# Big Data Hadoop Training

Project: USA Crime Analysis Solution

**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

**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.**

**It is a .csv file.**

**Input Dataset stored in /My\_Downloads/Assignments\_BDH/USA\_Crime\_Analysis\_Pig\_Project/Crimes\_-\_2001\_to\_present.csv' file in local system.**

****

**(a) Write a pig program to calculate the number of cases investigated under each FBI code**

A) **Pig Latin Script**:

REGISTER '/usr/lib/pig/piggybank.jar';

crime\_dataset = LOAD './Crimes\_-\_2001\_to\_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'UNIX') AS (ID:int,Case\_Number:chararray,Date:chararray,Block:chararray,IUCR:int,Primary\_Type:chararray,Description:chararray,Location\_Description:chararray,Arrest:chararray,Domestic:chararray,Beat:int,District:int,Ward:int,Community\_Area:int,FBICode:chararray,X\_Coordinate:int,Y\_Coordinate:int,Year:int,Updated\_On:chararray,Latitude:double,Longitude:double,Location:tuple(Latitude:double,Longitude:double));

FBI\_group = GROUP crime\_dataset BY FBICode;

no\_cases\_per\_FBICode = FOREACH FBI\_group GENERATE group, COUNT(crime\_dataset);

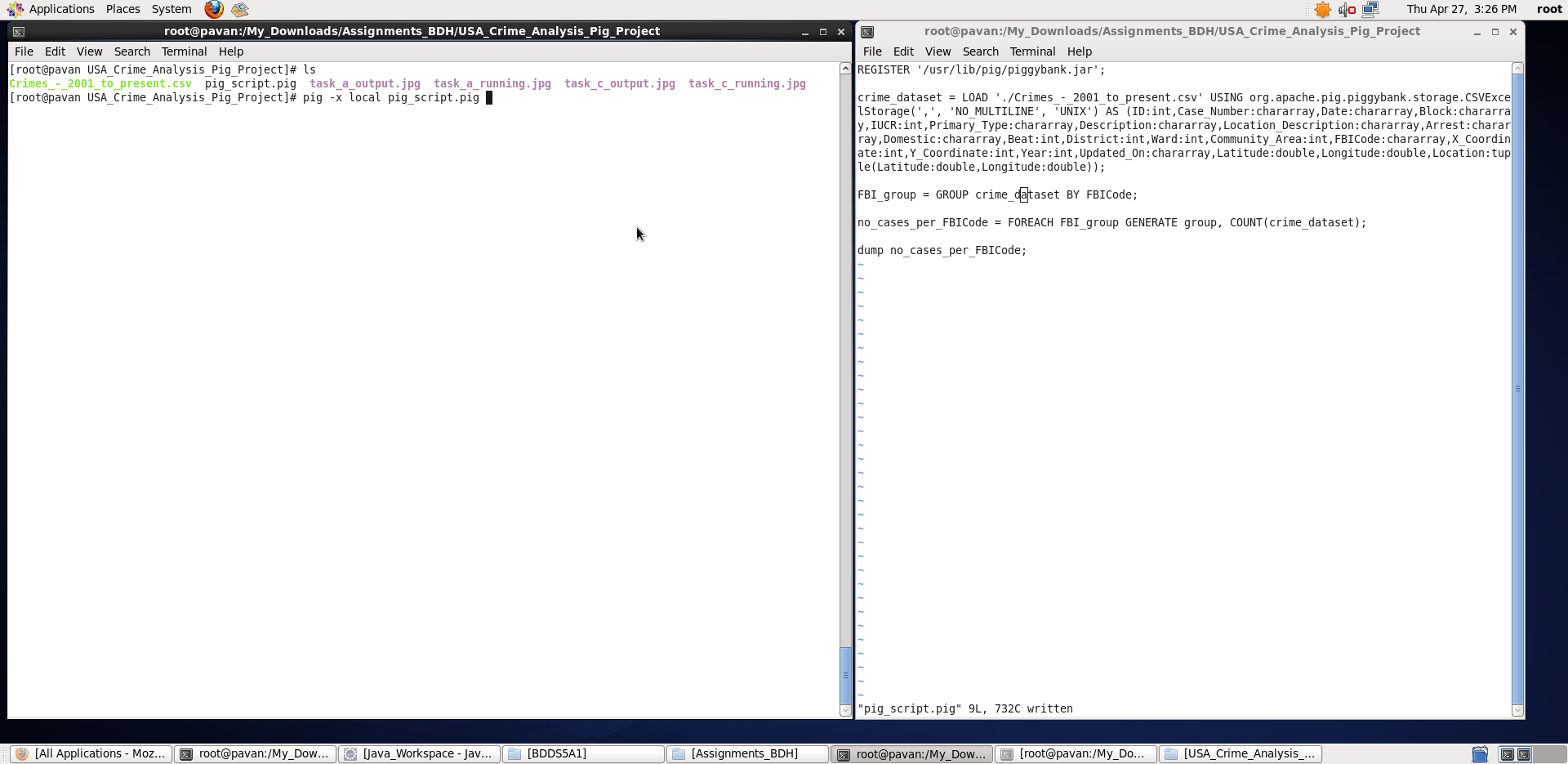
dump no\_cases\_per\_FBICode;

**Explanation :**

* Loaded data from Crimes\_-\_2001\_to\_present.csv into “crime\_dataset” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question
* Then, all records are grouped on basis of FBICode.
* For each FBICode (group-key) , we calculated the values – number of cases investigated under that particular FBICode using COUNT function.
* **Outputting only “FBICode and number of cases investigated under that FBICode” as per Question using dump command.**

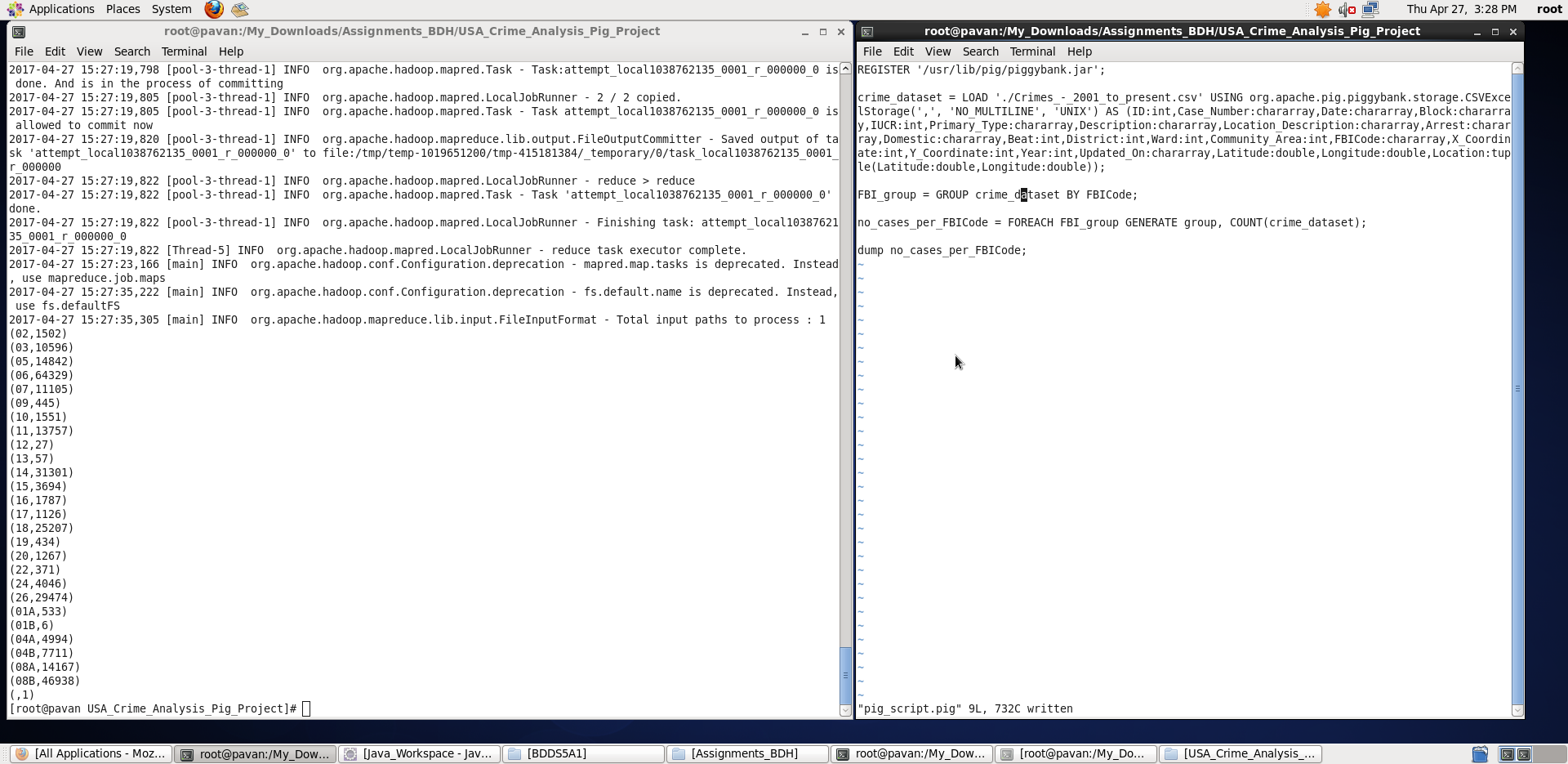
Screenshots

Running **Pig Latin Script in Local Mode** inside /My\_Downloads/Assignments\_BDH/USA\_Crime\_Analysis\_Pig\_Project/ directory



Output displayed: **<FBICode> <number\_of\_cases>**

**Note : We can store the output if required in a file using “store” command.**



P.S: There is one invalid (last) entry (only ID = 9 it has) in the dataset with no FBICode – the last ( ,1) corresponds to that which can be neglected .

**b) Write a pig program to calculate the number of cases investigated under FBI code 32.**

**Pig Latin Script:**

REGISTER '/usr/lib/pig/piggybank.jar';

crime\_dataset = LOAD './Crimes\_-\_2001\_to\_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'UNIX') AS (ID:int,Case\_Number:chararray,Date:chararray,Block:chararray,IUCR:int,Primary\_Type:chararray,Description:chararray,Location\_Description:chararray,Arrest:chararray,Domestic:chararray,Beat:int,District:int,Ward:int,Community\_Area:int,FBICode:chararray,X\_Coordinate:int,Y\_Coordinate:int,Year:int,Updated\_On:chararray,Latitude:double,Longitude:double,Location:tuple(Latitude:double,Longitude:double));

FBI\_code\_32 = FILTER crime\_dataset BY FBICode=='32';

FBI\_code\_32\_group = GROUP FBI\_code\_32 ALL;

no\_cases\_under\_FBICode\_32 = FOREACH FBI\_code\_32\_group GENERATE group, COUNT(FBI\_code\_32);

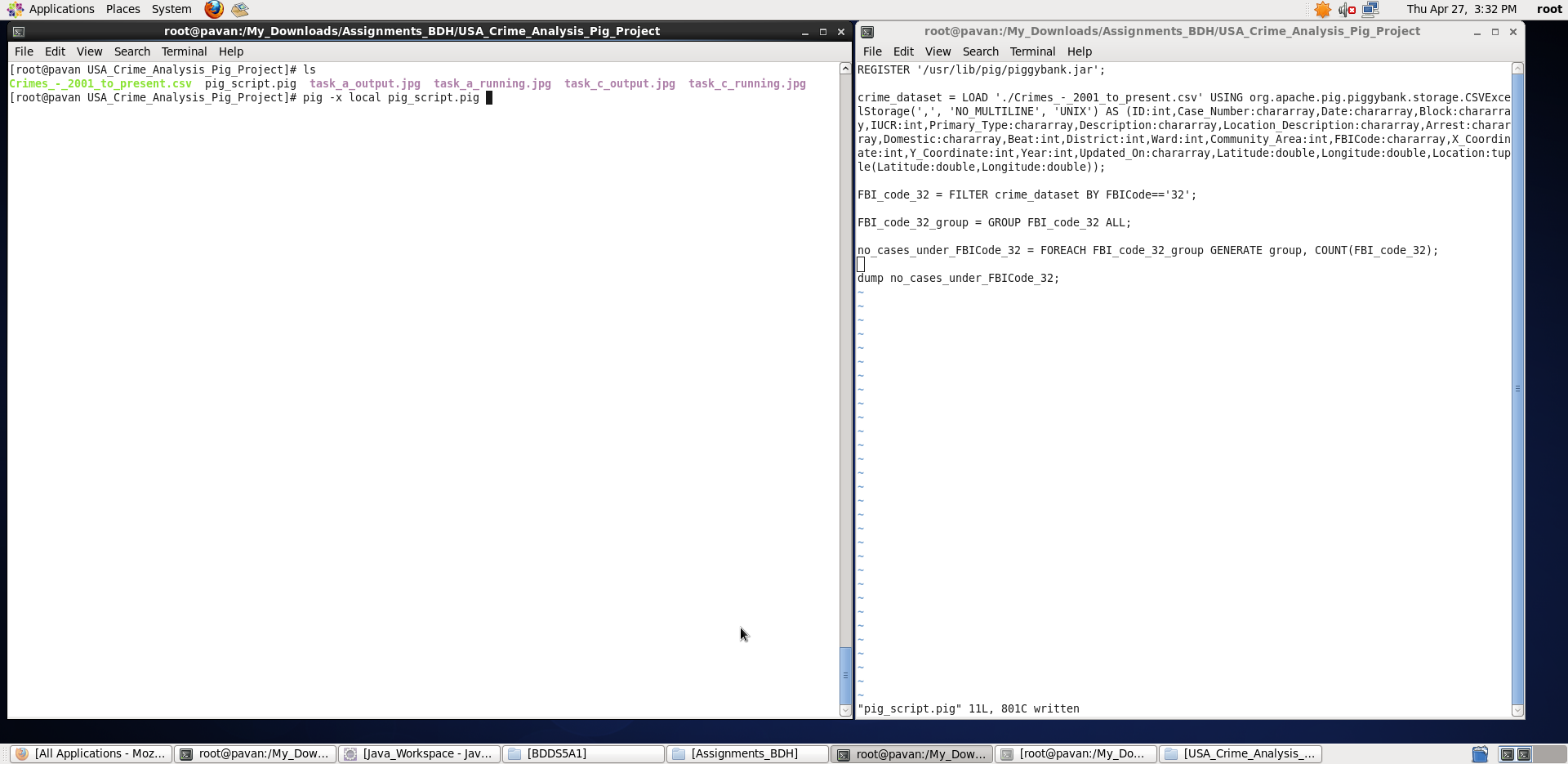
dump no\_cases\_under\_FBICode\_32;

**Explanation :**

* Loaded data from Crimes\_-\_2001\_to\_present.csv into “crime\_dataset” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question
* Then, all records are filtered on the basis of expression FBICode==’32’.
* Now, for counting the number of cases investigated under FBICode 32, we group all the above filtered records and then using COUNT function, we calculate the number of cases investigated under FBICode 32.
* **Outputting only “FBICode 32 and the number of cases investigated under FBICode 32” as per Question using dump command.**

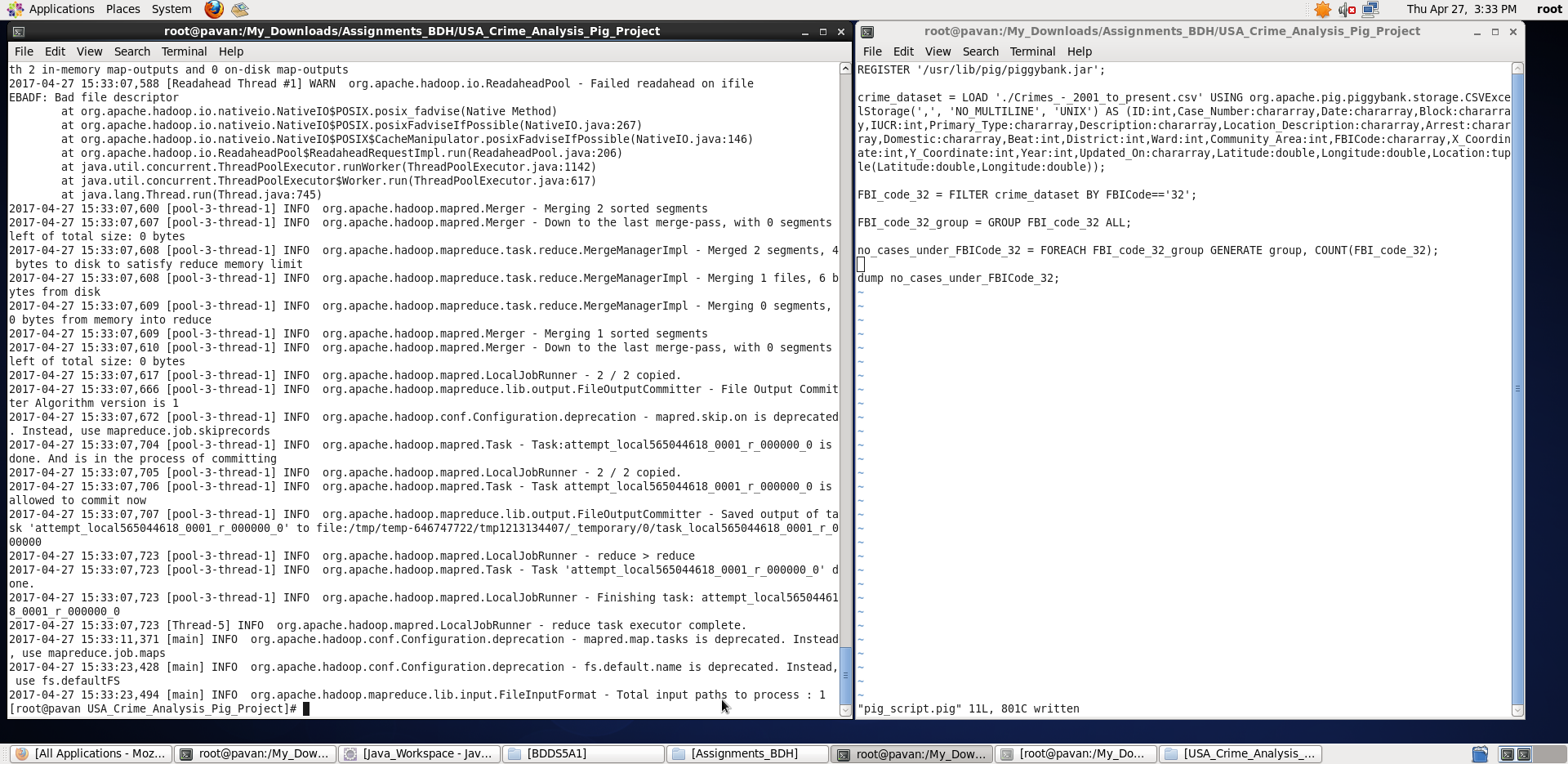
Screenshots

Running **Pig Latin Script in Local Mode** inside /My\_Downloads/Assignments\_BDH/USA\_Crime\_Analysis\_Pig\_Project/ directory



Output displayed: **<FBICode (i.e. 32)> <** **the number of cases investigated under FBICode 32>**

**Note: We can store the output if required in a file using “store” command.**



***Observe, there is no output indicating there are no cases investigated under FBICode 32.***

c) **Write a pig program to calculate the number of arrests in theft district wise**

**Pig Latin Script:**

REGISTER '/usr/lib/pig/piggybank.jar';

crime\_dataset = LOAD './Crimes\_-\_2001\_to\_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'UNIX') AS (ID:int,Case\_Number:chararray,Date:chararray,Block:chararray,IUCR:int,Primary\_Type:chararray,Description:chararray,Location\_Description:chararray,Arrest:chararray,Domestic:chararray,Beat:int,District:int,Ward:int,Community\_Area:int,FBICode:chararray,X\_Coordinate:int,Y\_Coordinate:int,Year:int,Updated\_On:chararray,Latitude:double,Longitude:double,Location:tuple(Latitude:double,Longitude:double));

Theft\_Arrest\_cases = FILTER crime\_dataset BY (Primary\_Type=='THEFT') AND (Arrest=='true');

district\_wise\_theft\_arrest\_group = GROUP Theft\_Arrest\_cases BY District;

no\_arrests\_theft\_district\_wise = FOREACH district\_wise\_theft\_arrest\_group GENERATE group, COUNT(Theft\_Arrest\_cases);

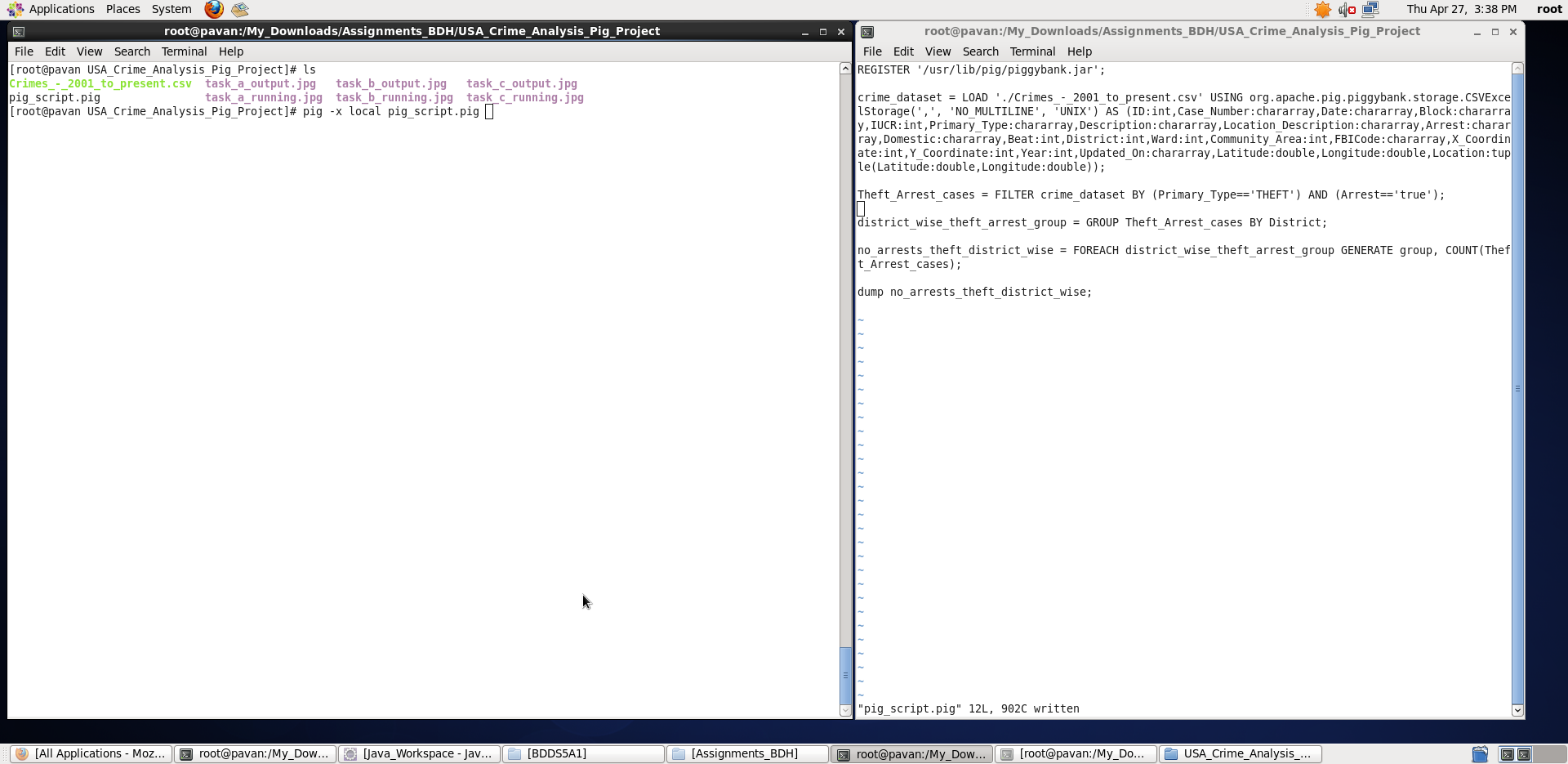
dump no\_arrests\_theft\_district\_wise;

**Explanation :**

* Loaded data from Crimes\_-\_2001\_to\_present.csv into “crime\_dataset” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question
* All the records are filtered based on the conditions – Primary\_Type == THEFT and Arrest==’true’. This gives us all theft cases in which arrest has been happened.
* Now, grouping the theft\_arrest\_cases with respect to District.
* Then, foreach key in the group, we calculated the number of values – theft\_arrest\_cases that happened corresponding to that District using COUNT function.
* **Outputting only “District and corresponding number of theft arrest cases” as per Question using dump command.**

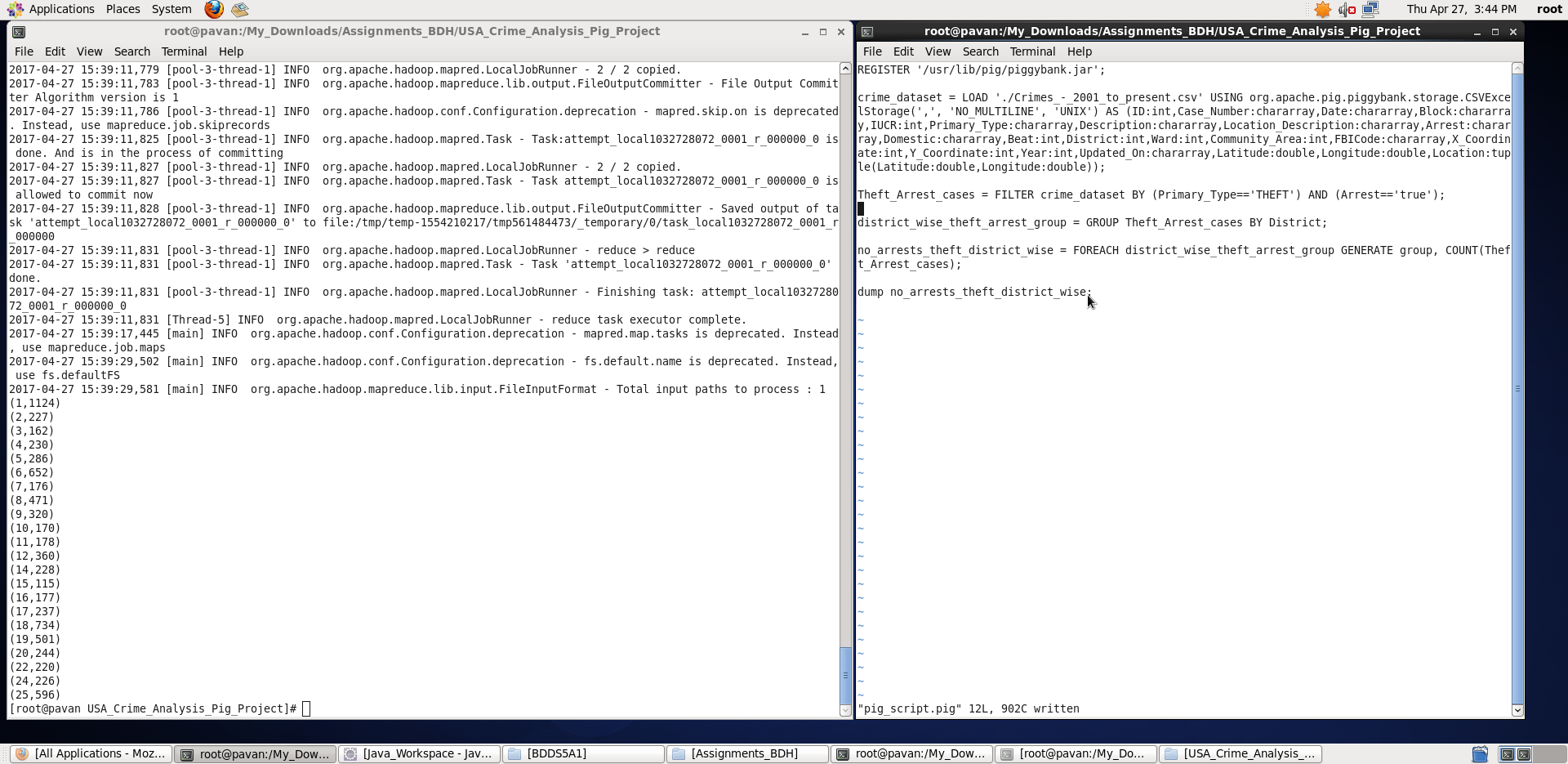
Screenshots

Running **Pig Latin Script in Local Mode** inside /My\_Downloads/Assignments\_BDH/USA\_Crime\_Analysis\_Pig\_project/ directory



Output displayed: **<District> <number of theft arrest cases in that particular district>**

**Note : We can store the output if required in a file using “store” command.**



d) **Write a pig program to calculate the number of arrests done between october 2014 and october 2015.**

**Pig Latin Script:**

REGISTER '/usr/lib/pig/piggybank.jar';

crime\_dataset = LOAD 'Crimes\_-\_2001\_to\_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'UNIX') AS (ID:int,Case\_Number:chararray,Date:chararray,Block:chararray,IUCR:int,Primary\_Type:chararray,Description:chararray,Location\_Description:chararray,Arrest:chararray,Domestic:chararray,Beat:int,District:int,Ward:int,Community\_Area:int,FBICode:chararray,X\_Coordinate:int,Y\_Coordinate:int,Year:int,Updated\_On:chararray,Latitude:double,Longitude:double,Location:tuple(Latitude:double,Longitude:double));

crime\_dataset\_proper\_date\_arrests = foreach crime\_dataset generate ToDate(Date, 'MM/dd/yyyy HH:mm:ss a', 'US/Pacific') AS Date\_Time, Arrest, ID;

Filter\_Arrest\_cases = FILTER crime\_dataset\_proper\_date\_arrests BY (Arrest=='true');

Filter\_Arrests\_by\_Date = filter Filter\_Arrest\_cases By (DaysBetween(Date\_Time, (datetime)ToDate('09/30/2014', 'MM/dd/yyyy')) >(long)0) AND (DaysBetween(Date\_Time, (datetime)ToDate('10/01/2015', 'MM/dd/yyyy')) <(long)0);

grouping\_to\_count = GROUP Filter\_Arrests\_by\_Date ALL;

num\_arrests\_btw\_Oct\_14\_Oct\_15 = FOREACH grouping\_to\_count GENERATE COUNT(Filter\_Arrests\_by\_Date);

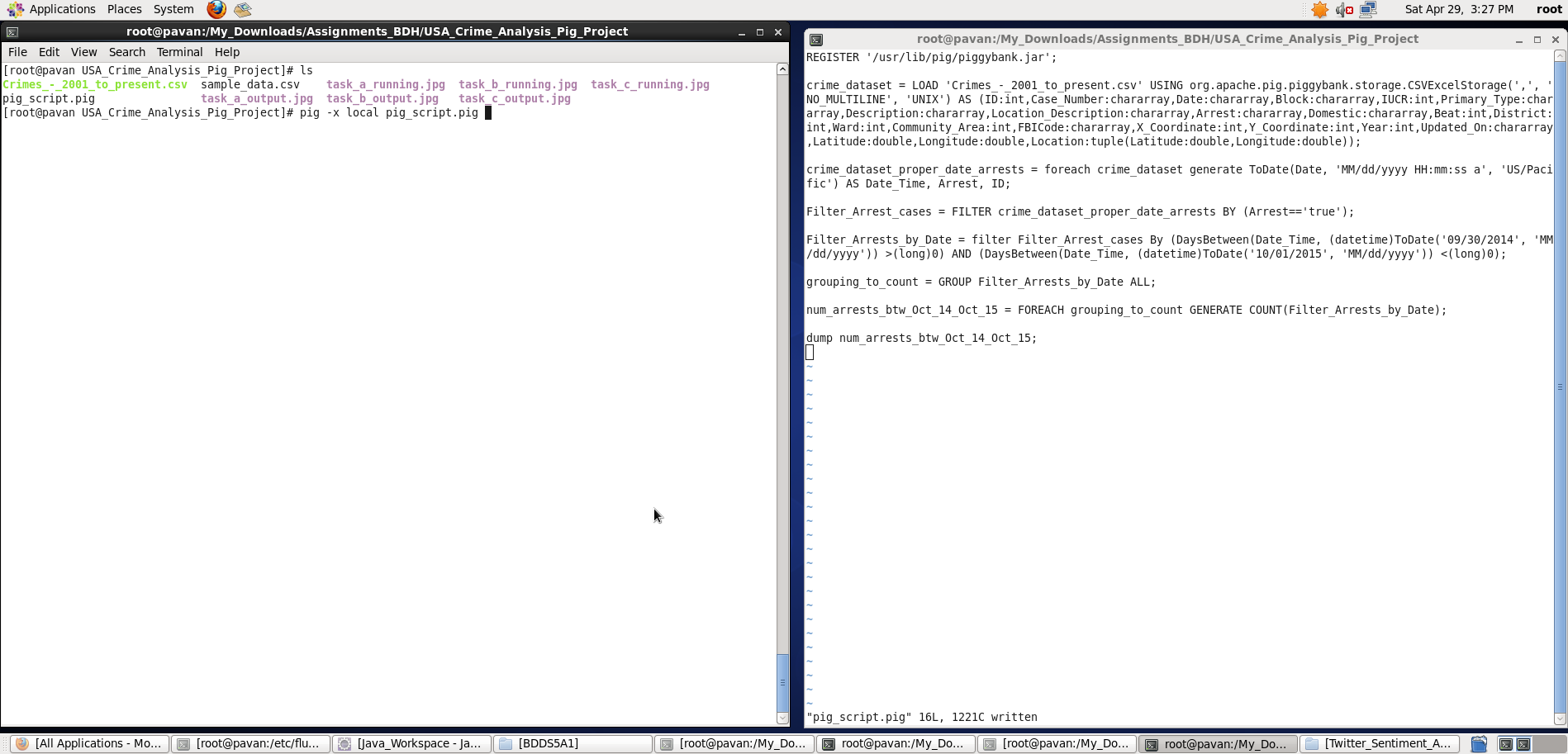
dump num\_arrests\_btw\_Oct\_14\_Oct\_15;

**Explanation :**

* Loaded data from Crimes\_-\_2001\_to\_present.csv into “crime\_dataset” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question
* For performing operations on Date, a DateTime object is required. So, with the help of ToDate() function we generated a DateTime object from the input crime\_dataset. Here, since dataset is from USA\_Crimes, I have specified timezone to be “US/Pacific” and input date format is ‘MM/dd/yyyy HH:mm:ss a’ observed from dataset.
* Now, we first filtered the records based on Arrest==’true’ condition.
* Then with the help of DaysBetween() function which accepts two date-time objects(dt1 and dt2) and calculates the number of days(**positive days if dt1 occurred after dt2** and **negative days if dt1 occurred before dt2** ) between the two given date-time objects, I have filtered the arrest cases such that they are done between October 2014 and October 2015 i.e. our input Date should be after '09/30/2014' and before ‘10/01/2015’.
* Now, for counting the number of arrests happened, we group all the above filtered records and then using COUNT function, we calculate the number of arrests done in between October 2014 and October 2015.
* **Outputting only the “number of arrest cases” as per Question using dump command.**

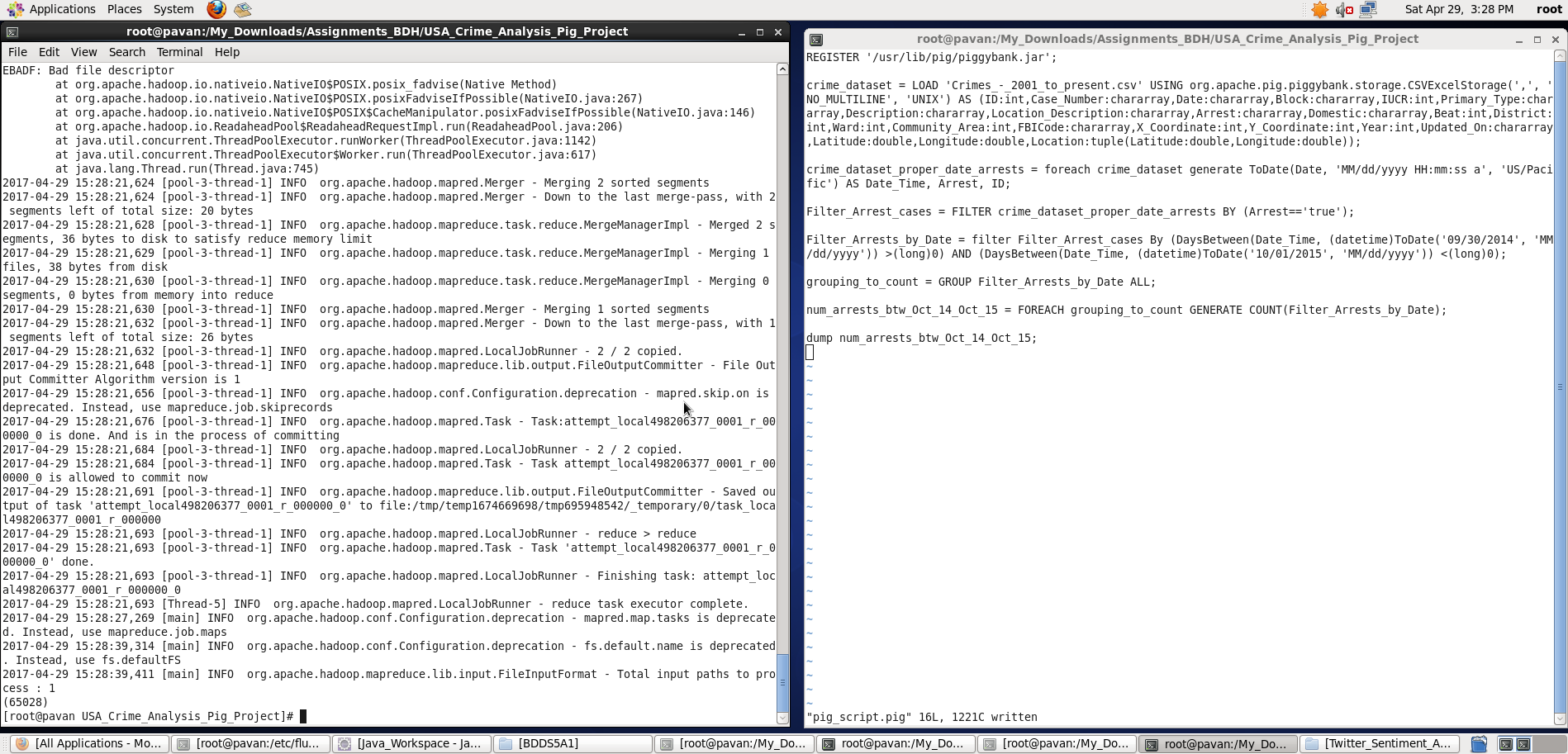
Screenshots

Running **Pig Latin Script in Local Mode** inside /My\_Downloads/Assignments\_BDH/USA\_Crime\_Analysis\_Pig\_Project/ directory



Output displayed: **<num\_of\_arrests\_btw\_Oct\_214\_Oct\_2015>**

**Note : We can store the output if required in a file using “store” command.**



**Thus, with the help of Pig Latin script, I have displayed the results for all the 4 tasks.**