**1. Introduction**

This dataset contains attributes related to crimes taking place in various areas like type of crime, FBI code related to that criminal case, arrest frequency, location of crime etc.

**2. Associated Data Files**

https://drive.google.com/file/d/0B1QaXx7tpw3SaUJHOHBZclBXWG8/view?usp=sharing

**Dataset Description:**

ID, Case Number, Date, Block, IUCR, Primary Type,Description,Location Description,Arrest,Domestic,Beat,District,Ward,Community Area,FBICode,X Coordinate,Y Coordinate,Year,Updated On,Latitude,Longitude,Location.

**3. Problem Statement**

I choose to work with pig. So all scripts for the problem statements will be in Pig latin.

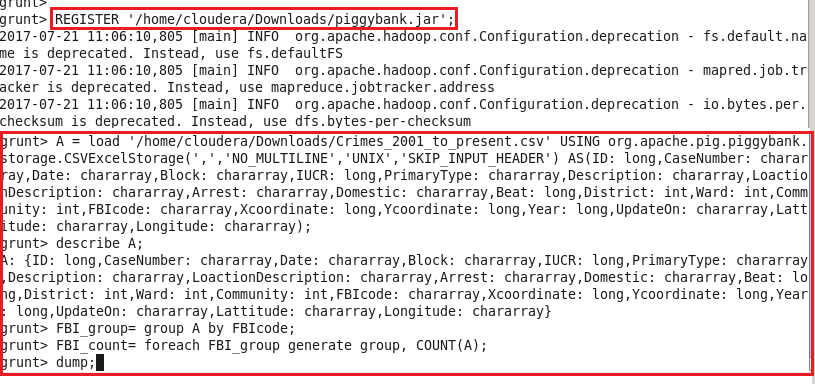
1. **Write a MapReduce/Pig program to calculate the number of cases investigated under each FBI code.**

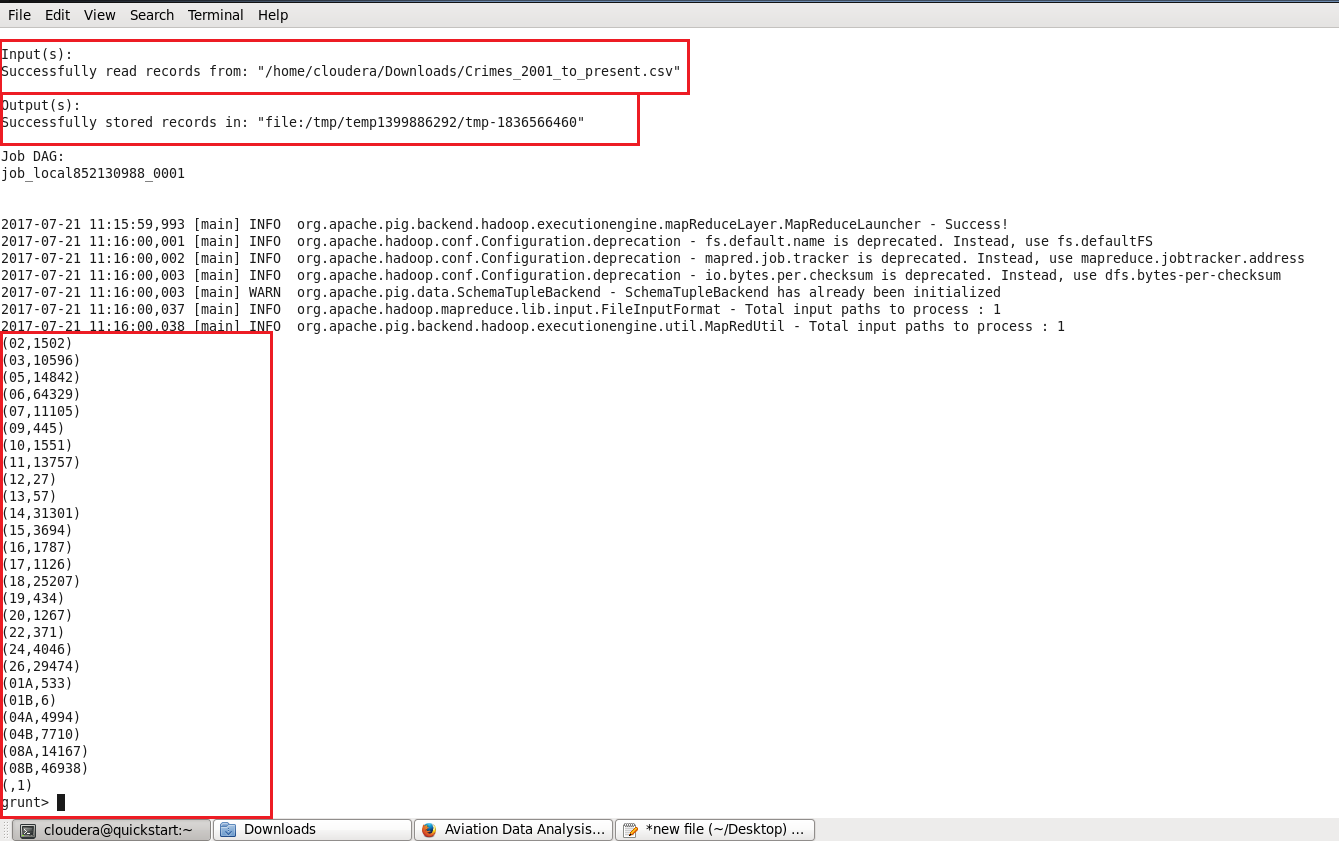
Step-1: We have to register the piggybank.jar file to get the functionalities of CSVExcelStorage which will be useful to store the Excel data.

Step-2: Now we have to load the data using one of the piggybank functionalities CSVExcelStorage and load the data into pig for analysis.

Step-3: To get the Number of cases investigated under each FBI code we need to group the data using FBI code and the find the count.

Step-4: Finally to get the result, we need to dump the data using dump command.



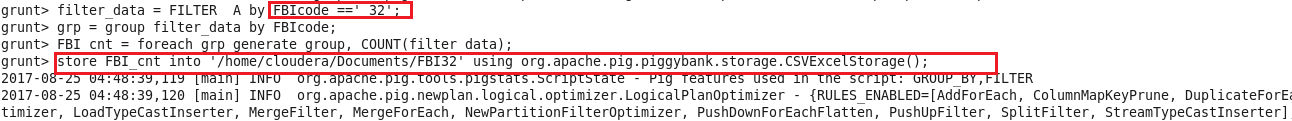


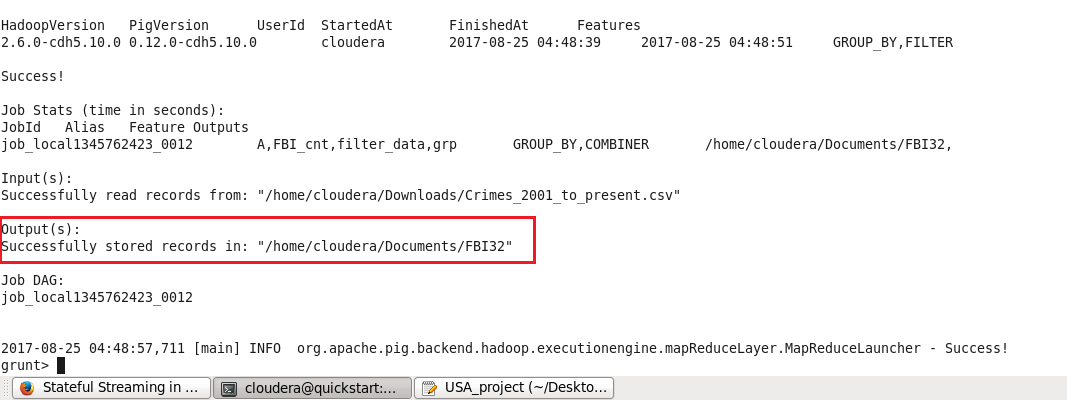
**Output:** In the output we can see the number of cases investigated under each FBI code.

1. **Write a MapReduce/Pig program to calculate the number of cases investigated under FBI code 32.**

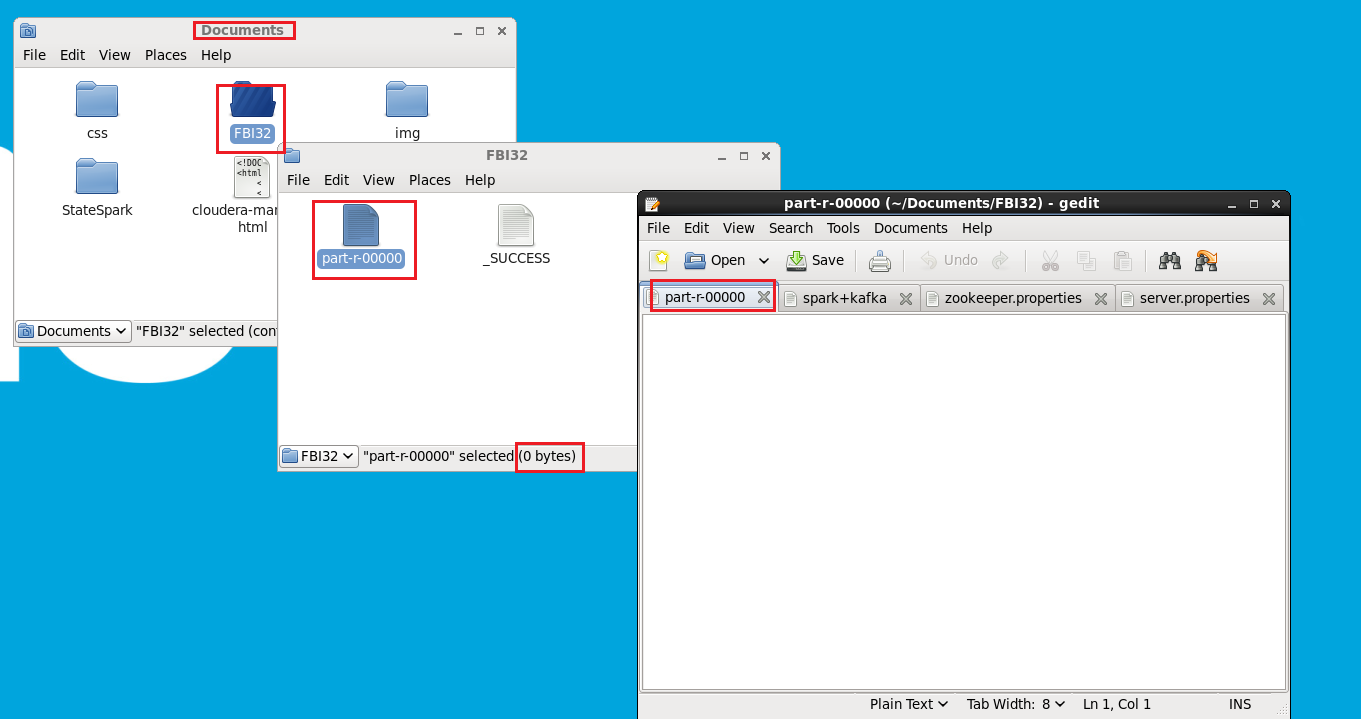
**Case 1:**

In this case we need find out number of cases investigated under FBI code 32 only.



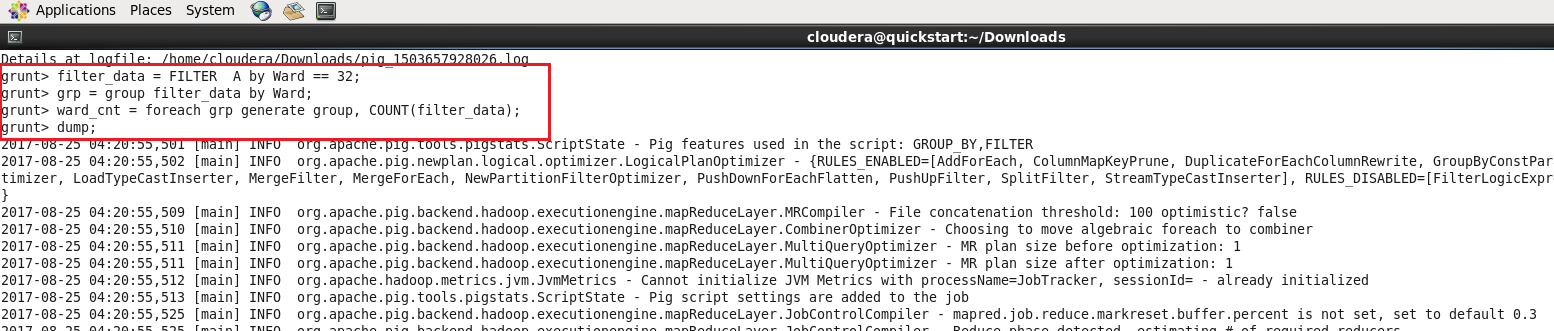


In the output, you can see the number of cases for FBI code 32 only is stored under this path /home/cloudera/Documents/FBI32

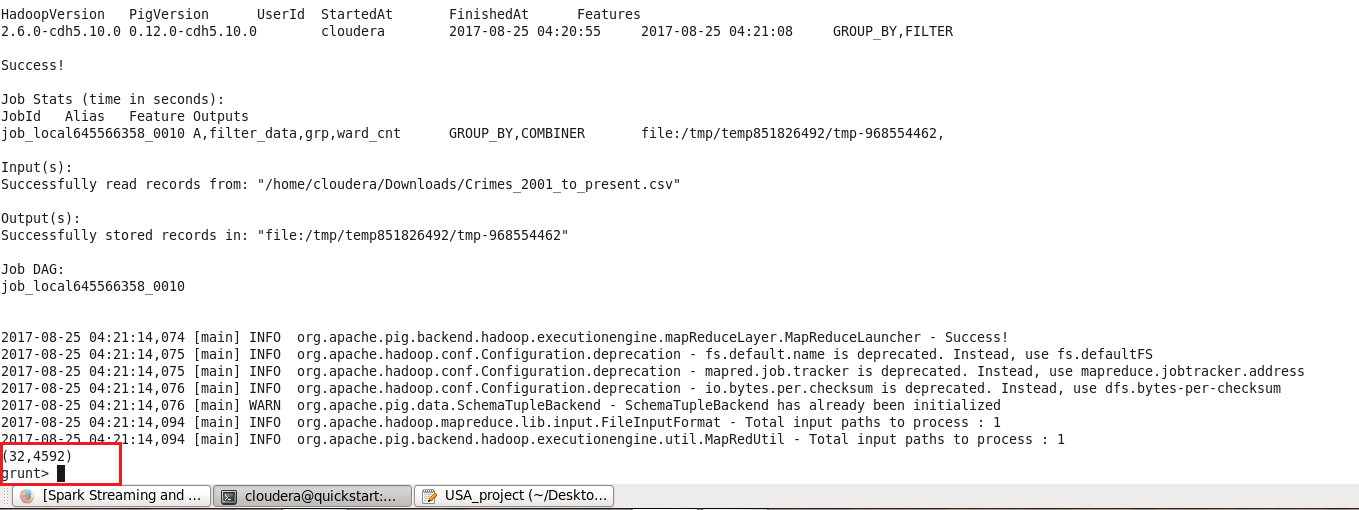


As you can see that there are no records for the FBI code 32 because there is no FBI code 32 in the given data. So we are getting zero records.

**Case-2: (suggested by support team to find number of cases investigated for ward 32).**



Here I am filtering records for ward == 32 and getting the required count.

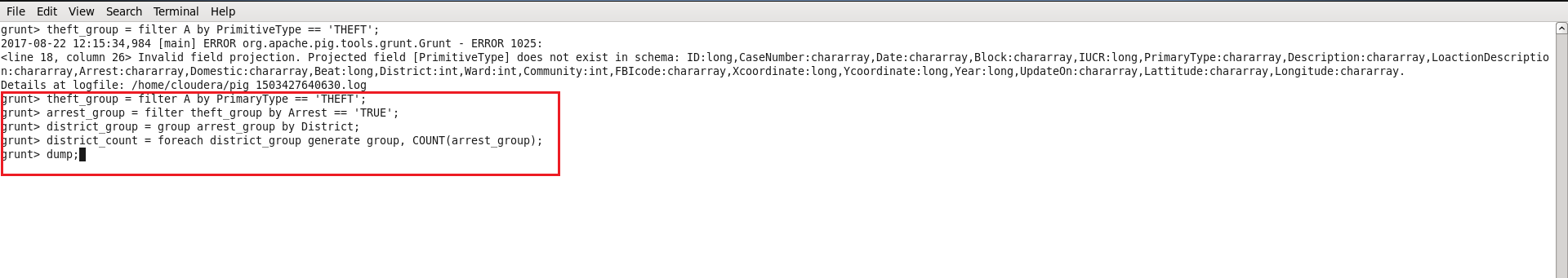


So in the output we got the number of cases investigated under Ward 32 is 4592.

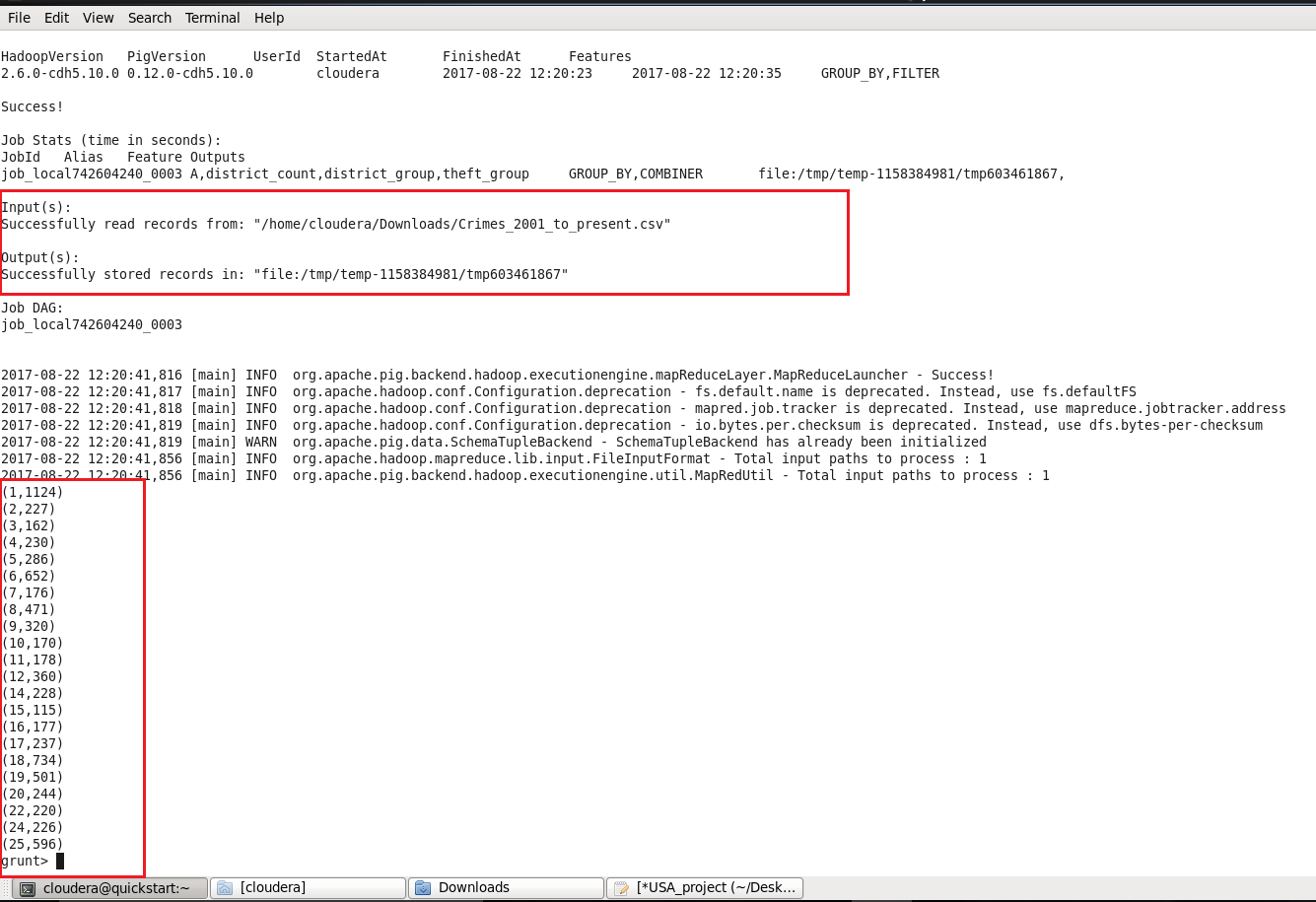
1. **Write a MapReduce/Pig program to calculate the number of arrests in theft district wise.**

In this problem 3 conditions are there. They are:

1. Arrests that fall under THEFT category.
2. Since it is arrest implies the Arrest field which is Boolean should be TRUE.
3. Last one is we should group the number of arrest based on District wise.

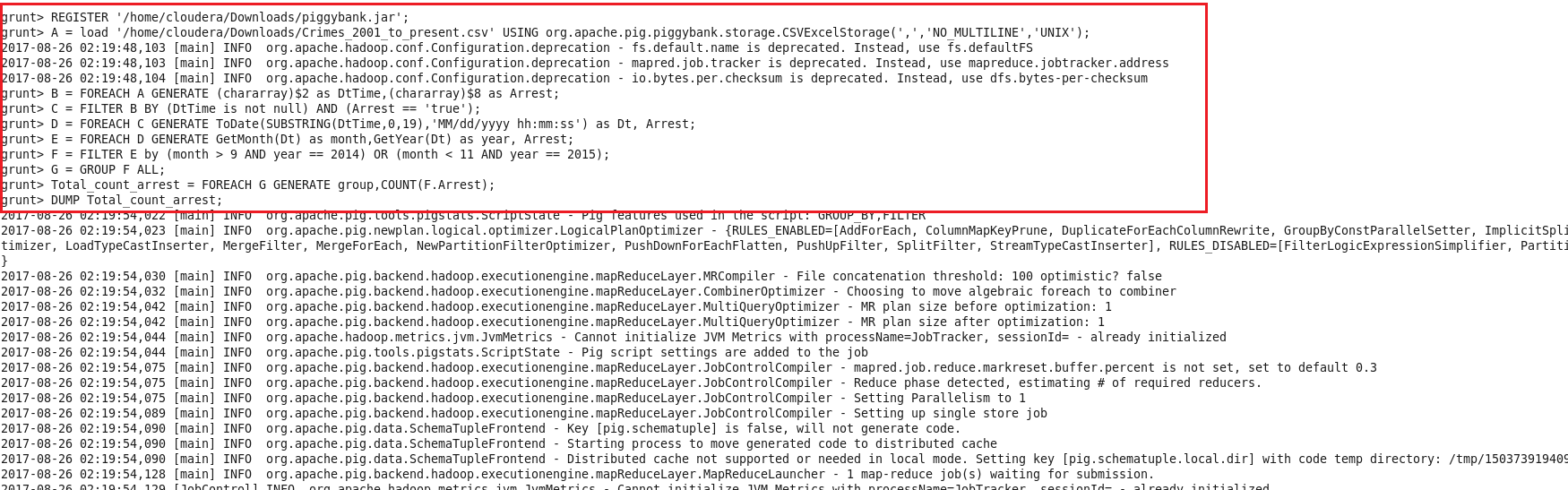


**Output:**



Here district is not shown in the output because for district 31 has only one case which has Arrest = false. So it won’t be included in the output.

1. **Write a MapReduce/Pig program to calculate the number of arrests done between October 2014 and October 2015.**

****

**Steps:**

In this query first we are taking the required columns date and arrest by using foreach.

Next step is filter records by Arrest = true, here we will get all the records that have Arrest column as True.

In this step we have convert chararray date field to date type using ToDate() function and get the required fields month and year.

Here we will filter data between OCT 2014 and OCT 2015.

Group data based on that and generate the count value.

**Output:**

