# **PROJECT 1.1**

# **USA Crime Analysis**

**Dataset :**

[**https://drive.google.com/file/d/0B1QaXx7tpw3SaUJHOHBZclBXWG8/view?usp=sharing**](https://drive.google.com/file/d/0B1QaXx7tpw3SaUJHOHBZclBXWG8/view?usp=sharing)

**About DataSet:**

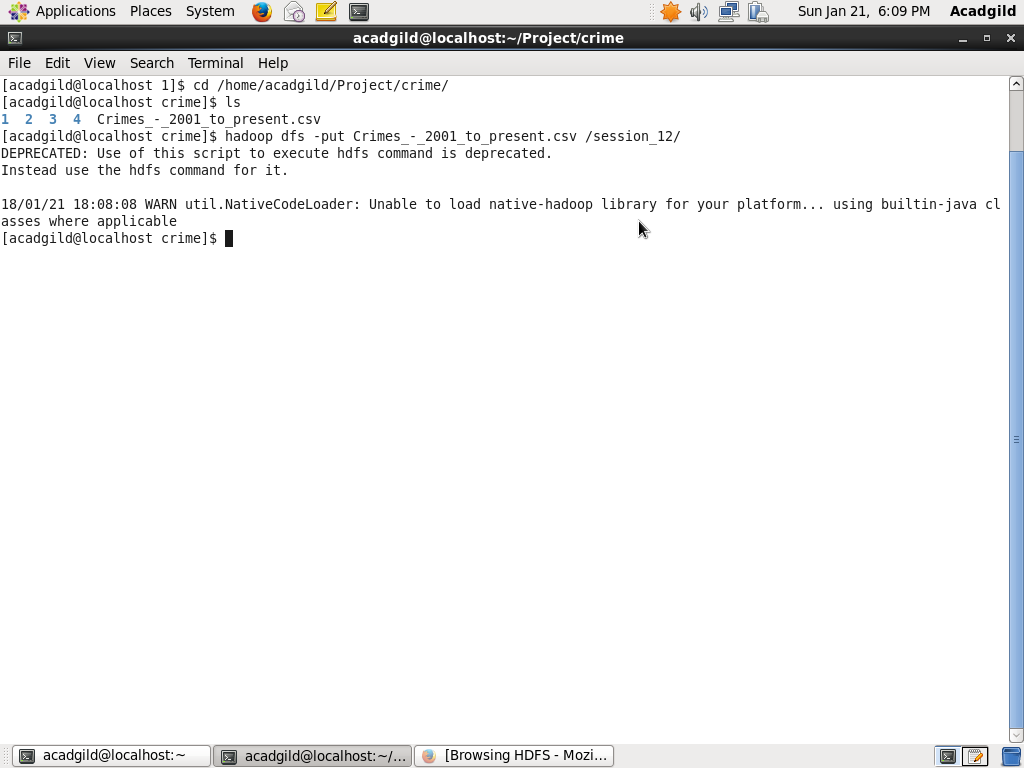
This dataset contains attributes related to crimes taking place in various areas like type of crime, FBI code related to that criminal case, arrest frequency, location of crime etc.

**Technology Used:**

* HDFS
* PIG

**Importing dataset to HDFS:**

Hadoop dfs -put Crimes\_-\_2001\_to\_present.csv /session\_12/

****

**Problem Statement 1 :**

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

**Code:**

**Crime\_1.pig**

A = load '/session\_12/Crimes\_-\_2001\_to\_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX');

B= FOREACH A GENERATE (chararray)$1 as case\_number, (chararray)$14 as FBI\_code;

C= GROUP B BY FBI\_code;

D= FOREACH C GENERATE group,COUNT(B.case\_number) AS cnt;

FILTER\_D = FILTER D BY group is not null;

E= ORDER FILTER\_D BY cnt;

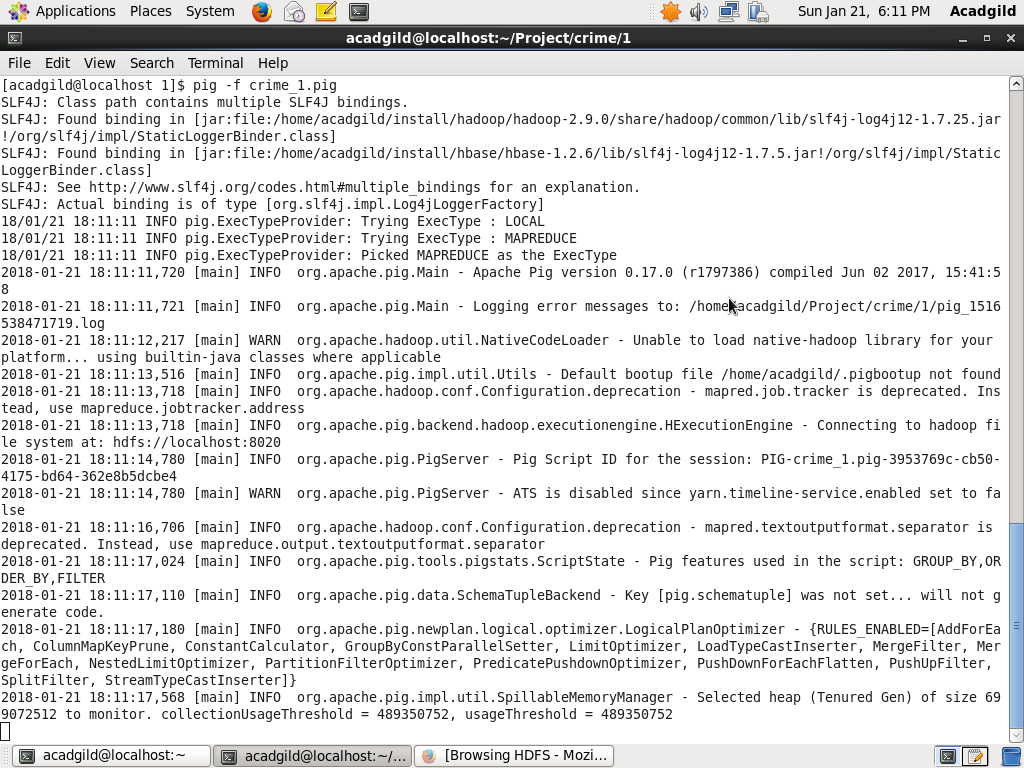
STORE E into '/session\_12/1/soln' USING PigStorage(',');

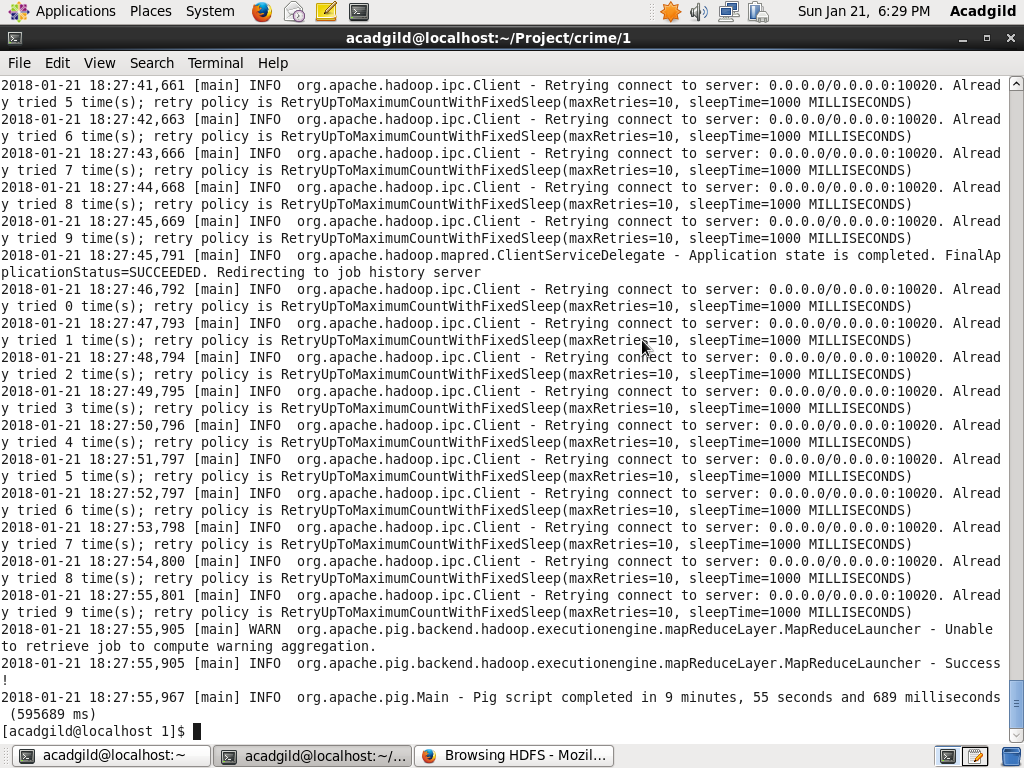
**Code Workflow:**

We are loading the dataset from HDFS to PIG table and generating schema for the fields which are required to solve the problem.To calculate number of cases firstly, we will group the data on the basis of FBI\_code(Table **C**). Then, foreach FBI\_code we will count the number of cases(Table **D**). Then, filtering out the invalid values(Table **Filter\_D**).

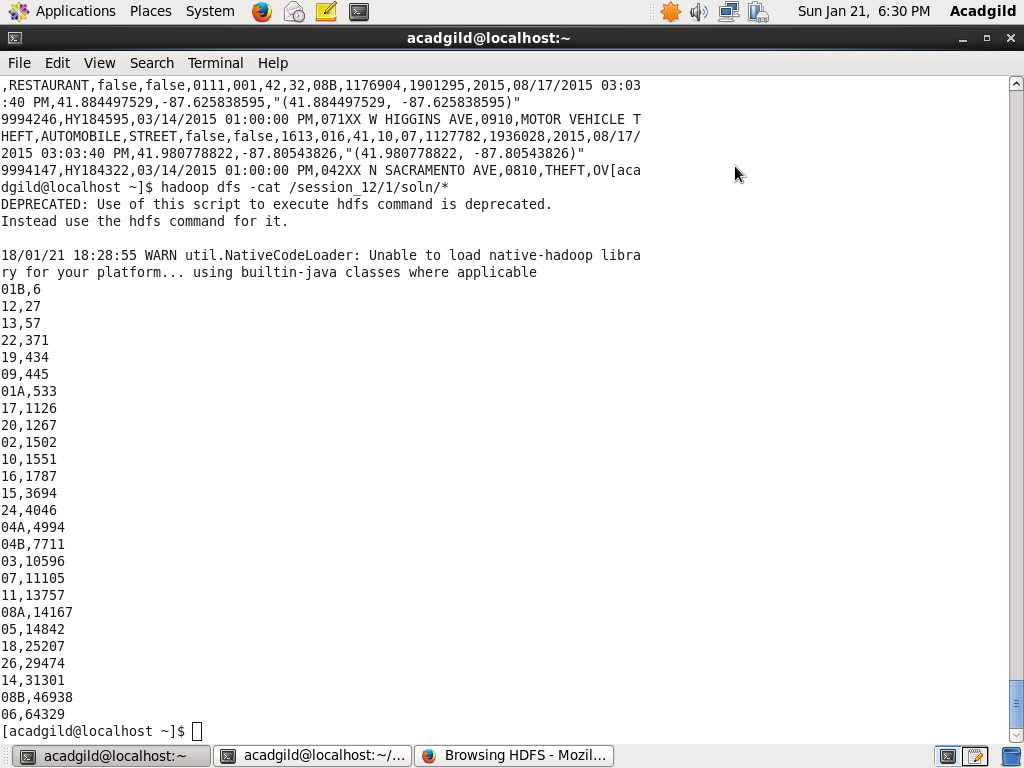
**Execution of code:**

Pig -f crime\_1.pig





**OUTPUT:**



**Problem Statement 2 :**

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

**Code:**

**Crime\_2.pig**

A = load '/session\_12/Crimes\_-\_2001\_to\_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX');

B= FOREACH A GENERATE (chararray)$1 as case\_number,(chararray)$5 as primary\_type, (chararray)$14 as FBI\_code;

FILTER\_B = FILTER B BY FBI\_code MATCHES '.\*32.\*';

C= GROUP FILTER\_B BY FBI\_code;

D= FOREACH C GENERATE group,COUNT(FILTER\_B.case\_number) AS cnt;

STORE D into '/session\_12/2/soln' USING PigStorage(',');

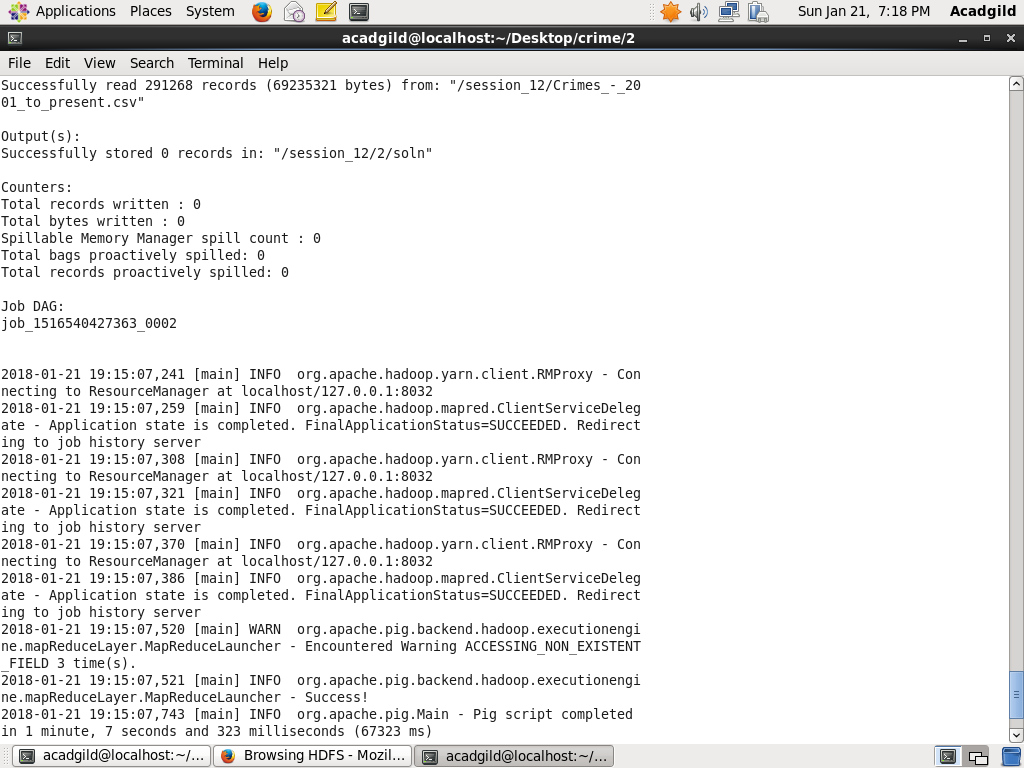
**Code Workflow:**

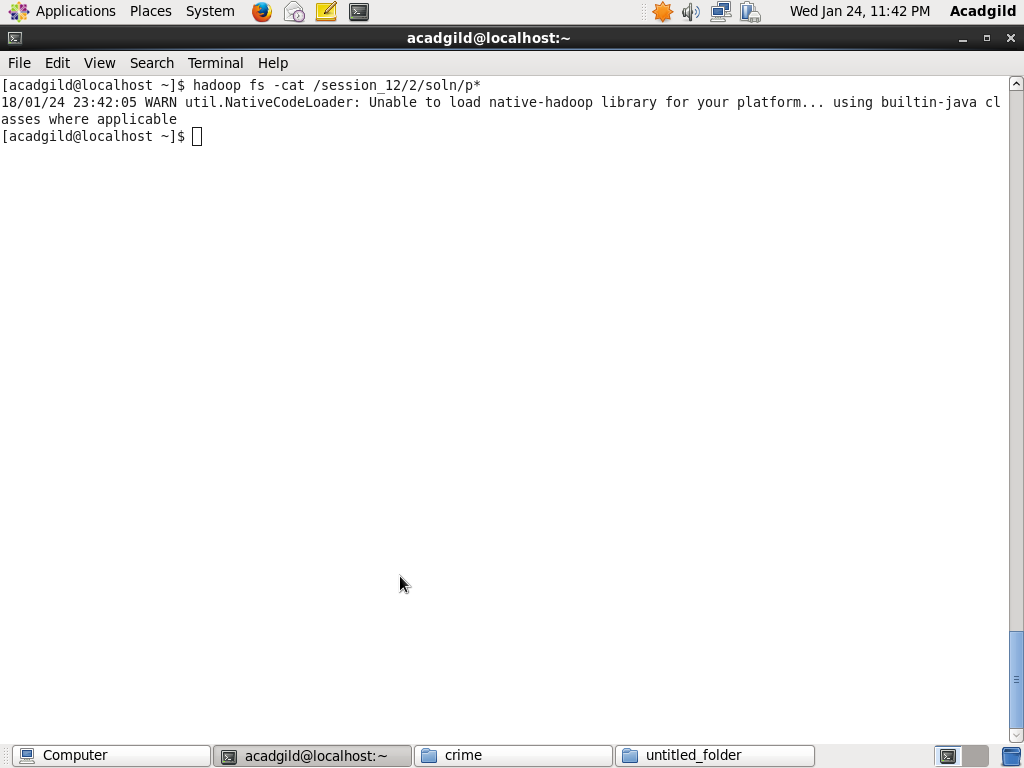
We are loading the dataset from HDFS to PIG table and generating schema for the fields which are required to solve the problem.To calculate number of cases firstly, we will filter the data with FBI\_Code =32 (Table **FILTER\_B**). Now, we will group by FBI\_code (Table **C**).Last step we will count the number of cases.

**Execution of code:**

Pig -f crime\_2.pig





OUTPUT:

**Problem Statement 3 :**

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

**Code:**

**Crime\_3.pig**

A = load '/session\_12/Crimes\_-\_2001\_to\_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX');

B= FOREACH A GENERATE (chararray)$8 as arrest, (chararray)$5 as primary, (chararray)$11 as district;

C = FILTER B BY primary == 'THEFT';

FILTER\_C = FILTER C BY arrest == 'true';

D= GROUP FILTER\_C BY district;

E= FOREACH D GENERATE group,COUNT(FILTER\_C) AS cnt;

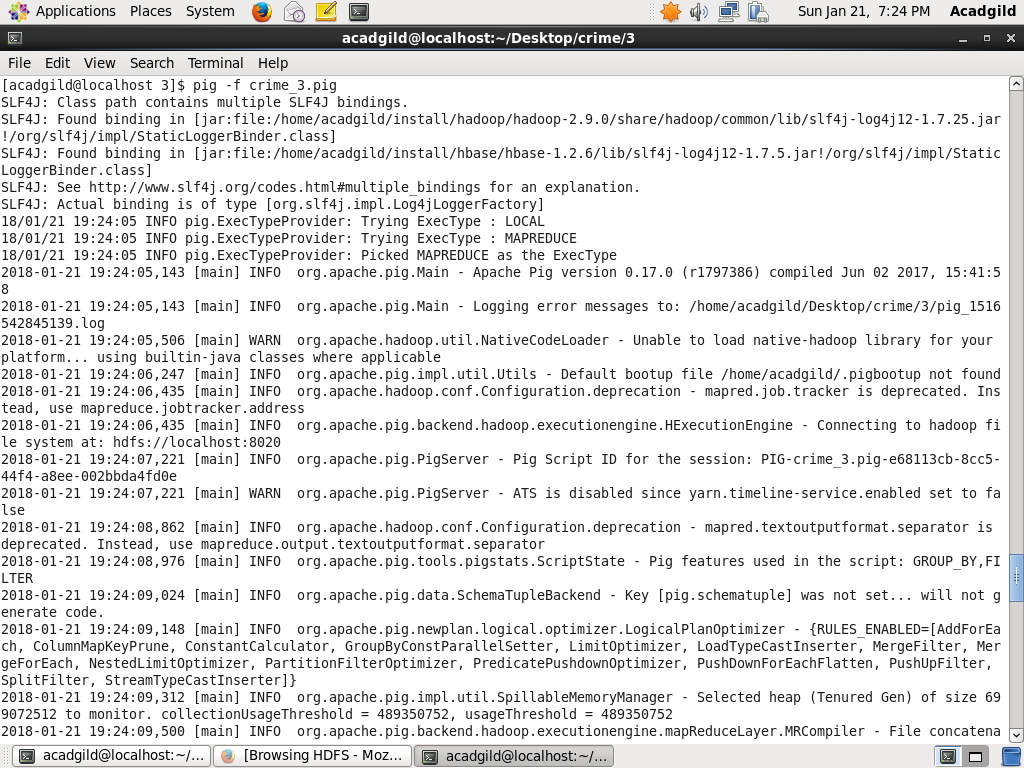
STORE E into '/session\_12/3/soln' USING PigStorage(',');

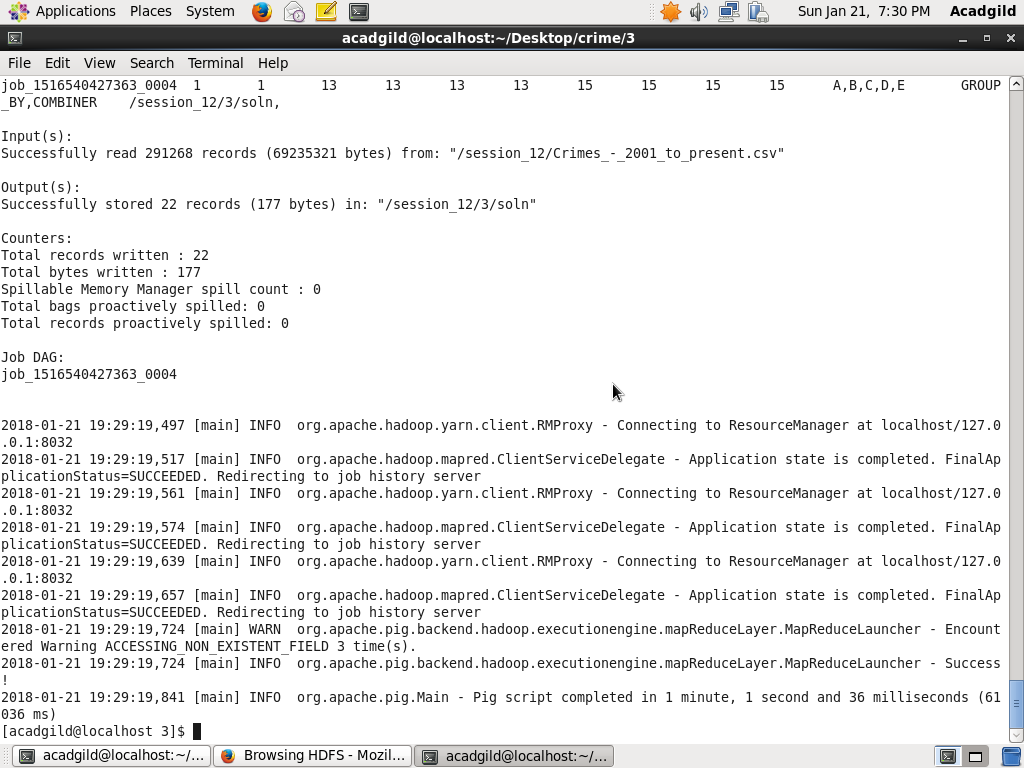
**Code Workflow:**

We are loading the dataset from HDFS to PIG table and generating schema for the fields which are required to solve the problem.To calculate number of cases firstly, filter data by primary =’theft’(Table **C**). Then, we need to find the records which satisfy the condition arrest=true (Table **Filter\_C**).Then grouping records by district and getting the district wise count of arrest(Table **E**).

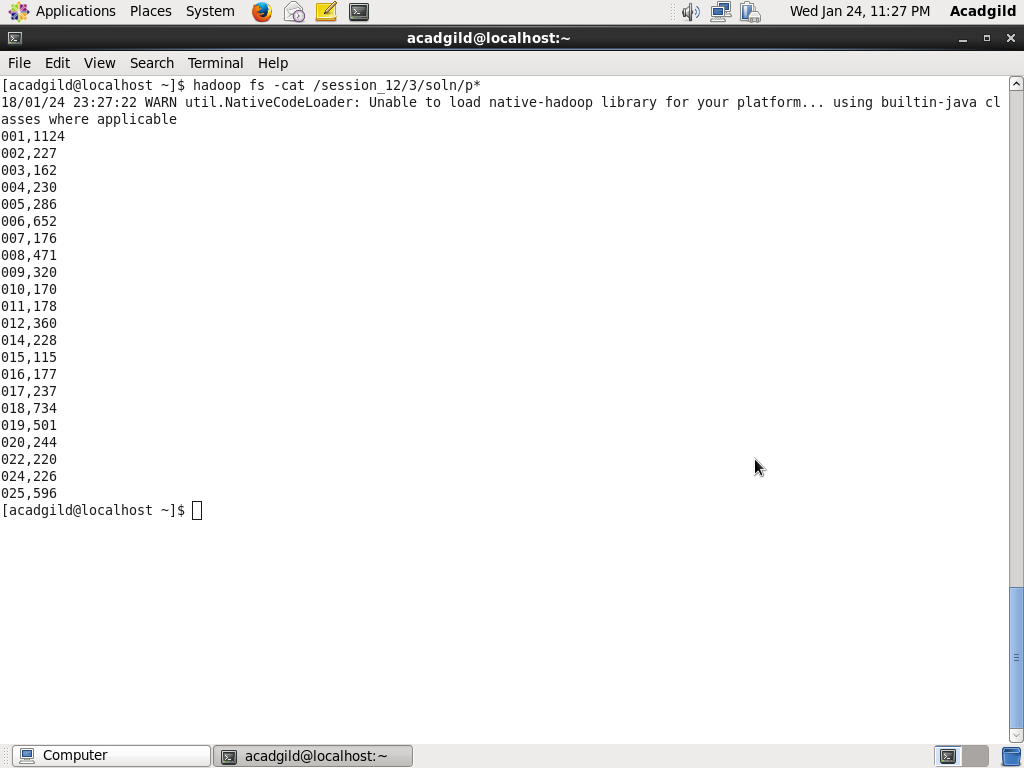
**Execution of code:**

Pig -f crime\_3.pig





**OUTPUT:**



**Problem Statement 4 :**

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

**Code:**

**Crime\_4.pig**

A = load '/session\_12/Crimes\_-\_2001\_to\_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX');

B= FOREACH A GENERATE (chararray)$8 as arrest, (chararray)$17 as year,(chararray)$18 as updated;

FILTER\_B = FILTER B BY year == '2014' or year == '2015';

C= FOREACH FILTER\_B GENERATE arrest,year,ToDate(updated,'MM/dd/yyyy hh:mm:ss aa') as updated;

FILTER\_C = FILTER C BY ((updated >= ToDate('10/01/2014','MM/dd/yyyy')) AND (updated <= ToDate('10/31/2015','MM/dd/yyyy')));

FILTER\_TRUE = FILTER FILTER\_C BY arrest == 'true';

D = GROUP FILTER\_TRUE ALL;

E = FOREACH D GENERATE COUNT(FILTER\_TRUE.arrest);

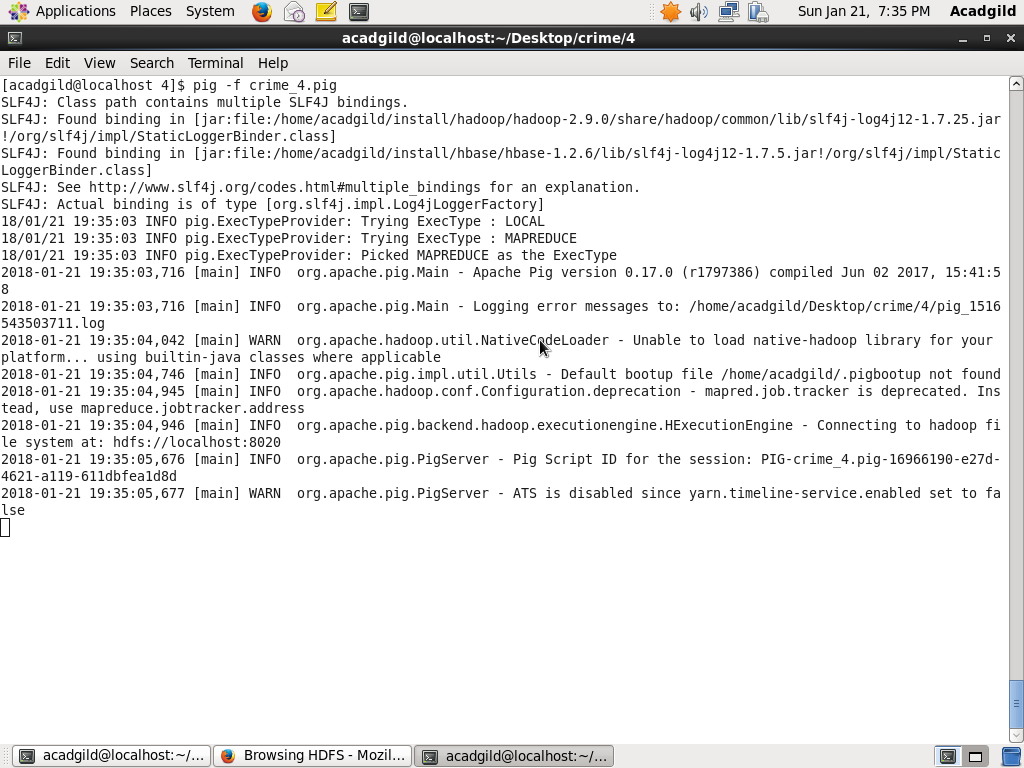
STORE E into '/session\_12/4/soln' USING PigStorage(',');

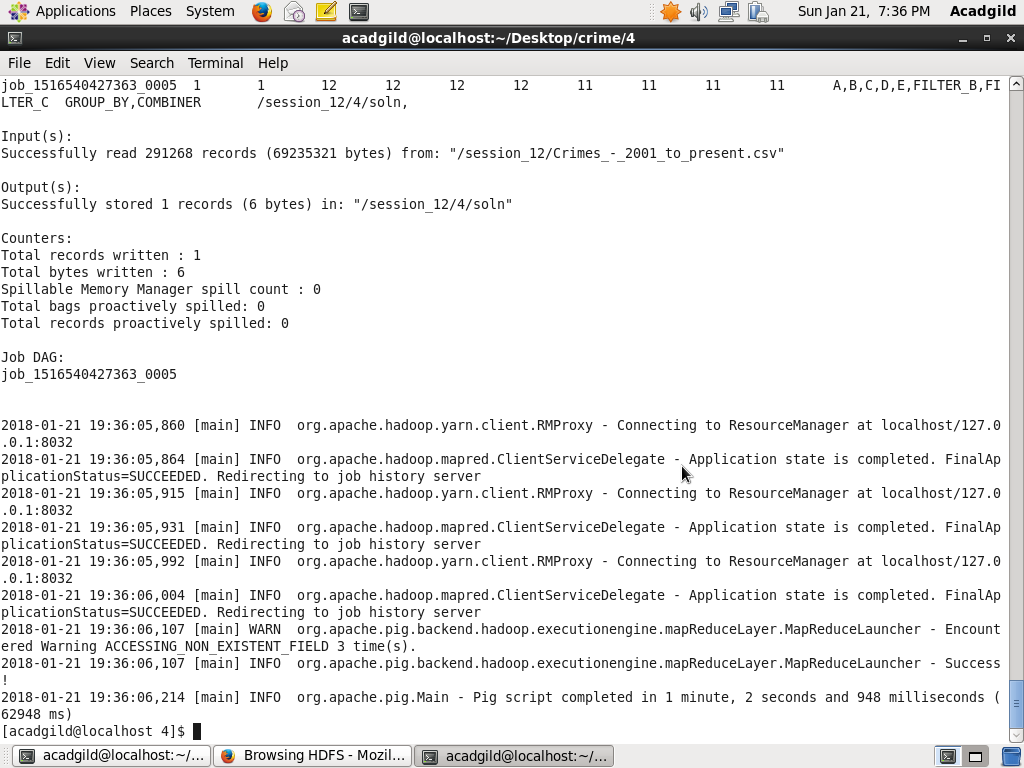
**Code Workflow:**

We are loading the dataset from HDFS to PIG table and generating schema for the fields which are required to solve the problem.To calculate number of cases firstly,we will filter records with year = 2014 or year = 2015(Table **FILTER\_B**).To find the records in which arrest were done during oct 2014 to oct 2015 we will firstly convert date from string type to DATE type and then filtering data(table **FILTER\_C**). Then filtering data in which arrest were done and counting the number of cases.

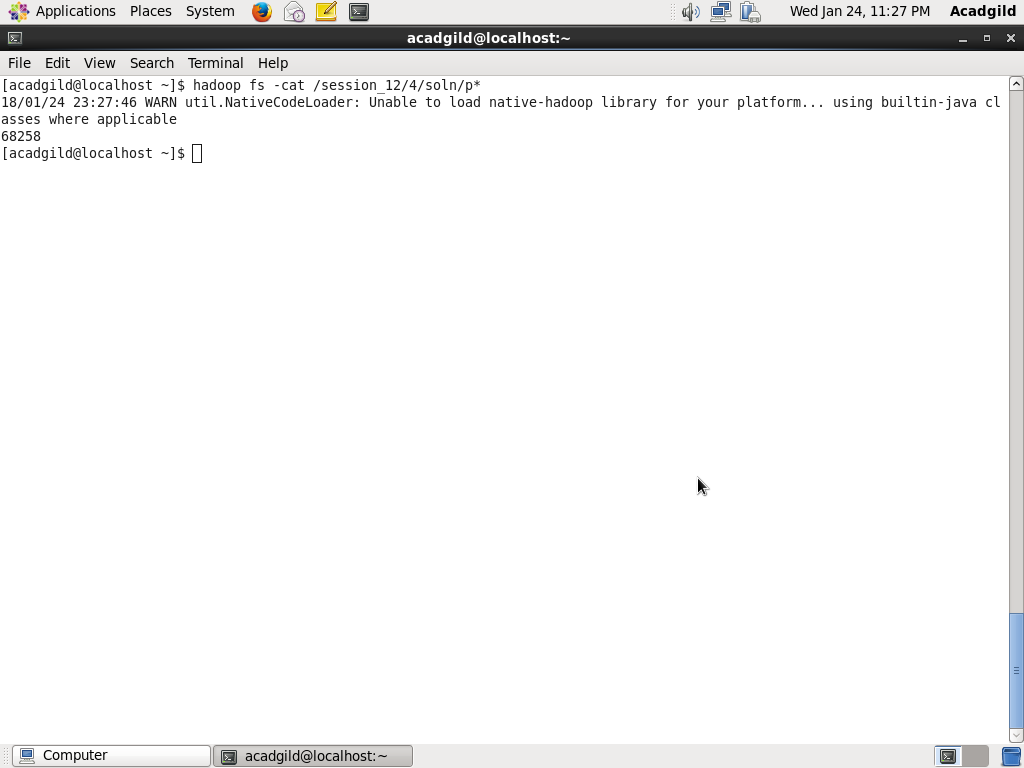
**Execution of code:**

Pig -f crime\_4.pig





**OUTPUT:**

****