider the FBI code which is the 14th split sent it as the key and value as the intwritable format which is one.

//reducer class



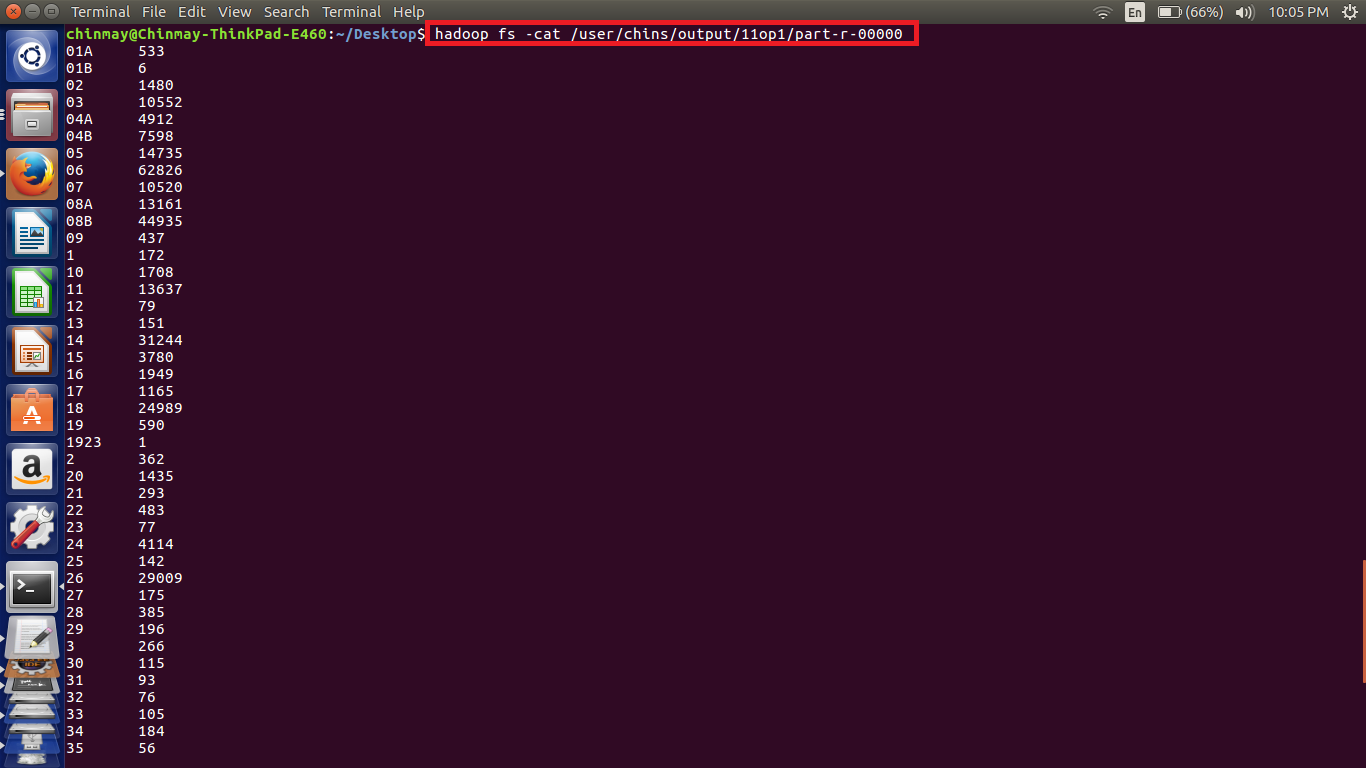
In the reducer the summing up the values which are related each key and passing the key and the sum as the value in the intwritable format.

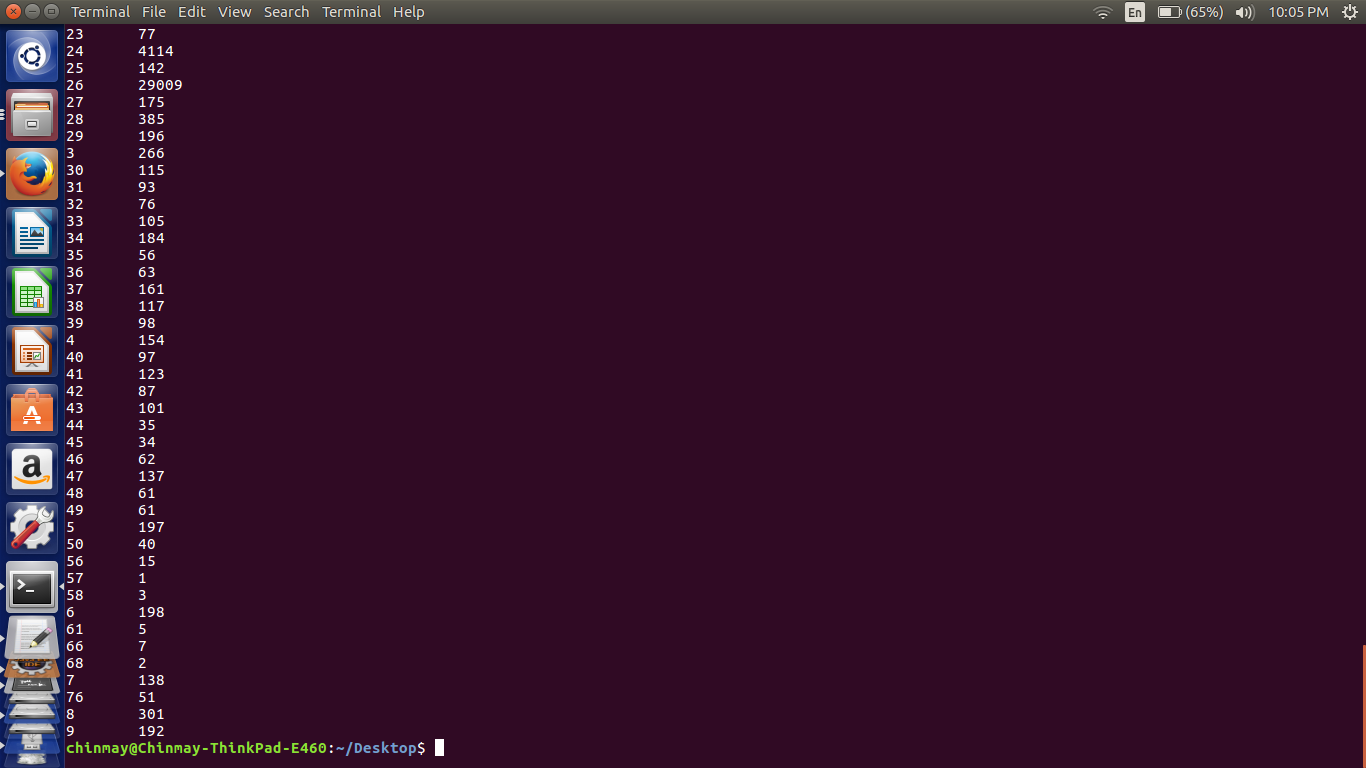
//Implementation



Here we can see the implementation the code using the highlighted command in which the input and output file paths are mentioned.

//Output





The outputs are in the format

FBI Code No of cases.

// Pig Query

a = load '/home/chinmay/Desktop/crimes.csv' USING PigStorage(',') as (id:int,caseno:chararray,date:chararray,block:chararray,IUCR:int,type:chararray,desc:chararray,locationdesc:chararray,arrest:chararray,dome:chararray,beat:int,dist:int,ward:int,Coarea:int,code:chararray,xco:int,yco:int,year:int,Upd:chararray,latitude:int,longitude:int,location:chararray);

b = group a by code;

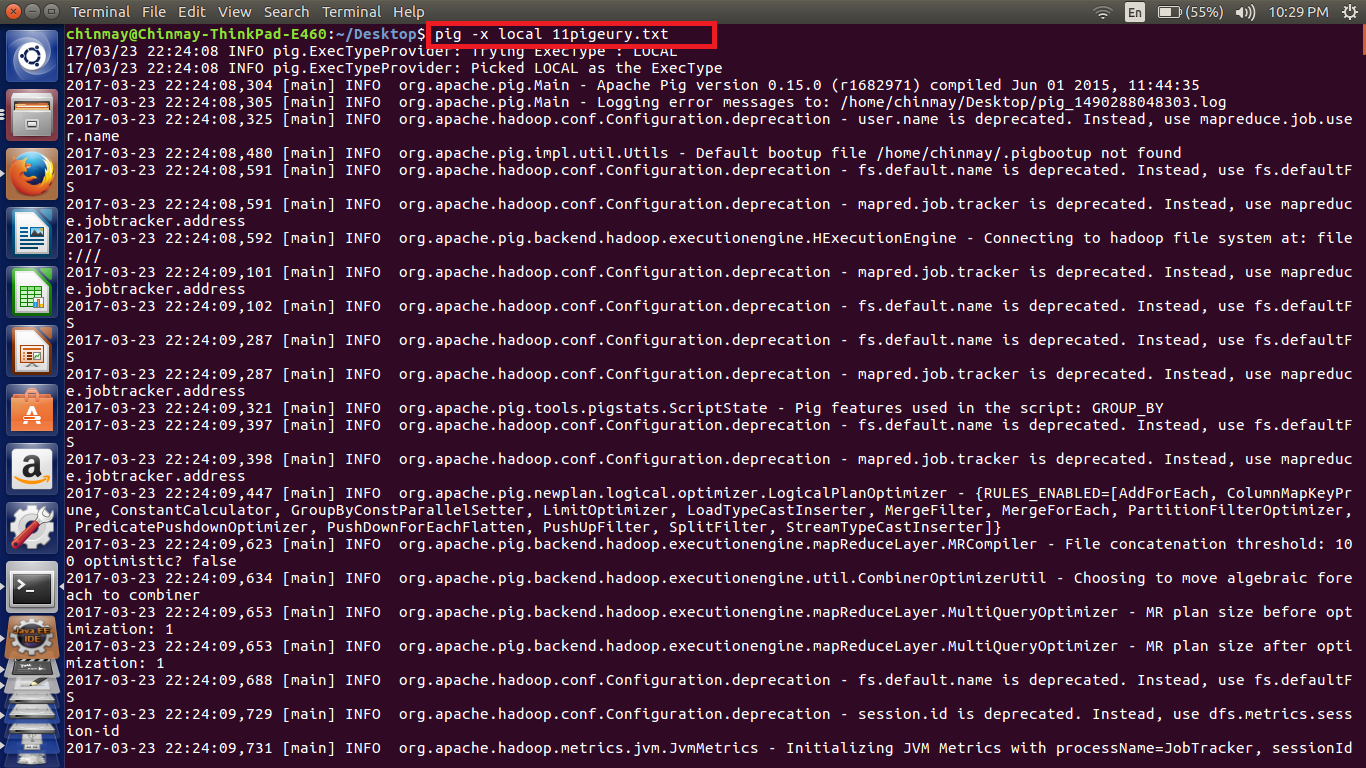
c = FOREACH b GENERATE group,COUNT(a.code);

dump c;

Here we have loaded the value with respective data type and then grouped by the FBIcode.

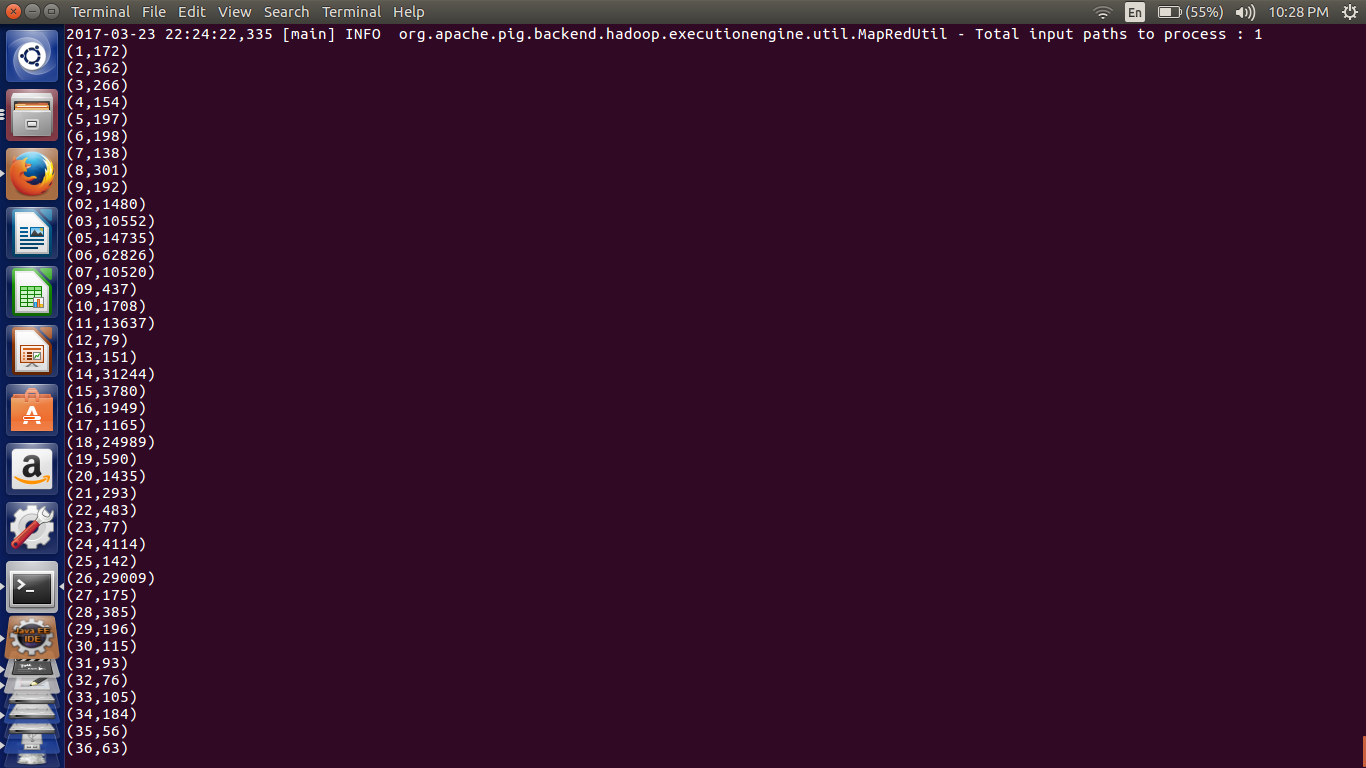
Then generated the count of the FBI codes in each group and found out the no of cases in each code.

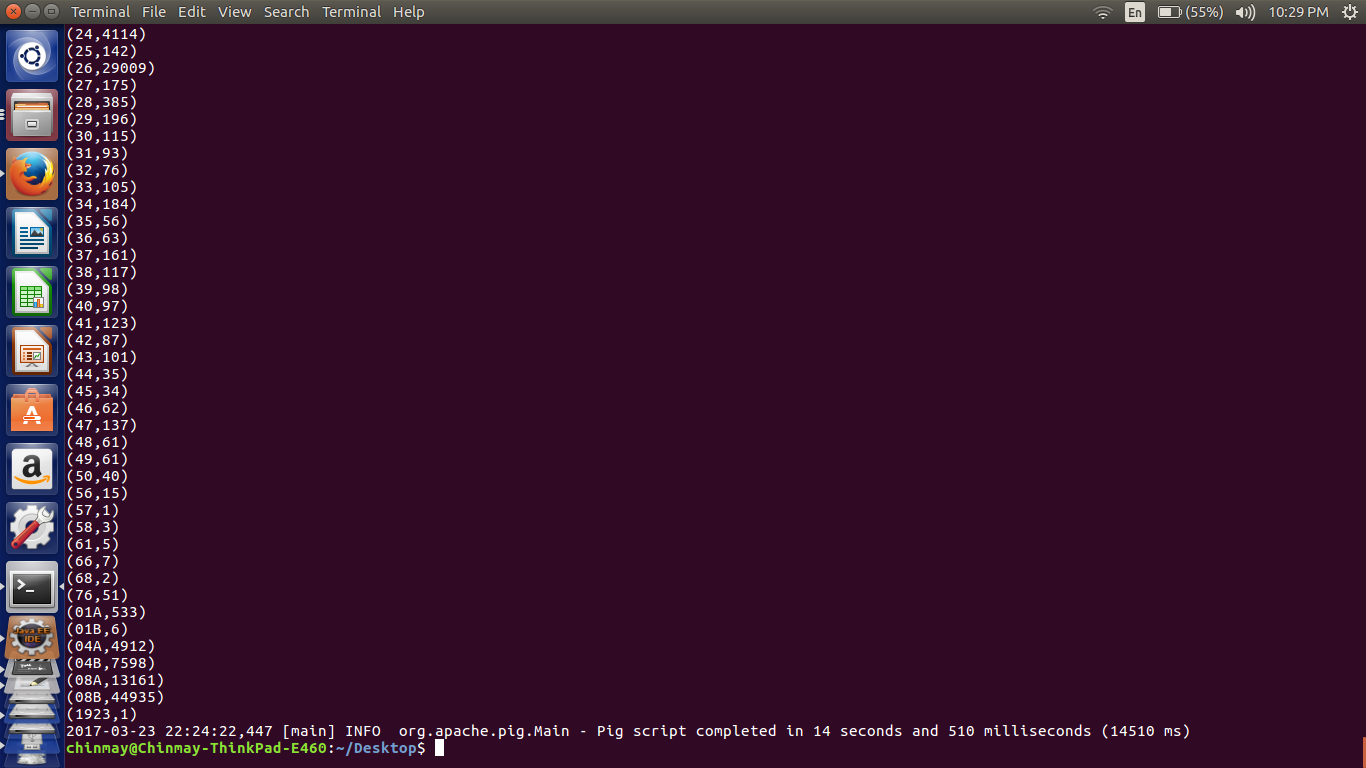
//Implementation



In the pig script first loaded the data in the comma delimited format then grouped by the FBI code and counted the number of codes in each group which gives us the number of cases under each code.

//Output





Here we can see that the output using the pig as well as the mapreduce program is same only the order of the output has changed

***Note-- Also attached the pig file used for the implementation.***