**PROJECT – 1**

**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. Prerequisites:**

You should have Hadoop cluster installed in your system.

**3. 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, FBI Code, X Coordinate, Y Coordinate, Year, Updated On, Latitude, Longitude, Location

**4. Problem Statement:**

* Write a mapreduce and pig program to calculate the number of cases investigated under each FBI code.
* Write a mapreduce and pig program to calculate the number of cases investigated under FBI code 32.
* Write a mapreduce and pig program to calculate the number of

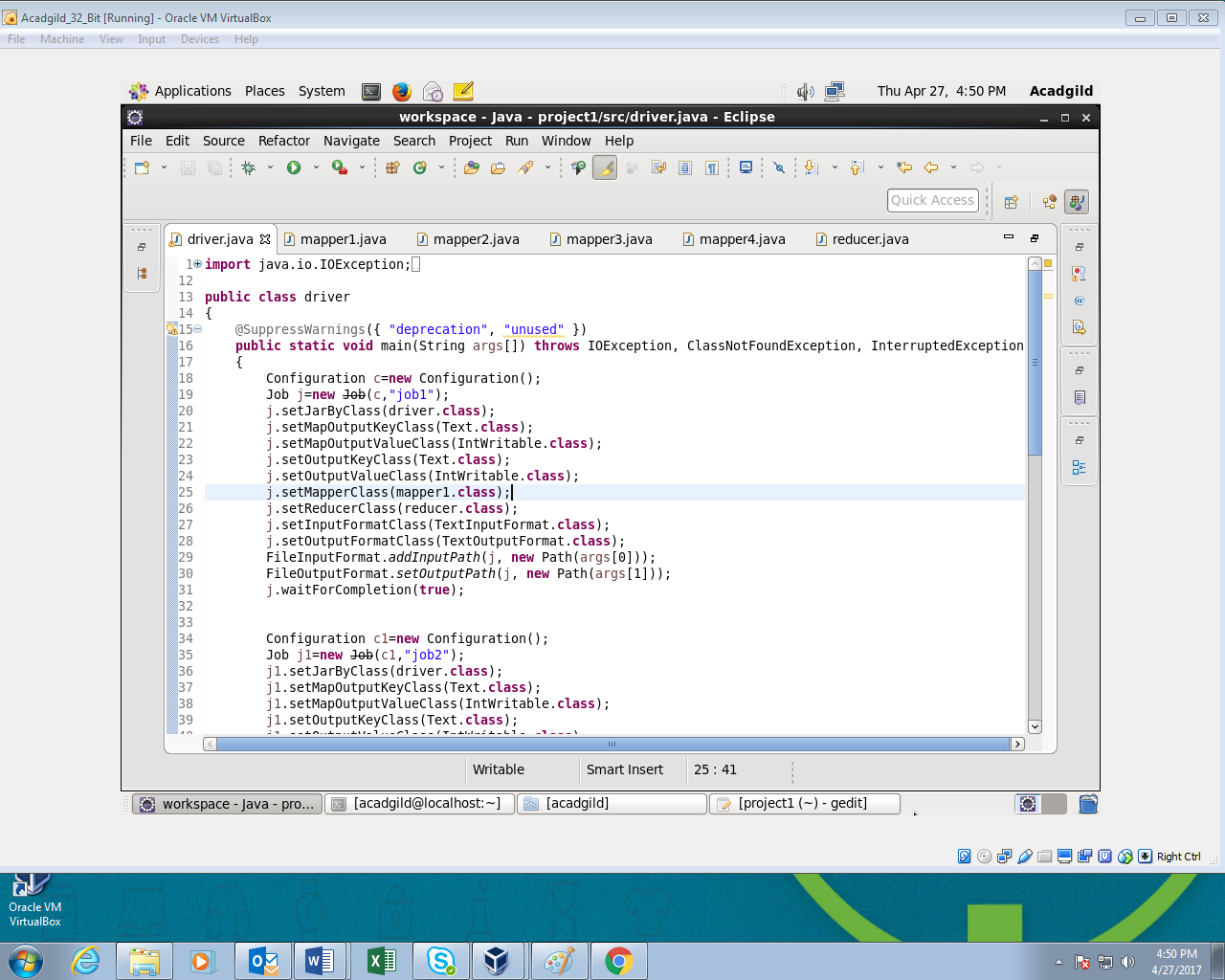
Arrests in theft district wise.

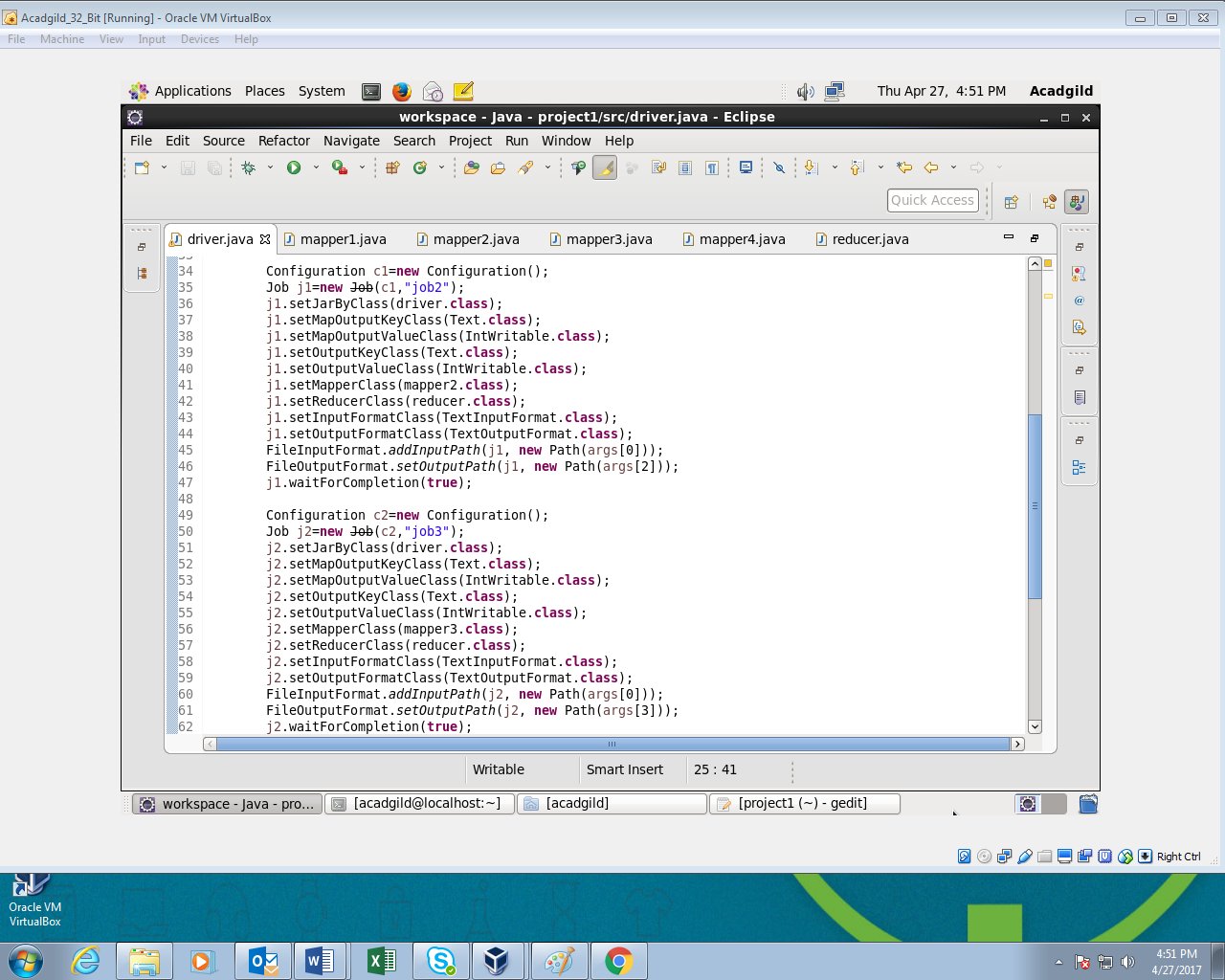
* Write a mapreduce and pig program to calculate the number of arrests done between October 2014 and October 2015.

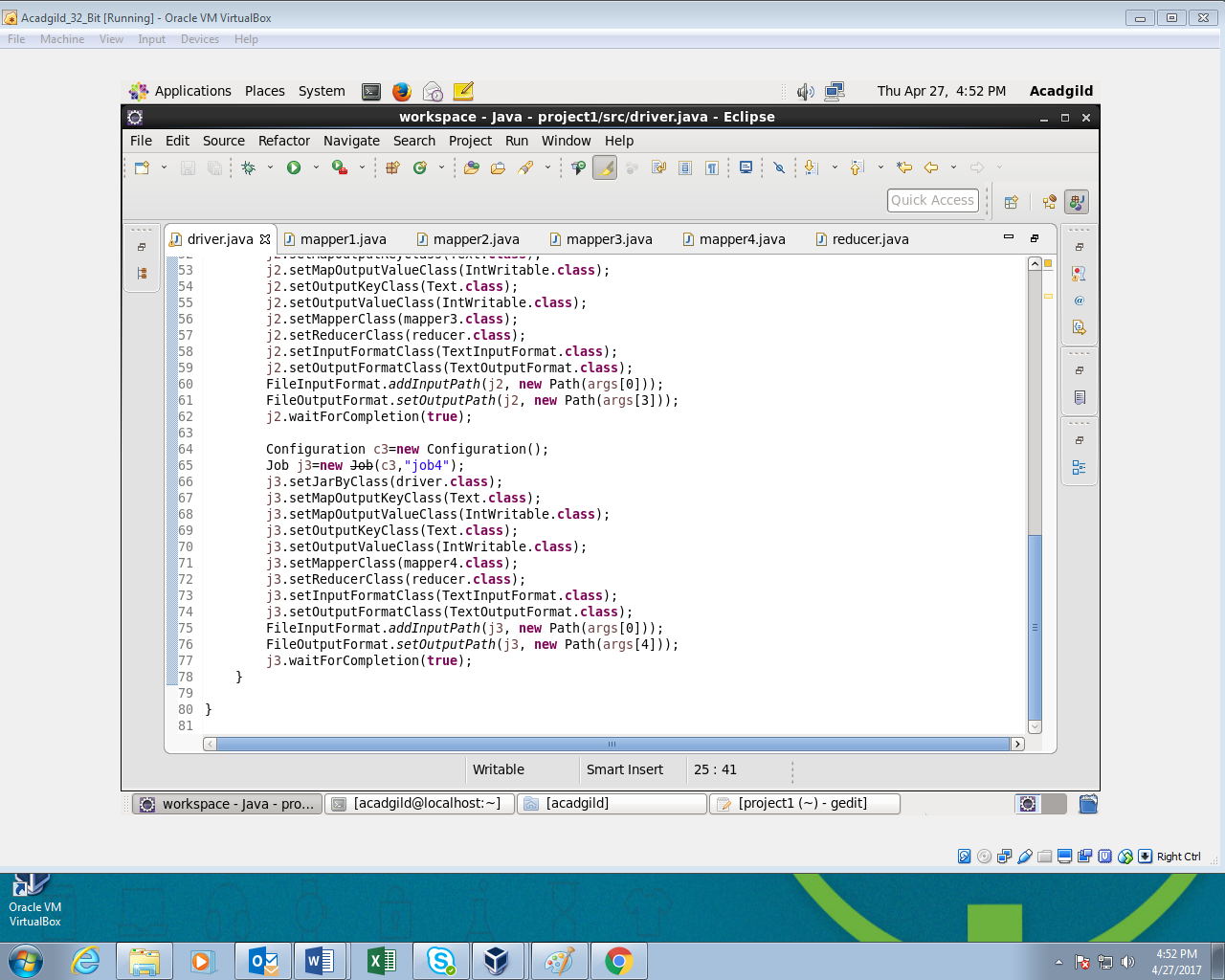
**5. Program:**

**MAPREDUCE APPROACH:**

**Driver class:**



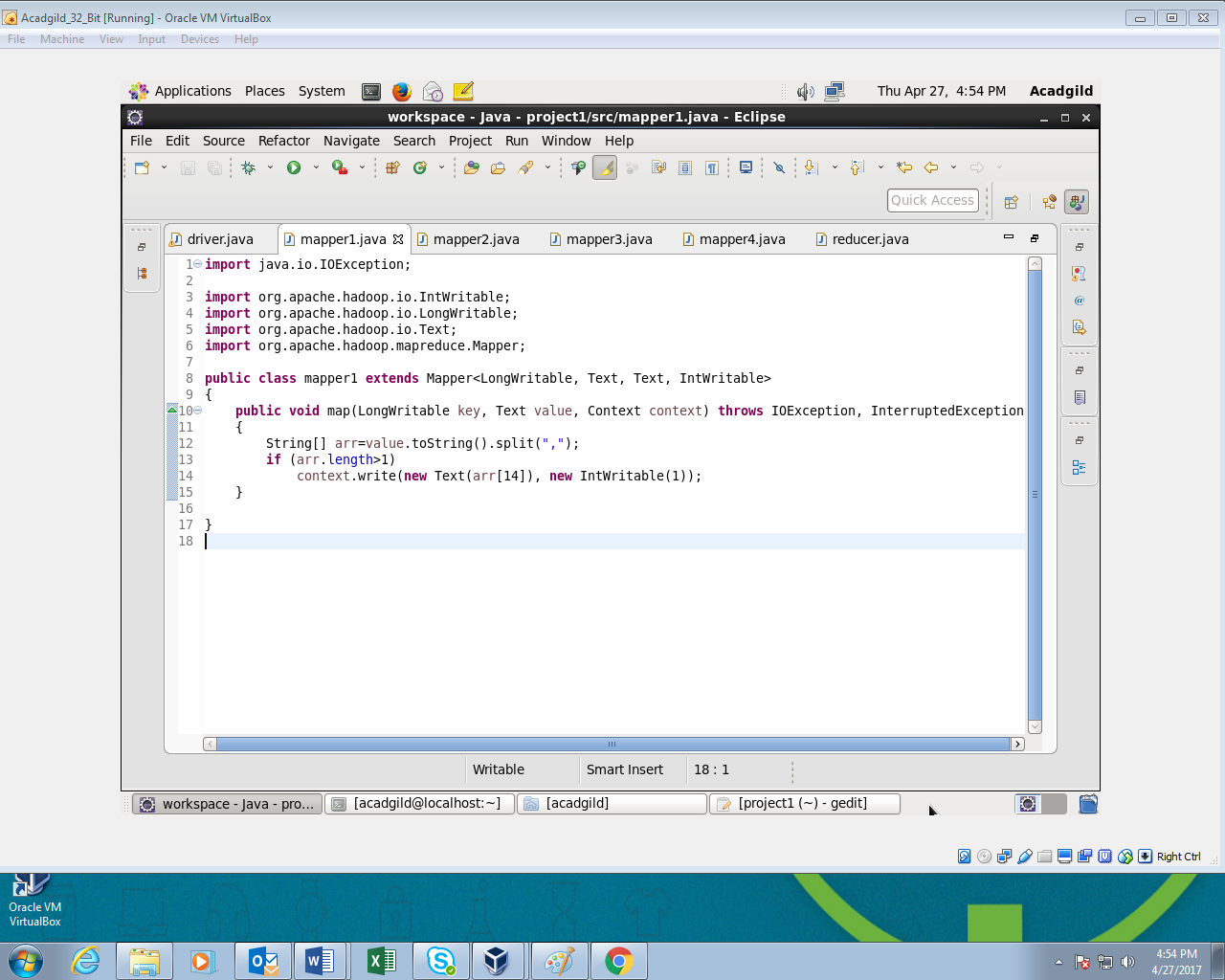




* In driver class four jobs are initiated: one for each problem respectively.
* The path arguments are set for input file and output files.
* In this class mapper and reducer for each problem is set.

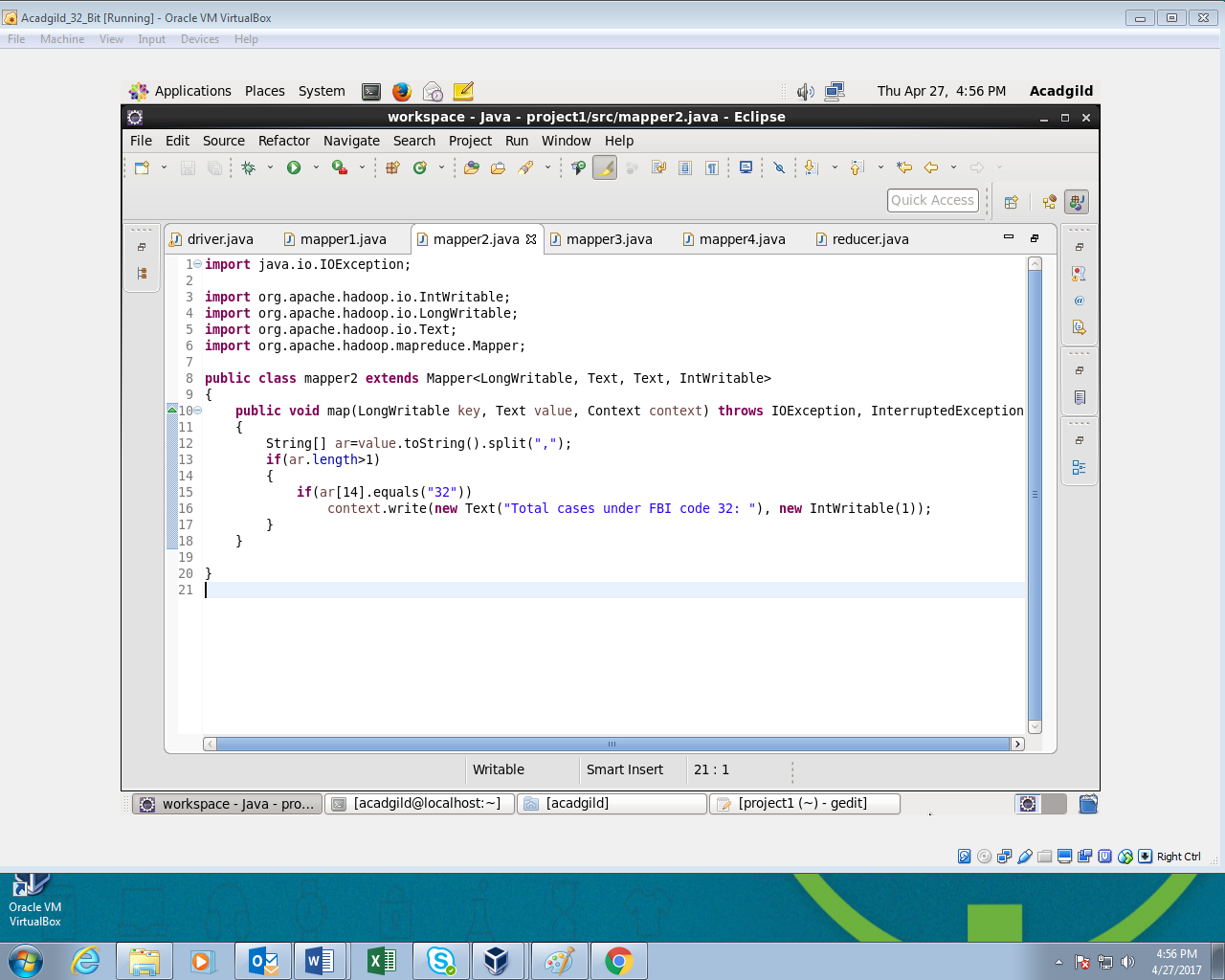
**Mapper class for problem 1:**

The mapper class is executed for each line in the input database. In this class FBI code is a key and 1 is a value for each FBI code.



**Mapper class for problem 2:**

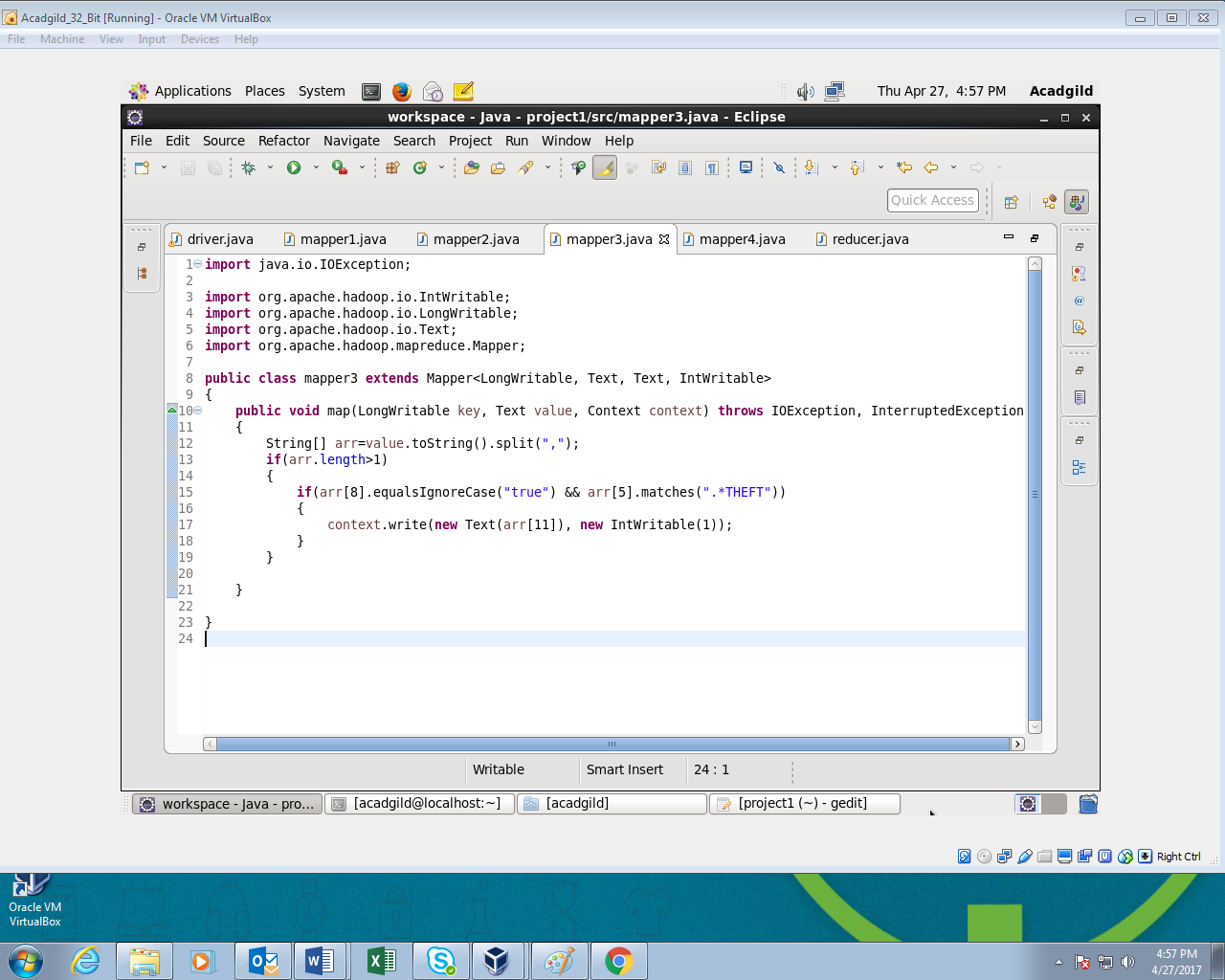
In this class the map method checks if FBI code is 32 and all the crime records with matching FBI code are given as an input to the reducer class. The reducer then calculates the count of total records in this category.



**Mapper class for problem 3:**

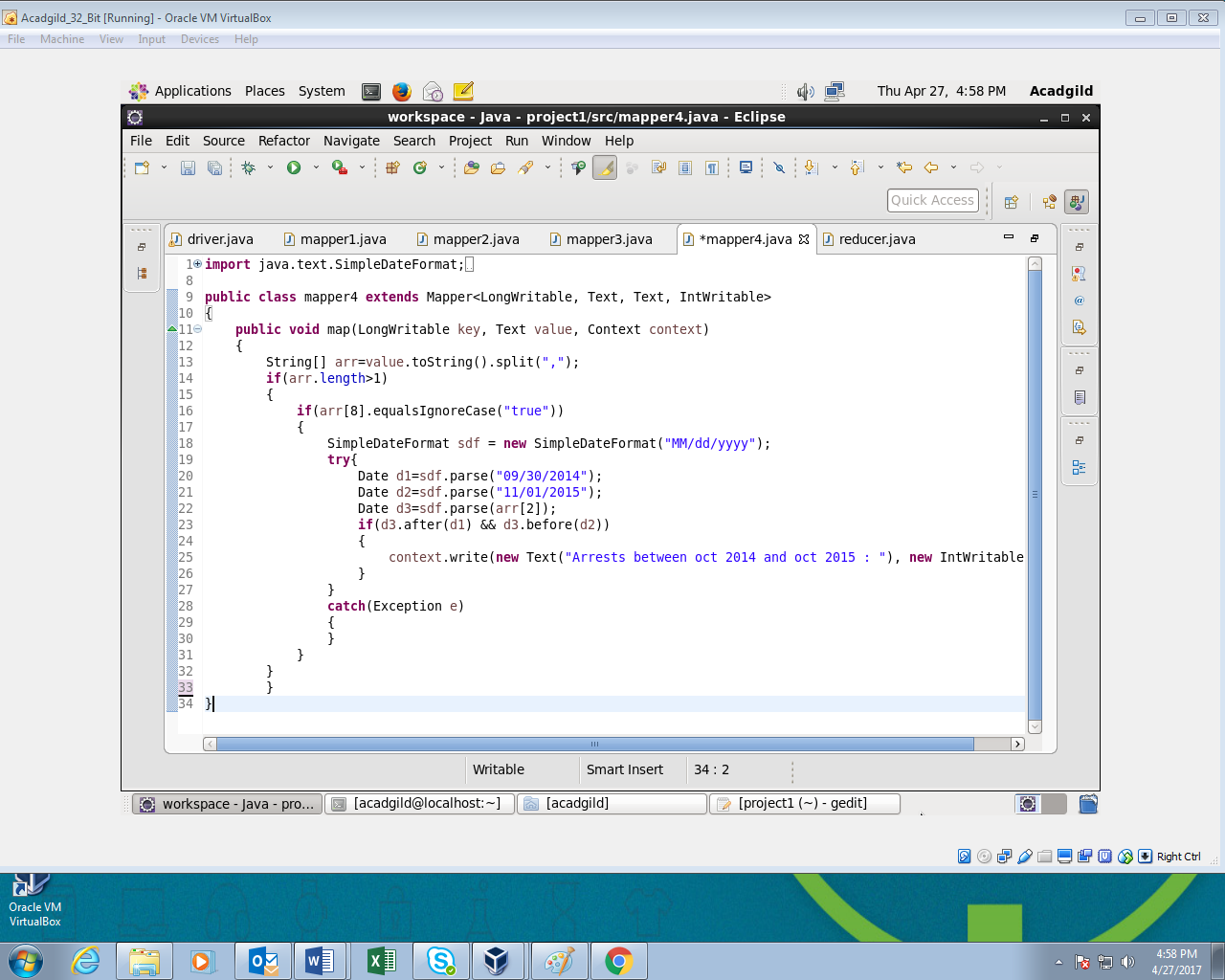
In this mapper class key is district code and value is 1.

In this class if primary type of the crime is “THEFT” then those records are filtered and given as an input to the reducer. The reducer then calculates the count of total records in this category district wise.



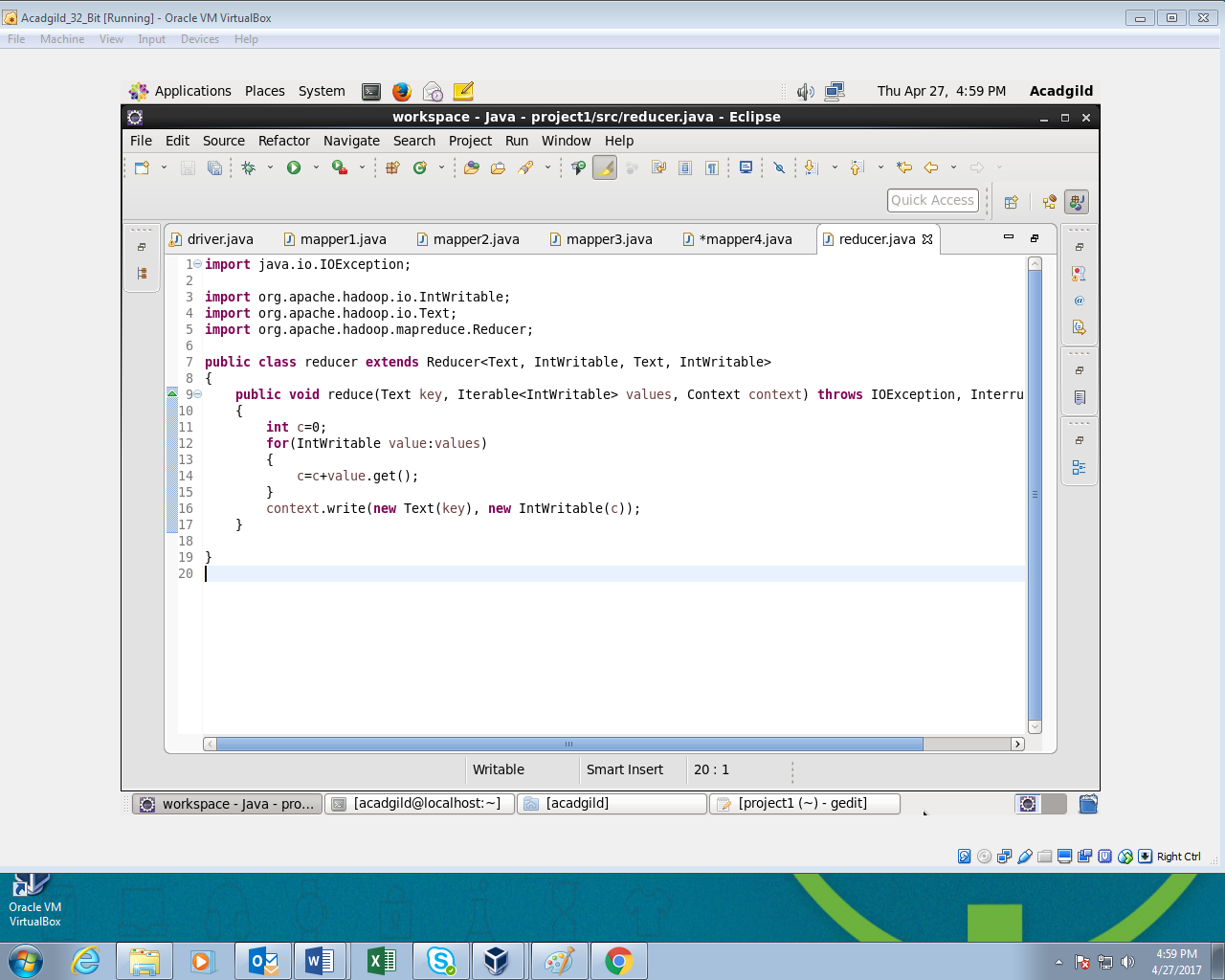
**Mapper class for problem 4:**

In this mapper class all the arrests between October 2014 and October 2015 are filtered and given as an input to the reducer.

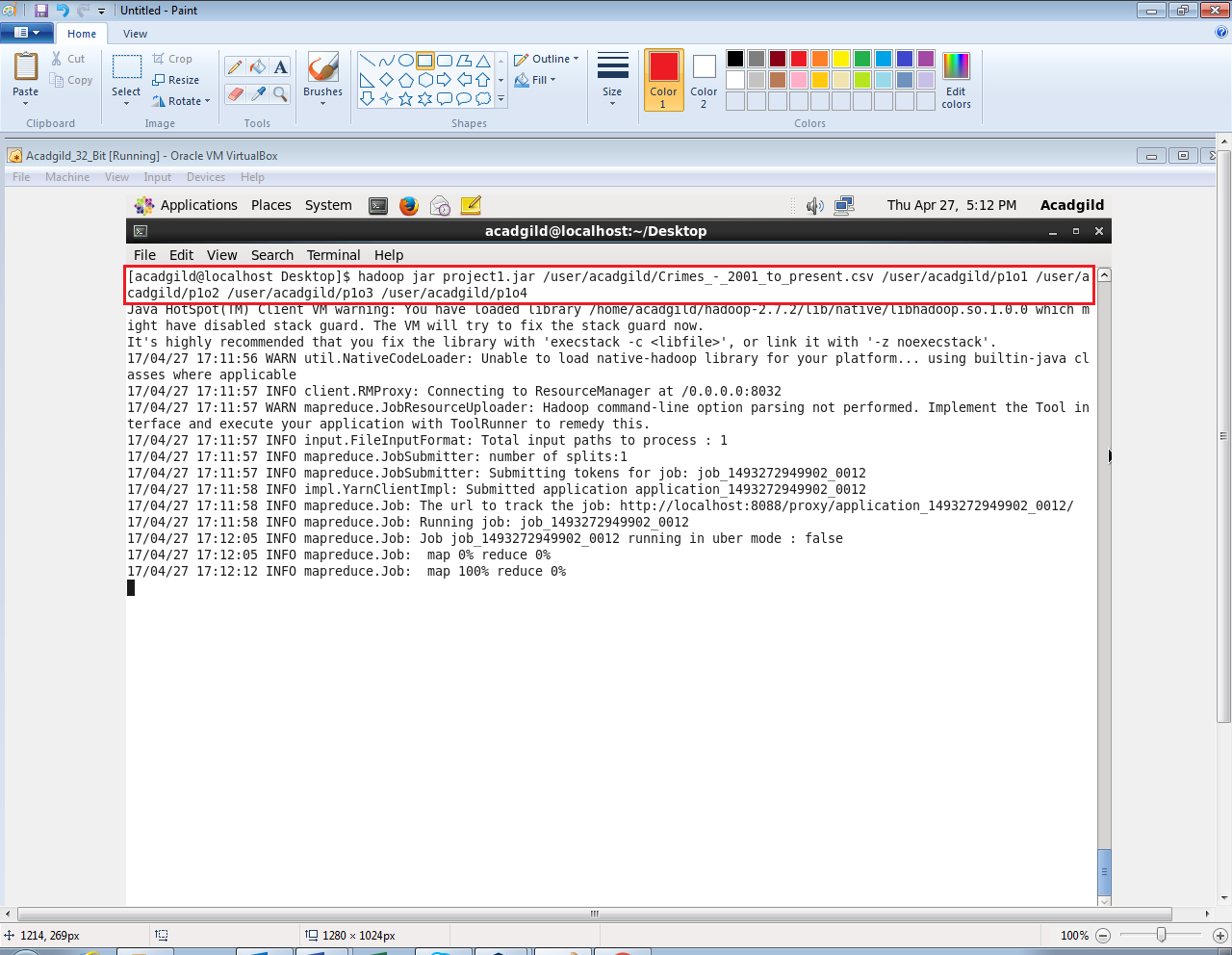


**Reducer class:**

The reducer is same for all the problems as the reducer logic required is same for all the problems.

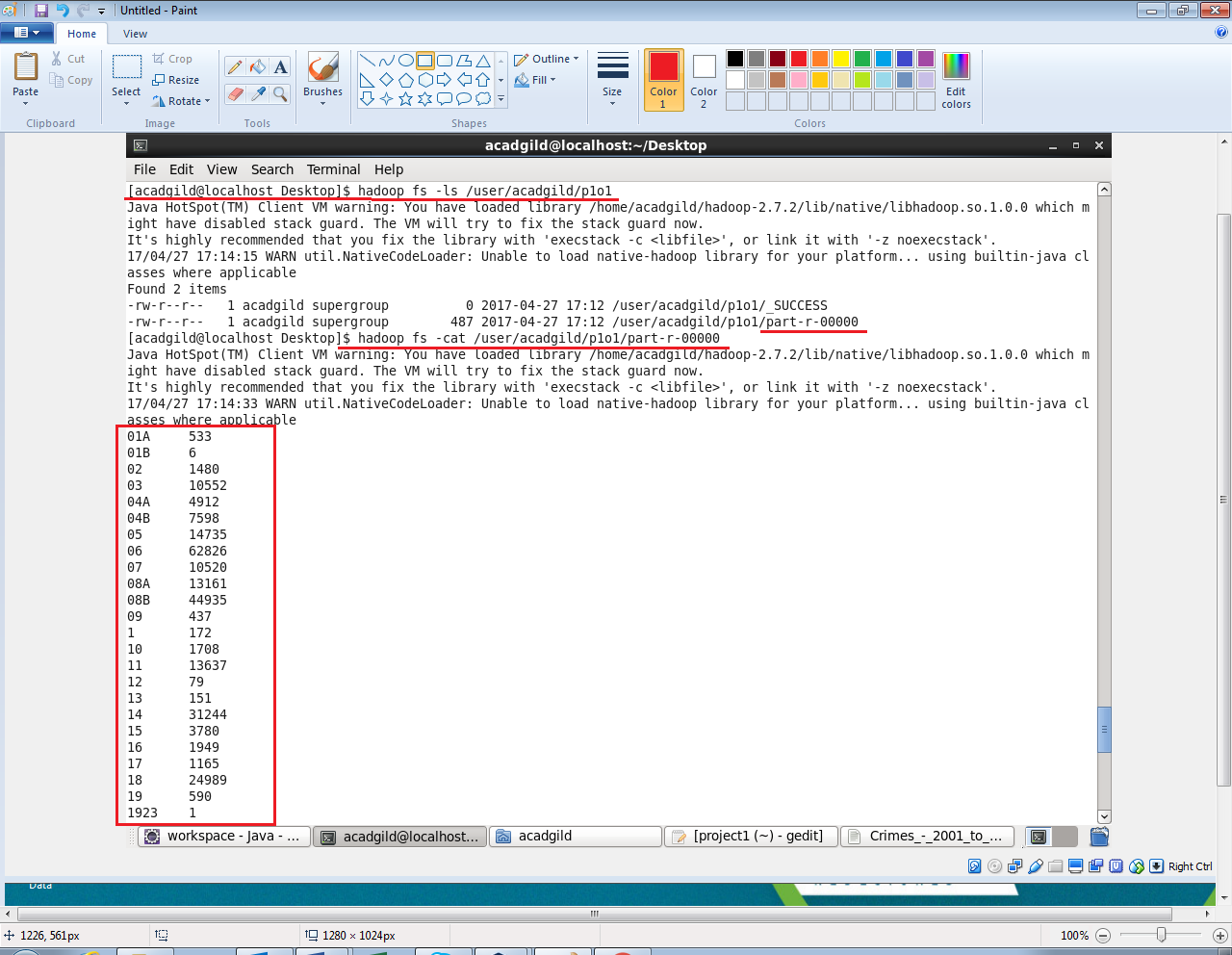


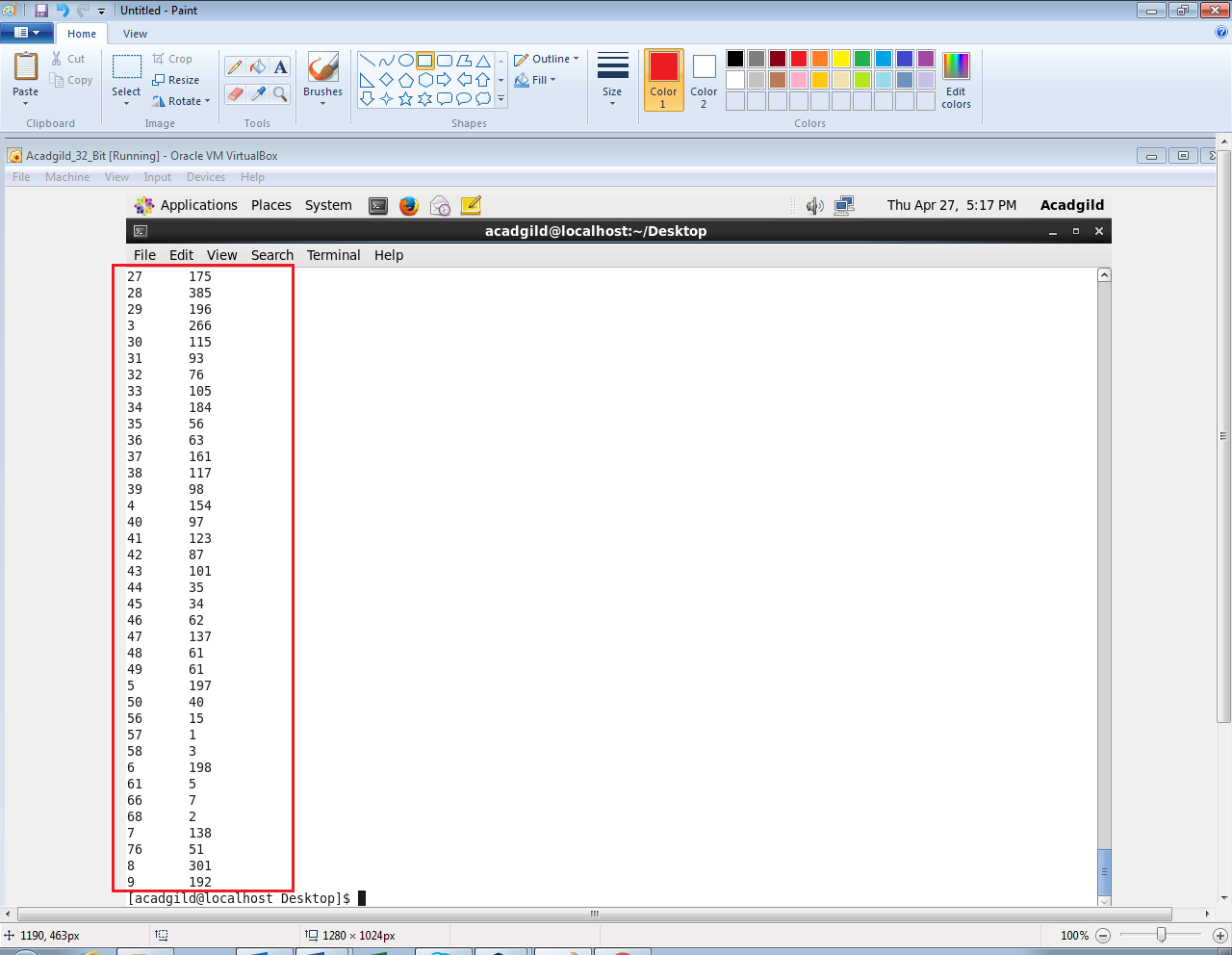
**OUTPUT:**



**Output 1:**

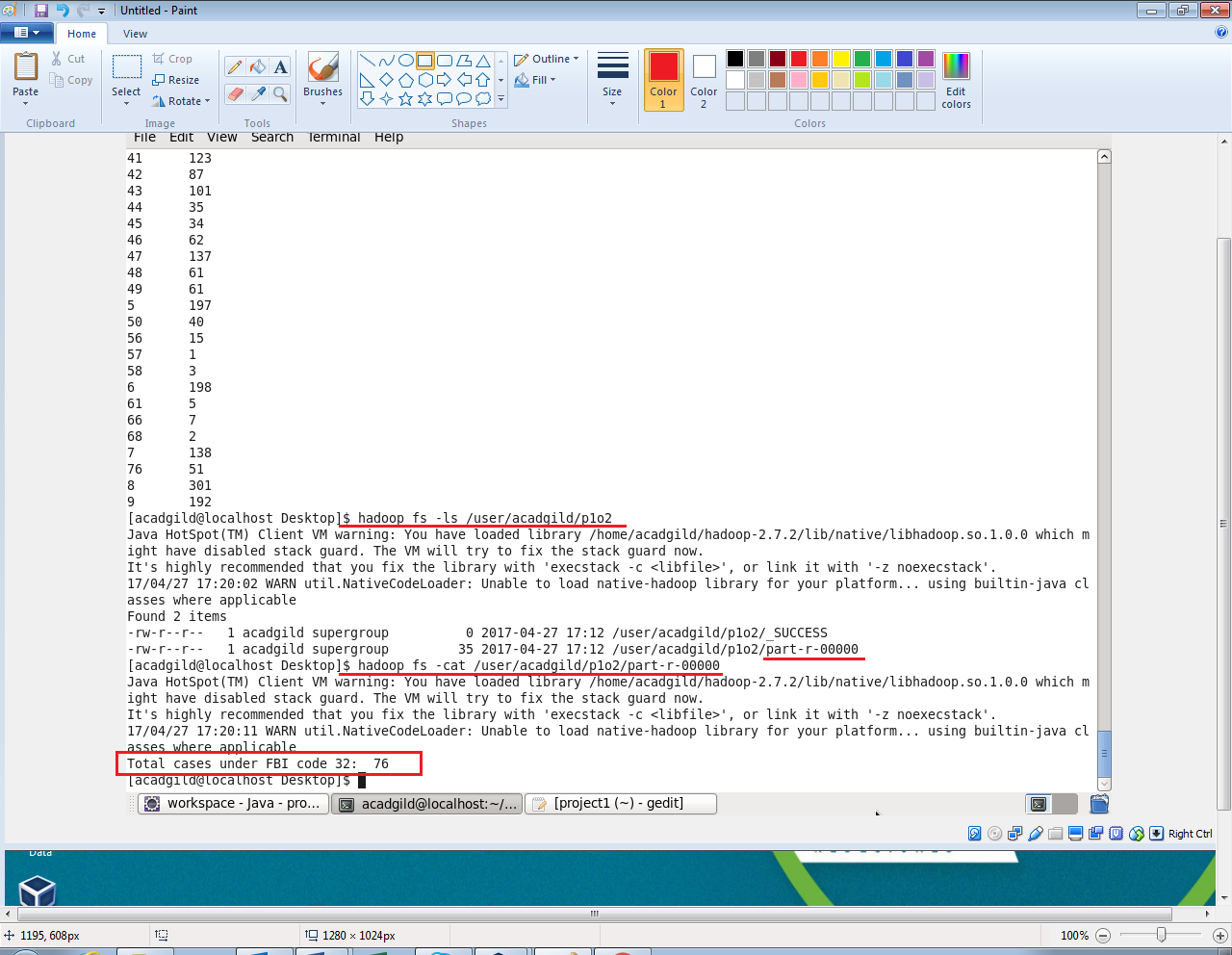
* It contains part file which holds the number of cases investigated under each FBI code.





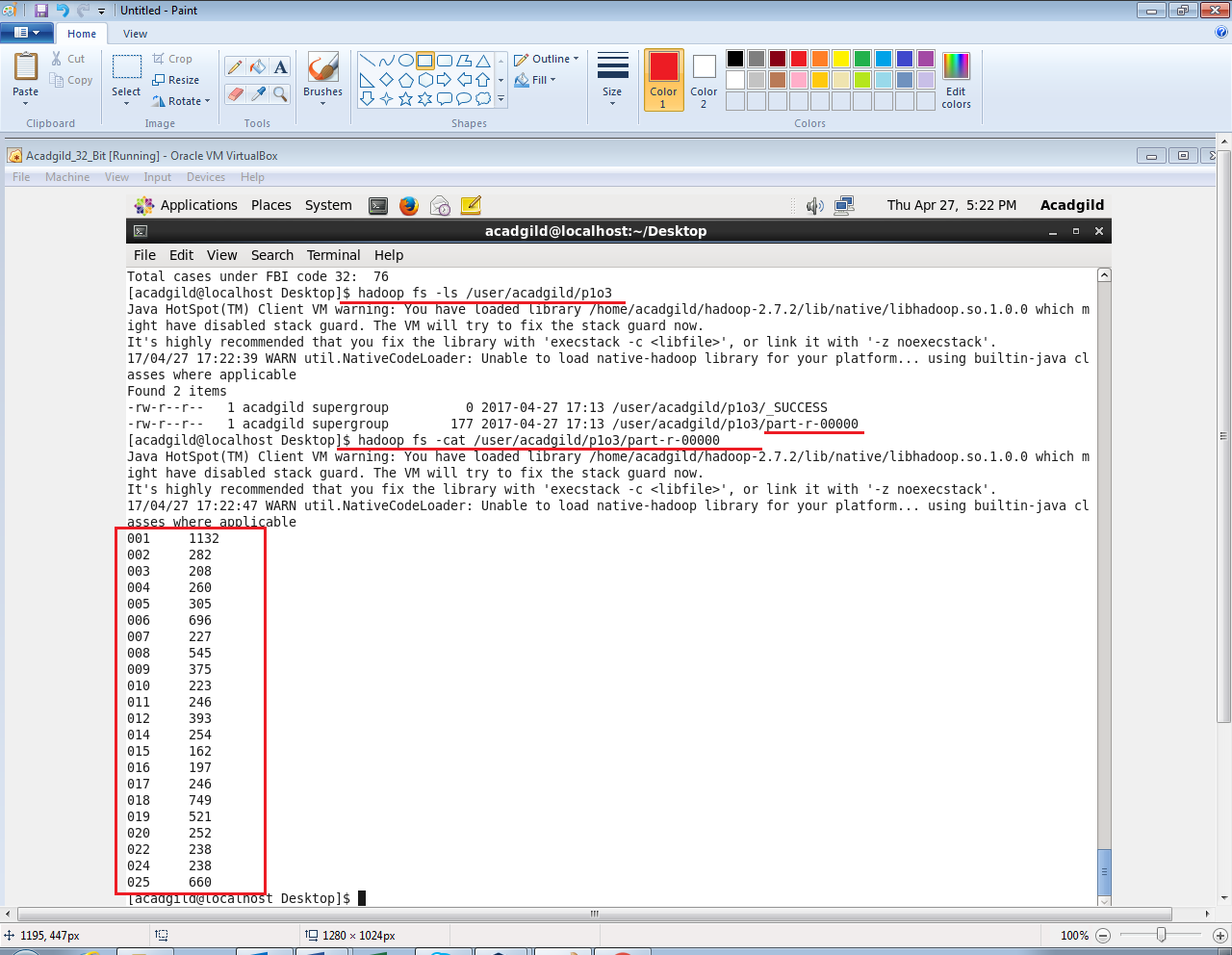
**Output 2:**

It contains the number of cases investigated under FBI code 32.



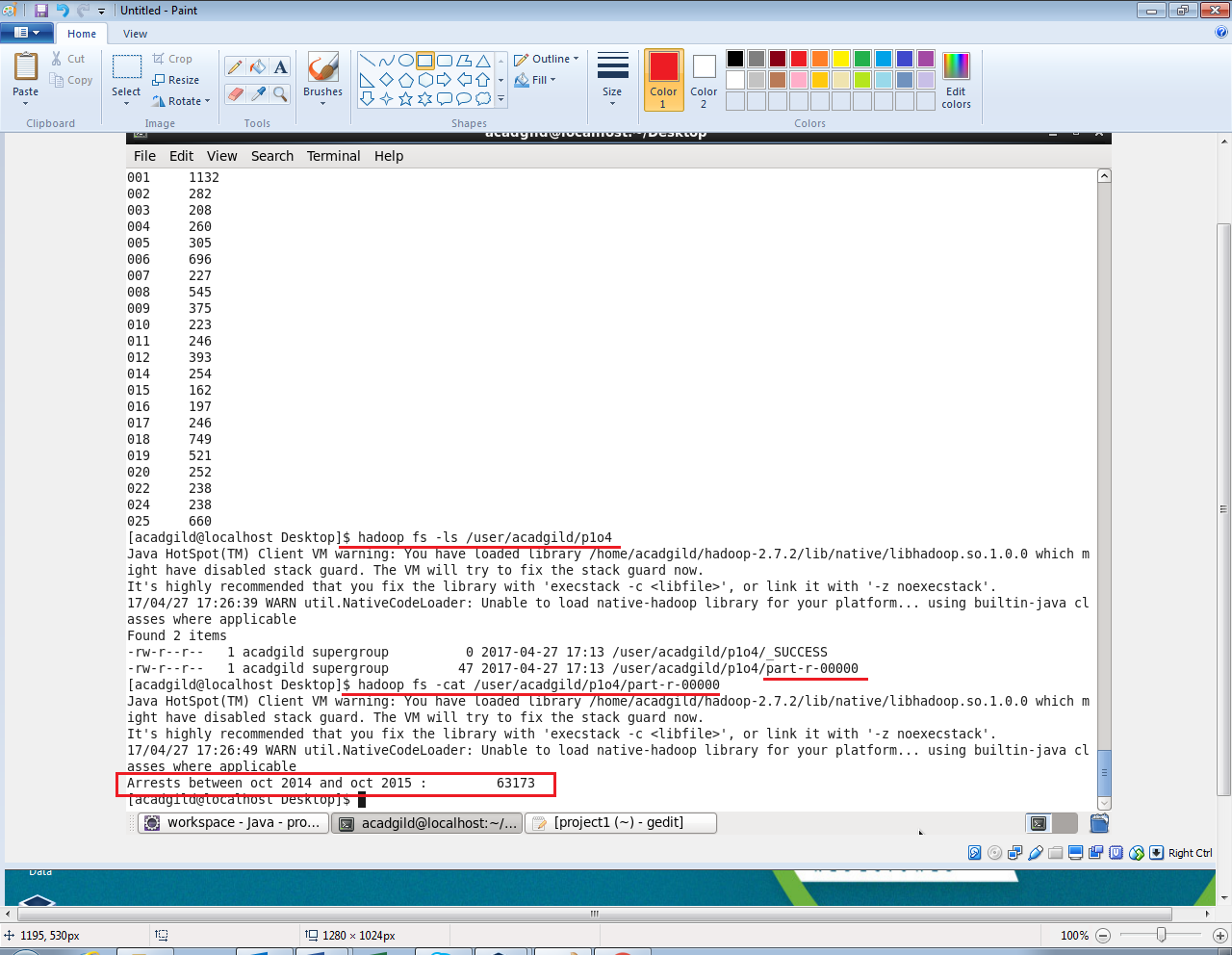
**Output 3:**

* It contains the number of cases investigated under FBI code 32.



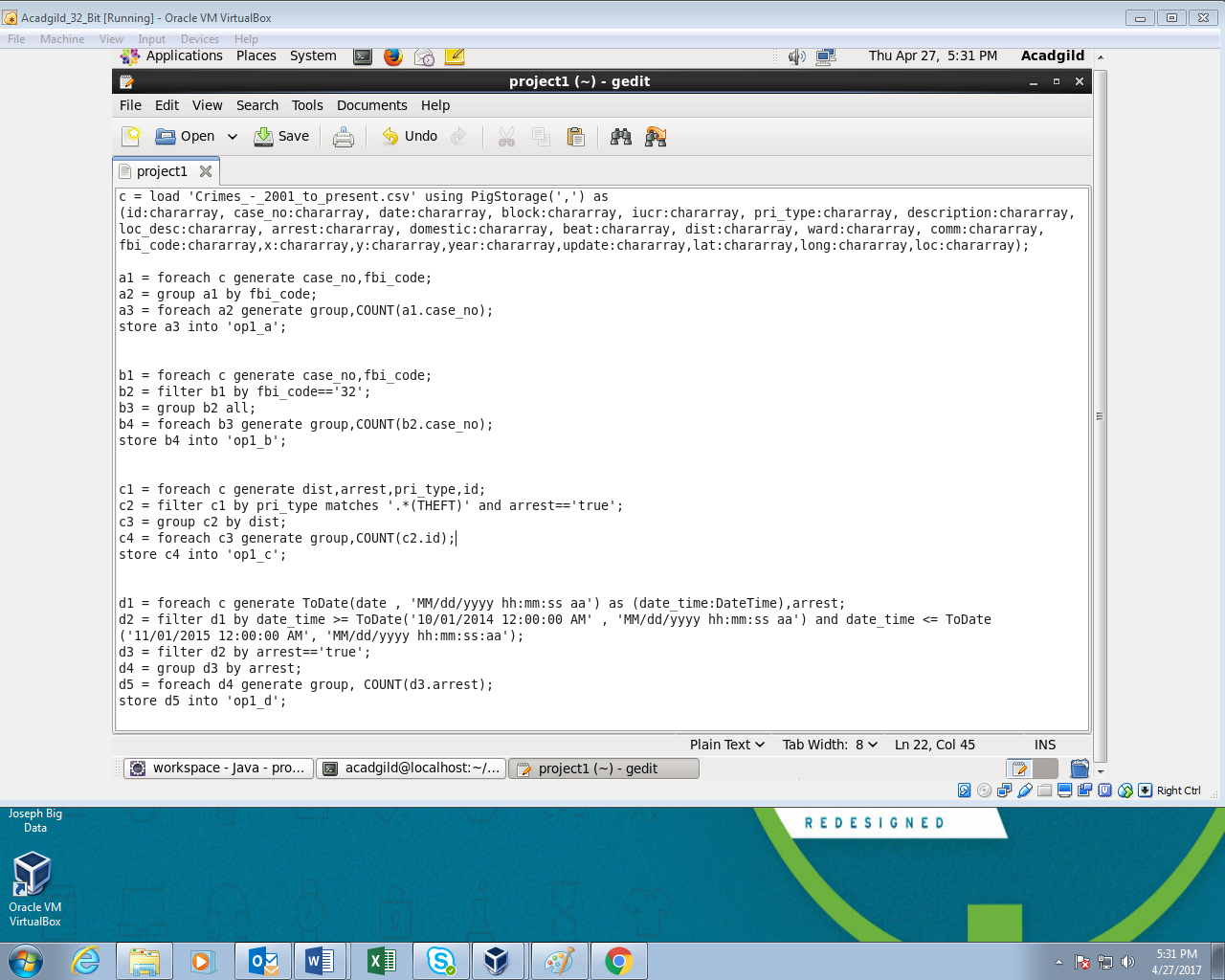
**Output 4:**

* It contains the part file which has the number of arrests done between October 2014 and October 2015.



**PIG APPROACH:**

**PIG SCRIPT:**



**DESCRIPTION FOR PIG SCRIPT:**

This script has solution for all the four problems in a sequence of problem statements.

* c: All the contents of the input dataset are loaded in ‘c’. The structure of the input data is defined in here. As the data is separated using ‘,’ it is provided in PigStorage().

**Solution for problem 1:**

* a1: The records of required fields (case number and FBI code) are loaded.
* a2: The records are then grouped with respect to FBI code.
* a3: For each FBI code total count of records is calculated.
* op1\_a: The output is stored in part file inside this folder.

**Solution for problem 2:**

* b1: The records of required fields (case number and FBI code) are loaded.
* b2: The records are then filtered where FBI code is 32.
* b3: The records are grouped in this step.
* b4: The total count of records is calculated here.
* op1\_b: The output is stored in part file inside this folder.

**Solution for problem 3:**

* c1: The records of required fields (district, arrest, primary type and id) are loaded.
* c2: The records are then filtered where primary type of crime is “THEFT” and arrest field contains value as “true”.
* c3: The records are grouped district wise.
* c4: The total count of records is calculated for each district.
* op1\_c: The output is stored in part file inside this folder.

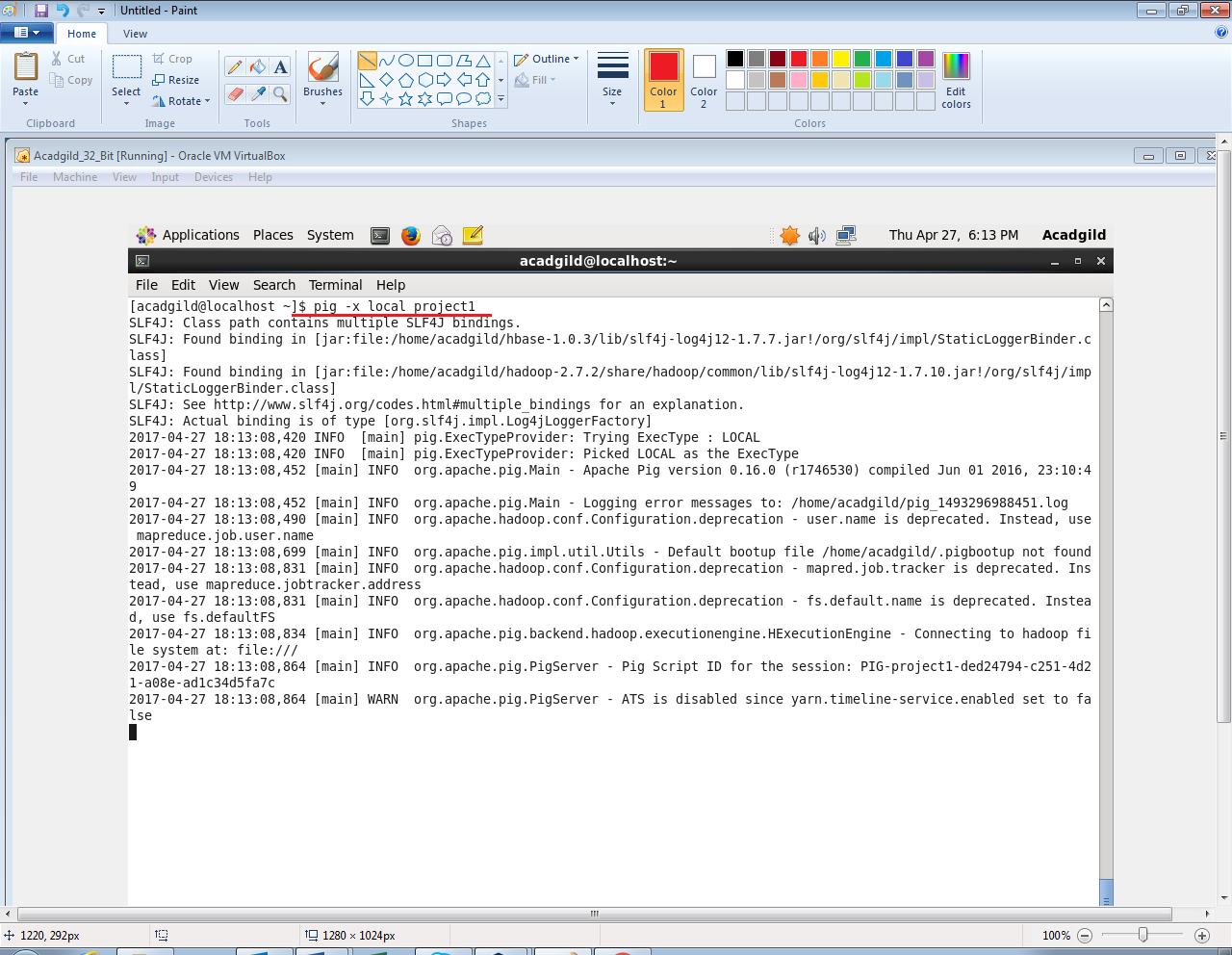
**Solution for problem 4:**

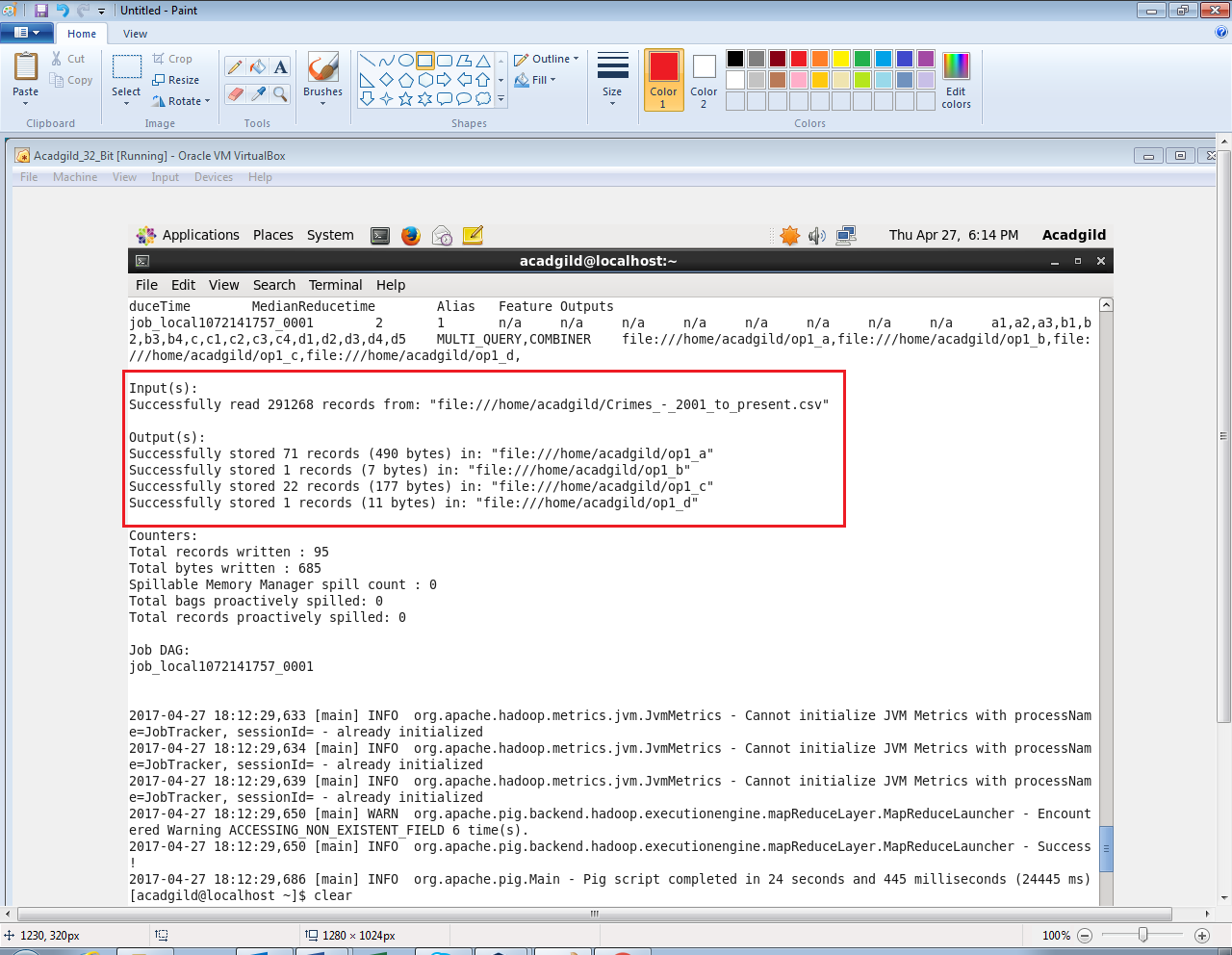
* d1: The records of required fields (date and arrest) are loaded. In this step the date is converted into defined format using “ToDate()”.
* d2: The records are then filtered where date is in between October 2014 and October 2015.
* d3: The records are filtered where arrest field contains value as “true”.
* d4: The records then are grouped by arrest field.
* d5: The count of arrests is generated in this step.
* op1\_d: The output is stored in part file inside this folder.

**OUTPUT:**

The pig script is executed using following command

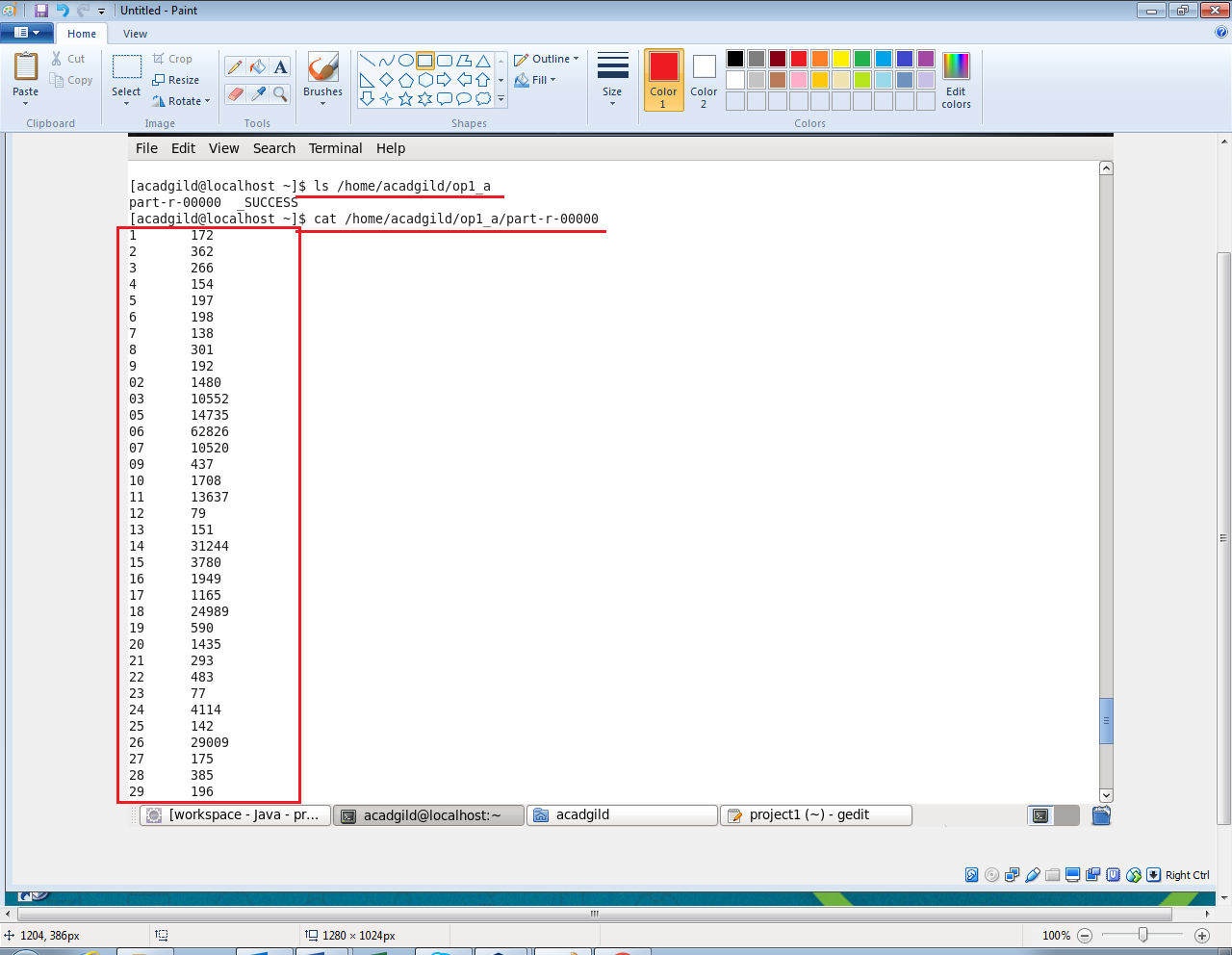
The contents of the output files are displayed using cat command:

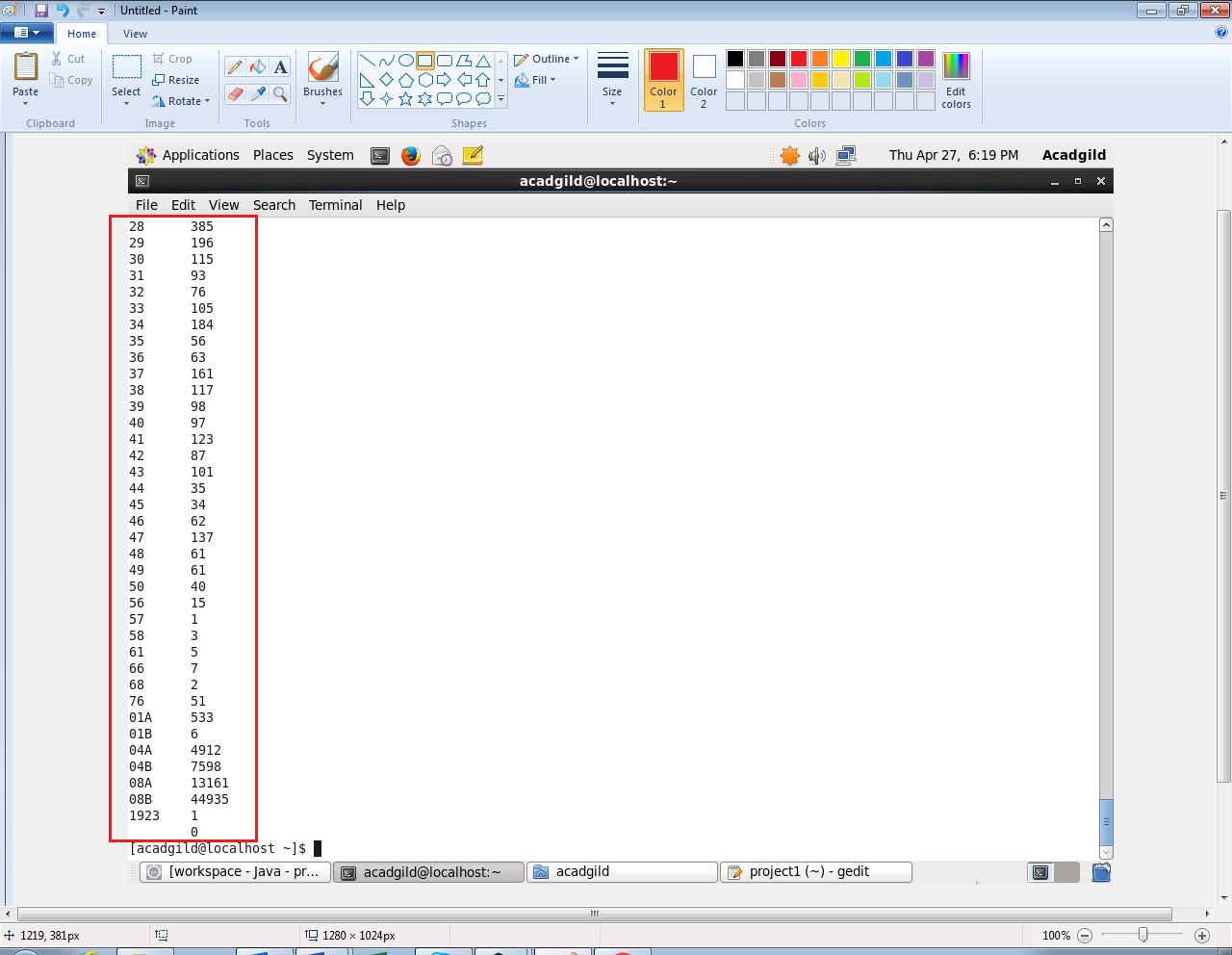




**Output 1:**

* op1\_1 contains part file which holds the number of cases investigated under each FBI code.





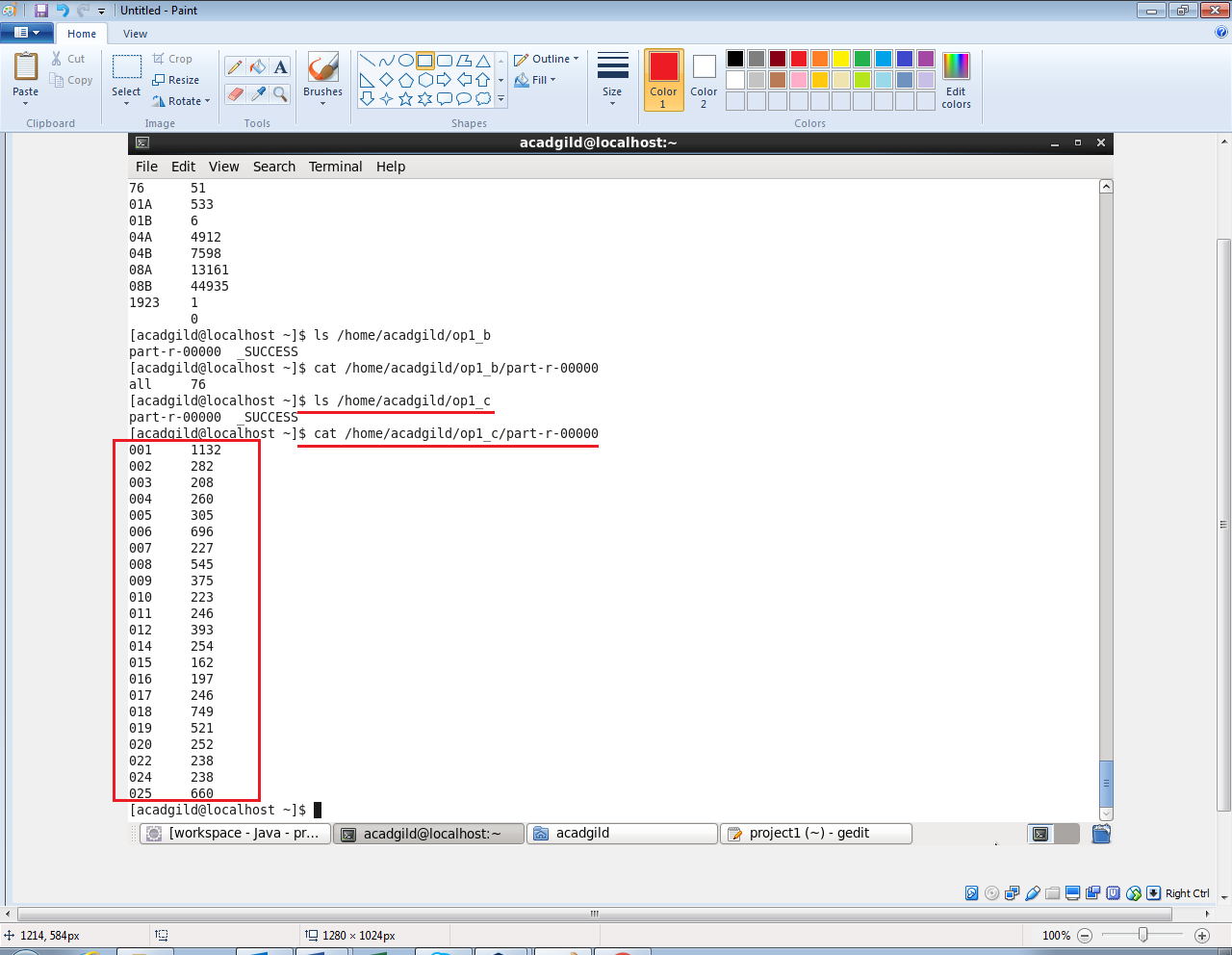
**Output 2:**

* op1\_b has part file which contains the number of cases investigated under FBI code 32.



**Output 3:**

* op1\_c contains the part file which has the number of arrests in theft district wise.



**Output 4:**

* op1\_d contains the part file which has the number of arrests done between October 2014 and October 2015.

