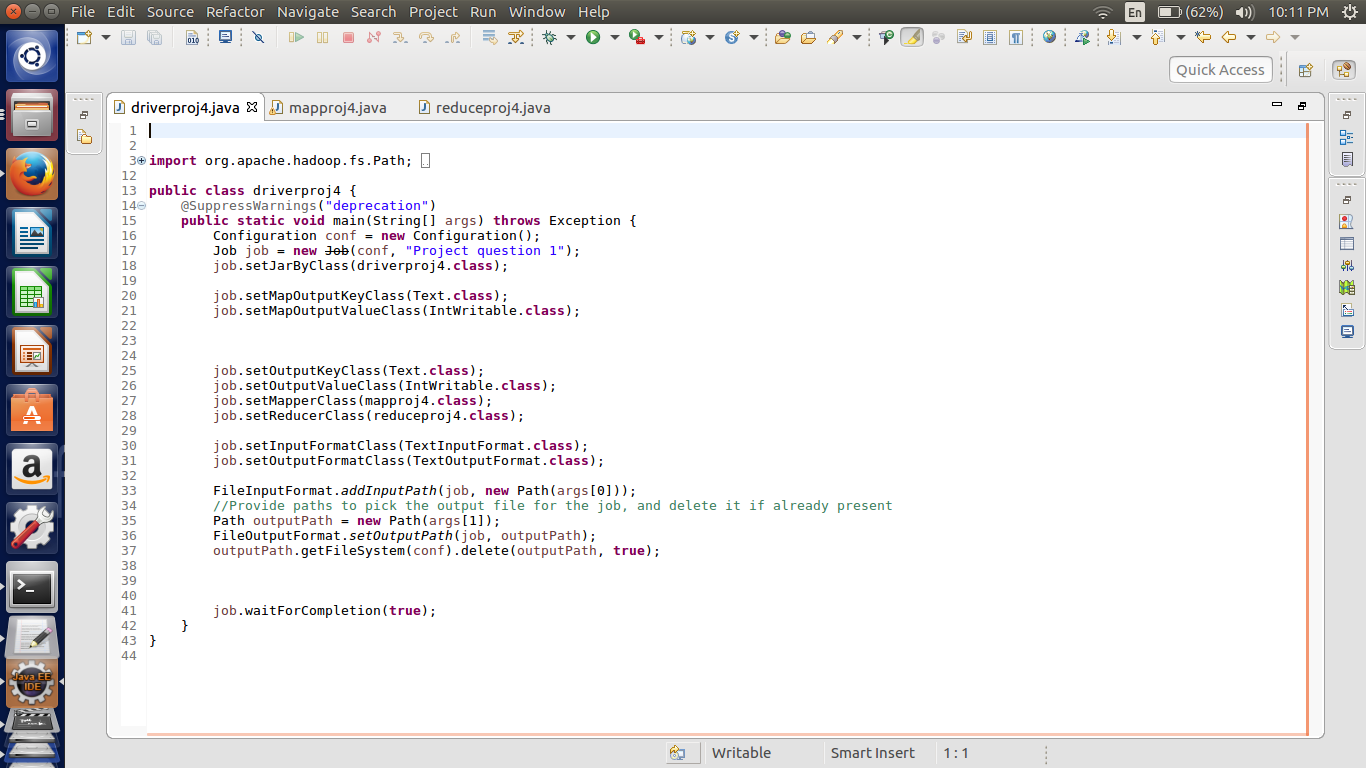
4. Write a mapreduce and pig program to calculate the number of arrests done between October 2014 and October 2015.

// Driver

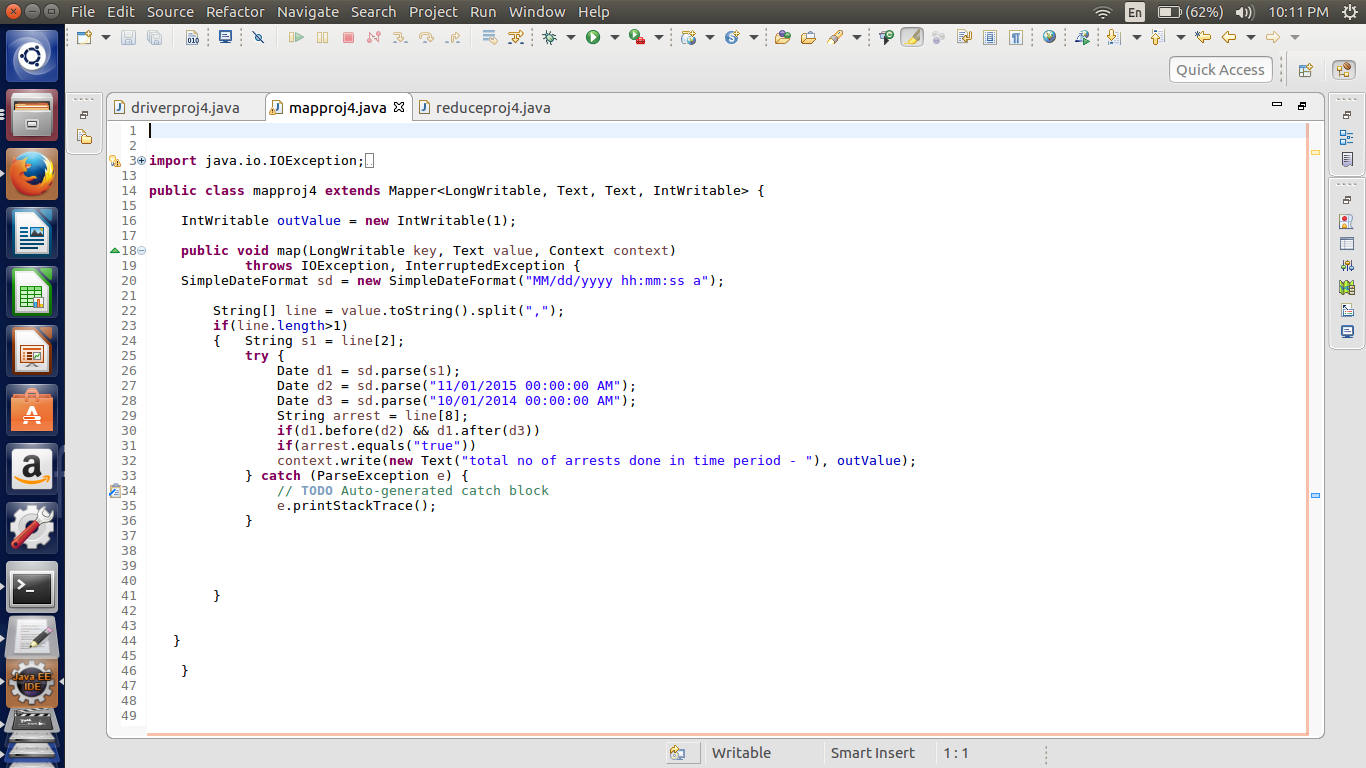


Here we are setting the job configurations and in the first question we can achieve the results by doing the map and reduce job hence setting the in the job configurations about the map class as well as the reducer class.

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 this mapper we have using simpledateformat. First selected the date and parsed the date to check the format and convert the given string into the date format.

The Date considered is

Strating – 1st NOV 2014

Ending – 1st Oct 2015

As the it is said between the October 14 and October 15. Hence the October moth is considered .

Interpretation of the statement depends on the individual. And if October moth is to be included just change the date in the program.

The Format of the date is Month/day/year.

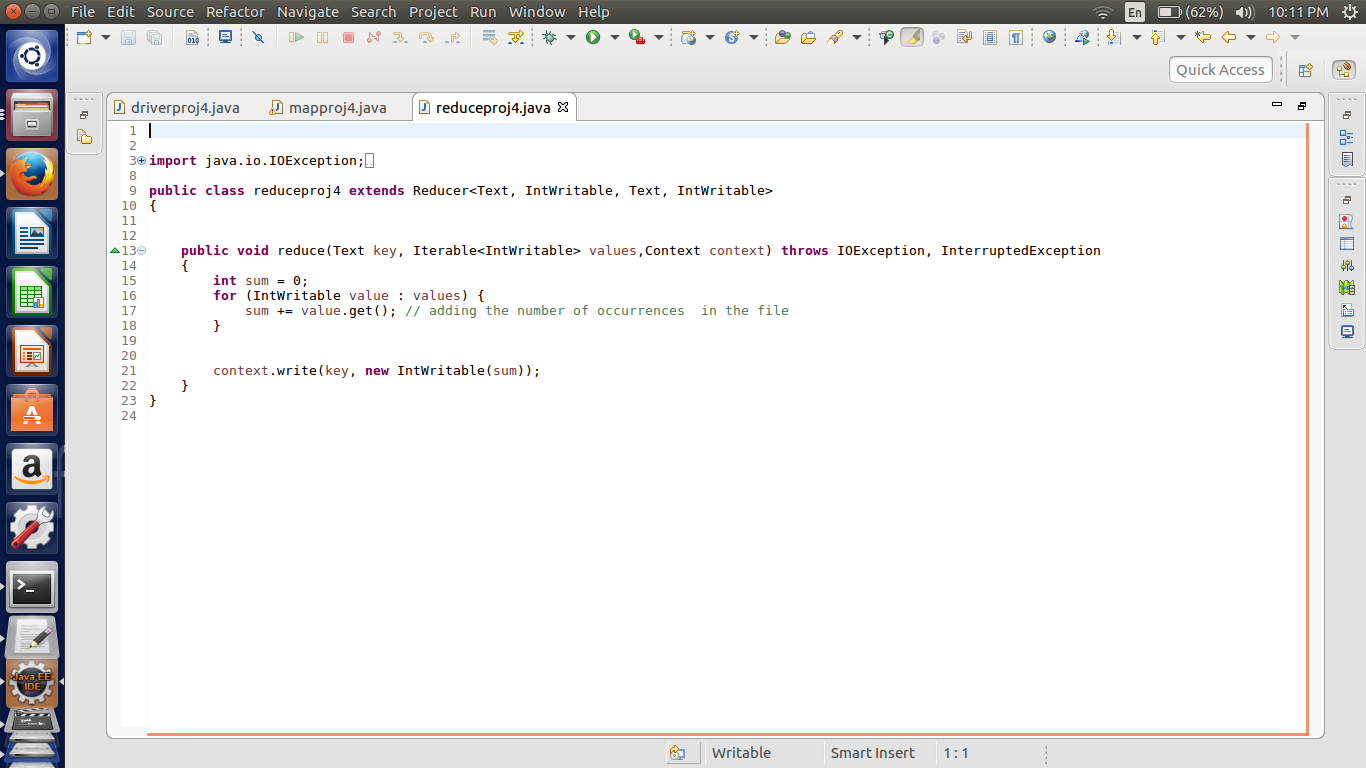
By using the condition of before and after functions date. We can check if the given date is between the specified date.

If the date is between the mentioned date then we are passing simple Text as key

“total no of arrests done in the time period”

The value passed is intwritable 1.

//Reducer

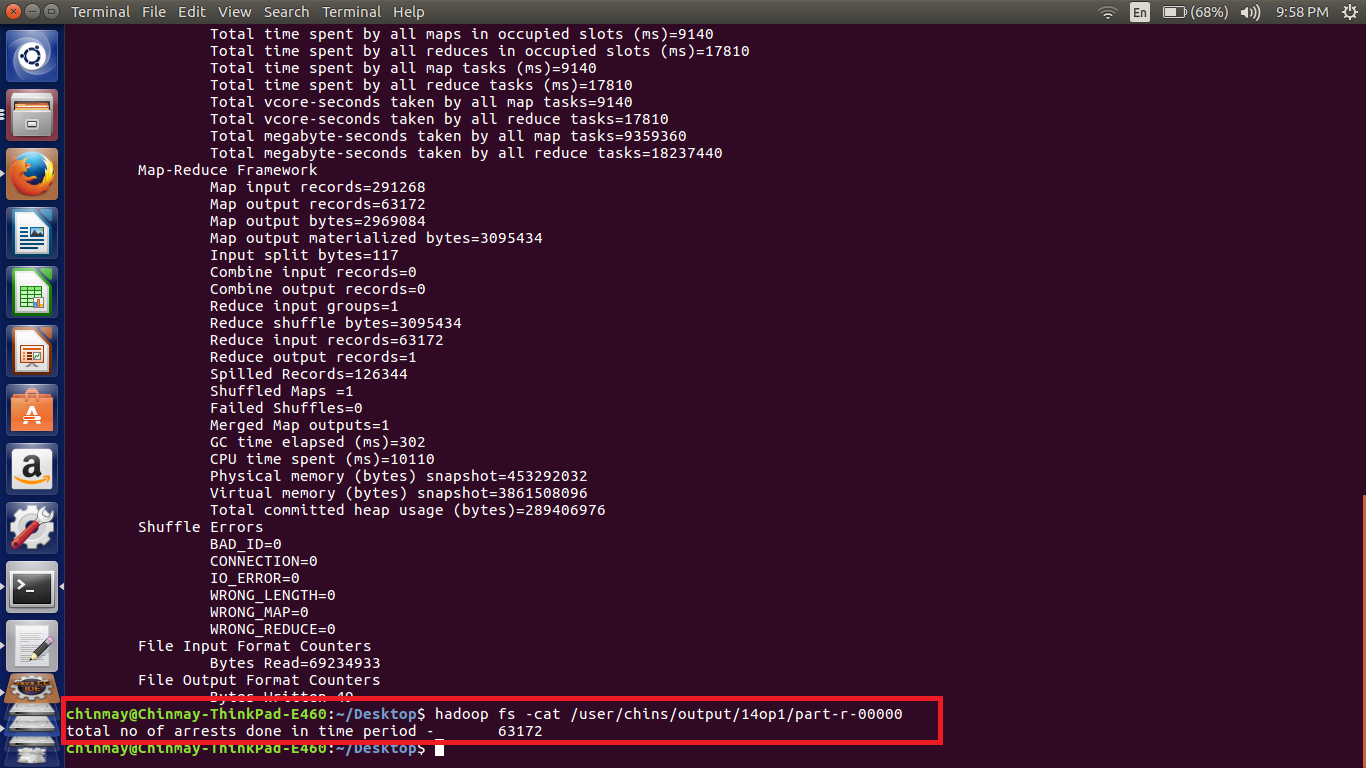


In the reducer the summing up the values which are related each key and passing the key and the sum as the value in the intwritable format.

In this case there is only one key.

At the reducer output we get the value of the no of arrests done in that period

//Output



Hence there are 63172 arrests done in that period.

Here important thing is that time considered 00:00:00 AM hence the answer may vary according to the time also.

Pig Implementation

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 = FILTER a by arrest=='true';

c = FILTER b by ToDate(date,'MM/dd/yyyy hh:mm:ss a')<ToDate('11/01/2015','MM/dd/yyyy') AND ToDate(date,'MM/dd/yyyy hh:mm:ss a')>ToDate('10/01/2014','MM/dd/yyyy');

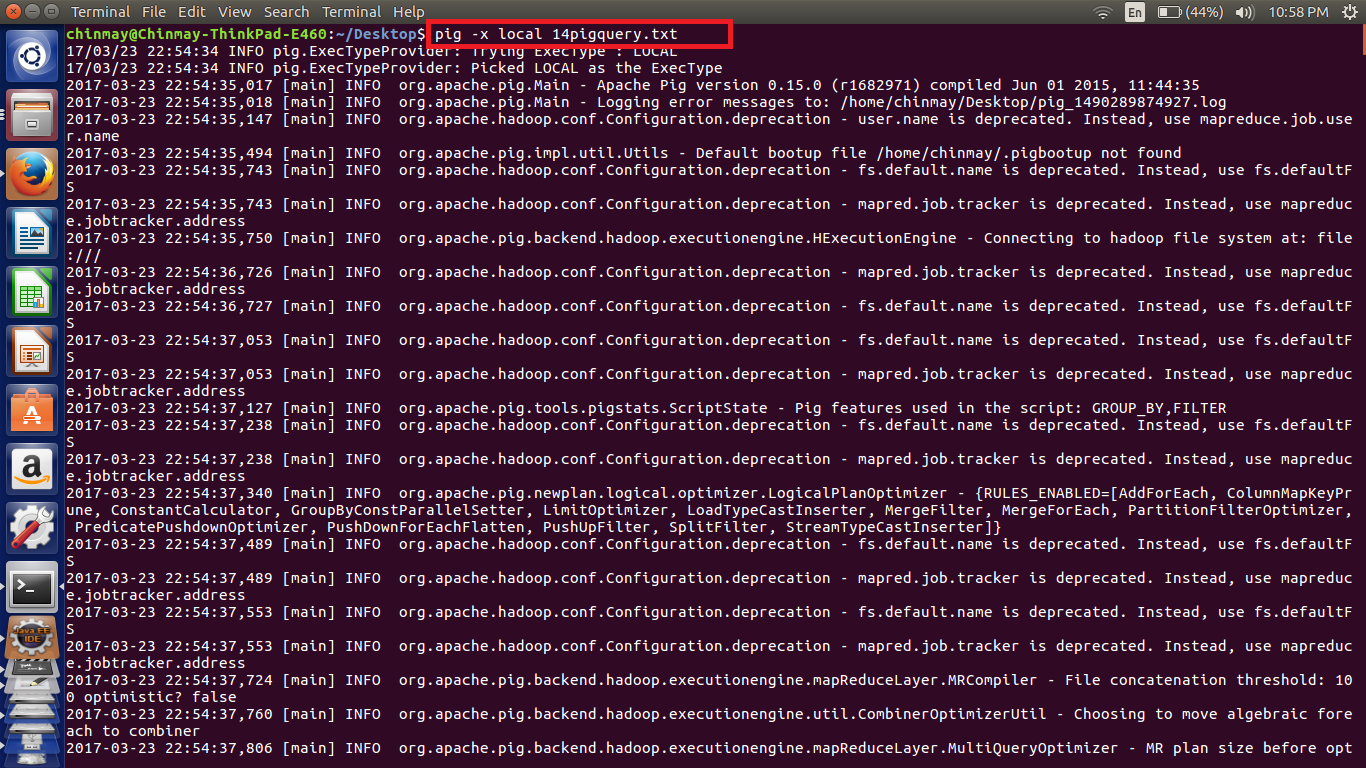
d = group c all;

e = FOREACH d GENERATE COUNT(c.id);

dump e;

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

//implementation



In this pig script first loaded the fields with respective data type.

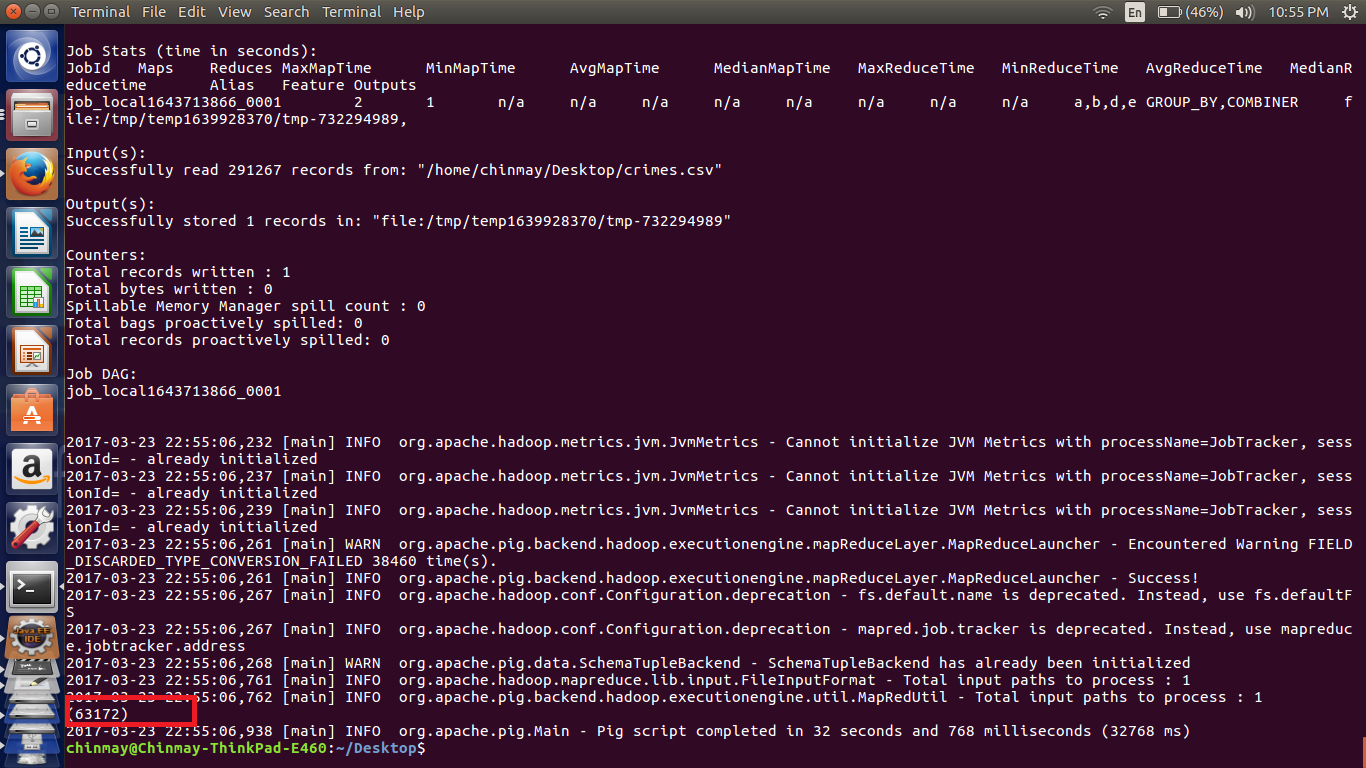
Filtered the data if the arrests are true

Then as we have to use the date function so convert the date fields into the date format using the TO DATE fuction in the pig.

Then filtered the dates with the AND operator to check before and after of the mentioned date.

Grouped by all the values and count the number of ids to get the arrests in that period.

//pig output



Here we can see that output of the simple mapreduce program is same as that of the pig implementation.