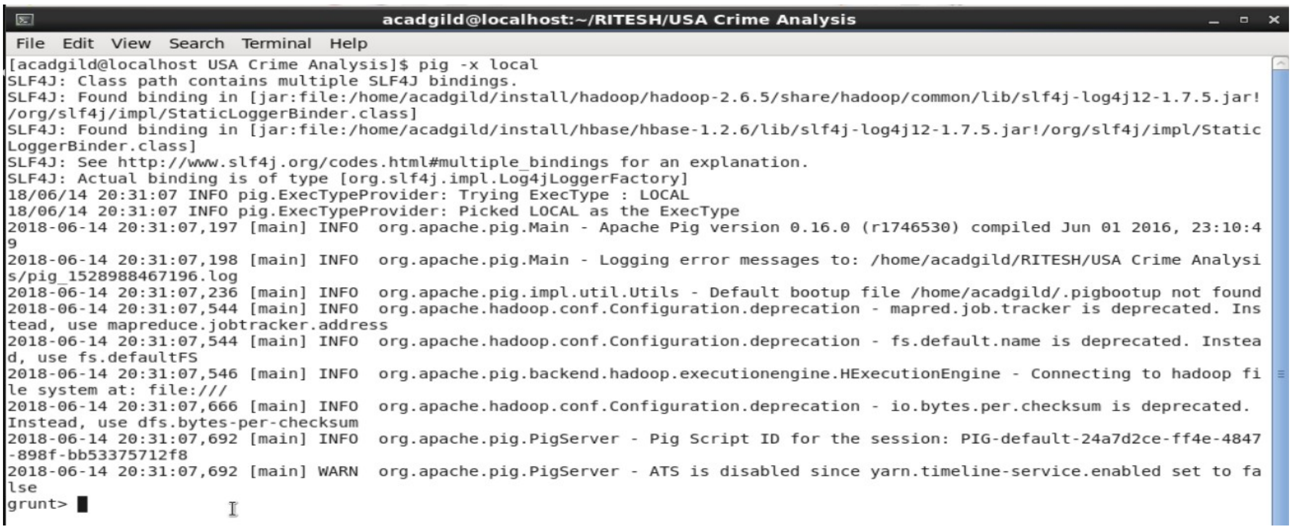
Project 1.1 - USA Crime Analysis

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

1. Write a MapReduce/Pig program to calculate the number of cases investigated under each FBI code
2. Write a MapReduce/Pig program to calculate the number of cases investigated under FBI code 32.
3. Write a MapReduce/Pig program to calculate the number of arrests in theft district wise.
4. Write a MapReduce/Pig program to calculate the number of arrests done between October 2014 and October 2015.

Common Steps for the 4 problems -

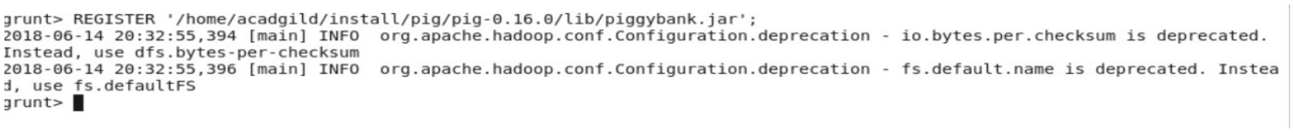
- Started the pig shell
  pig -x local



acadgild@localhost:~/RITESH/USA Crime Analysis
File Edit View Search Terminal Help

```
[acadgild@localhost USA Crime Analysis]$ pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!
/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/06/14 20:31:07 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/06/14 20:31:07 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-06-14 20:31:07,197 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:4
9
2018-06-14 20:31:07,198 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/RITESH/USA Crime Analysi
s/pig_1528988467196.log
2018-06-14 20:31:07,236 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-06-14 20:31:07,544 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2018-06-14 20:31:07,544 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-14 20:31:07,546 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
le system at: file:///
2018-06-14 20:31:07,666 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-14 20:31:07,692 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-24a7d2ce-ff4e-4847
-898f-bb53375712f8
2018-06-14 20:31:07,692 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to fa
lse
grunt> █
```

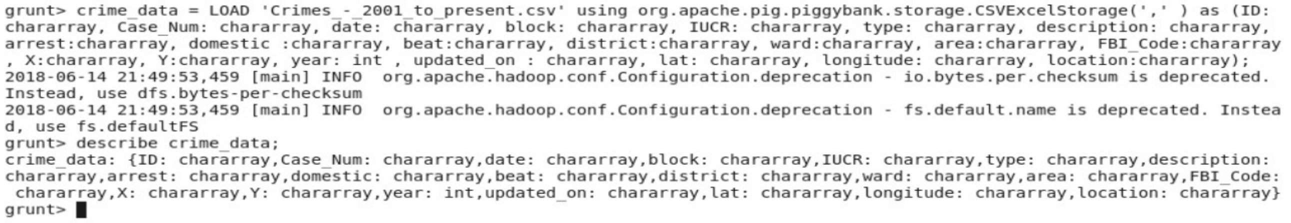
- External piggybank jar was added to pig shell for handling csv data.
  REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';



```
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
2018-06-14 20:32:55,394 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-14 20:32:55,396 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> █
```

- Loaded the csv data into a relation

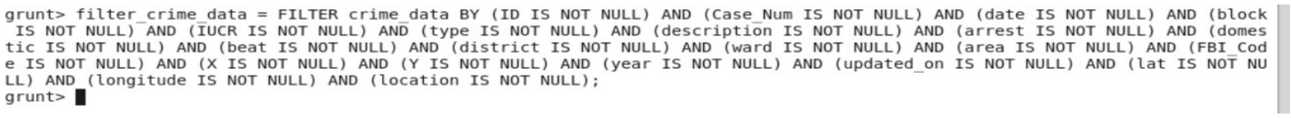
  crime_data = LOAD 'Crimes_-_2001_to_present.csv' using org.apache.pig.piggybank.storage.CSVExcelStorage(',') as (ID: chararray, Case_Num: chararray, date: chararray, block: chararray, IUCR: chararray, type: chararray, description: chararray, arrest:chararray, domestic :chararray, beat:chararray, district:chararray, ward:chararray, area:chararray, FBI_Code:chararray, X:chararray, Y:chararray, year: int , updated_on : chararray, lat: chararray, longitude: chararray, location:chararray);



```
grunt> crime_data = LOAD 'Crimes_-_2001_to_present.csv' using org.apache.pig.piggybank.storage.CSVExcelStorage(',' ) as (ID:
chararray, Case_Num: chararray, date: chararray, block: chararray, IUCR: chararray, type: chararray, description: chararray,
arrest:chararray, domestic :chararray, beat:chararray, district:chararray, ward:chararray, area:chararray, FBI_Code:chararray
, X:chararray, Y:chararray, year: int , updated_on : chararray, lat: chararray, longitude: chararray, location:chararray);
2018-06-14 21:49:53,459 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-14 21:49:53,459 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> describe crime_data;
crime_data: {ID: chararray,Case_Num: chararray,date: chararray,block: chararray,IUCR: chararray,type: chararray,description:
chararray,arrest: chararray,domestic: chararray,beat: chararray,district: chararray,ward: chararray,area: chararray,FBI_Code:
 chararray,X: chararray,Y: chararray,year: int,updated_on: chararray,lat: chararray,longitude: chararray,location: chararray}
grunt> █
```

- Filtered the crime data wherever in any tuple's column there is null and stored in an another relation.

**filter_crime_data = FILTER crime_data BY (ID IS NOT NULL) AND (Case_Num IS NOT NULL) AND (date IS NOT NULL) AND (block IS NOT NULL) AND (IUCR IS NOT NULL) AND (type IS NOT NULL) AND (description IS NOT NULL) AND (arrest IS NOT NULL) AND (domestic IS NOT NULL) AND (beat IS NOT NULL) AND (district IS NOT NULL) AND (ward IS NOT NULL) AND (area IS NOT NULL) AND (FBI_Code IS NOT NULL) AND (X IS NOT NULL) AND (Y IS NOT NULL) AND (year IS NOT NULL) AND (updated_on IS NOT NULL) AND (lat IS NOT NULL) AND (longitude IS NOT NULL) AND (location IS NOT NULL);
noCases_fbiCode;**



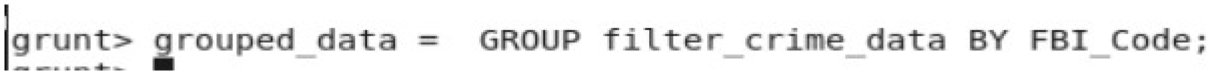
```
grunt> filter_crime_data = FILTER crime_data BY (ID IS NOT NULL) AND (Case_Num IS NOT NULL) AND (date IS NOT NULL) AND (block
 IS NOT NULL) AND (IUCR IS NOT NULL) AND (type IS NOT NULL) AND (description IS NOT NULL) AND (arrest IS NOT NULL) AND (domes
tic IS NOT NULL) AND (beat IS NOT NULL) AND (district IS NOT NULL) AND (ward IS NOT NULL) AND (area IS NOT NULL) AND (FBI_Cod
e IS NOT NULL) AND (X IS NOT NULL) AND (Y IS NOT NULL) AND (year IS NOT NULL) AND (updated_on IS NOT NULL) AND (lat IS NOT NU
LL) AND (longitude IS NOT NULL) AND (location IS NOT NULL);
grunt> █
```

1. To calculate the number of cases investigated under each FBI code.

Ans -

- First grouped the filtered crime data by fbi code.

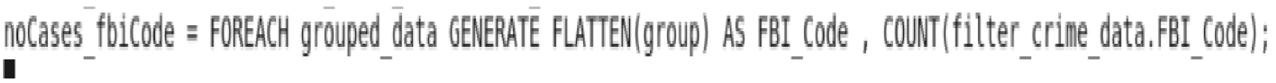
grouped_data = GROUP filter_crime_data BY FBI_Code;



```
grunt> grouped_data =  GROUP filter_crime_data BY FBI_Code;
```

- For each fbi code group , tuples are generated and no of tuples is counted by flattening it.

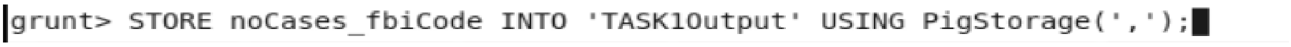
noCases_fbiCode = FOREACH grouped_data GENERATE FLATTEN(group) AS FBI_Code ,COUNT(filter_crime_data.FBI_Code);



```
noCases_fbiCode = FOREACH grouped_data GENERATE FLATTEN(group) AS FBI_Code , COUNT(filter_crime_data.FBI_Code);
```

- Stored the data(result) into a folder

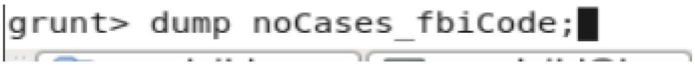
STORE noCases_fbiCode INTO 'TASK1Output' USING PigStorage(',');dump



```
grunt> STORE noCases_fbiCode INTO 'TASK1Output' USING PigStorage(',');
```

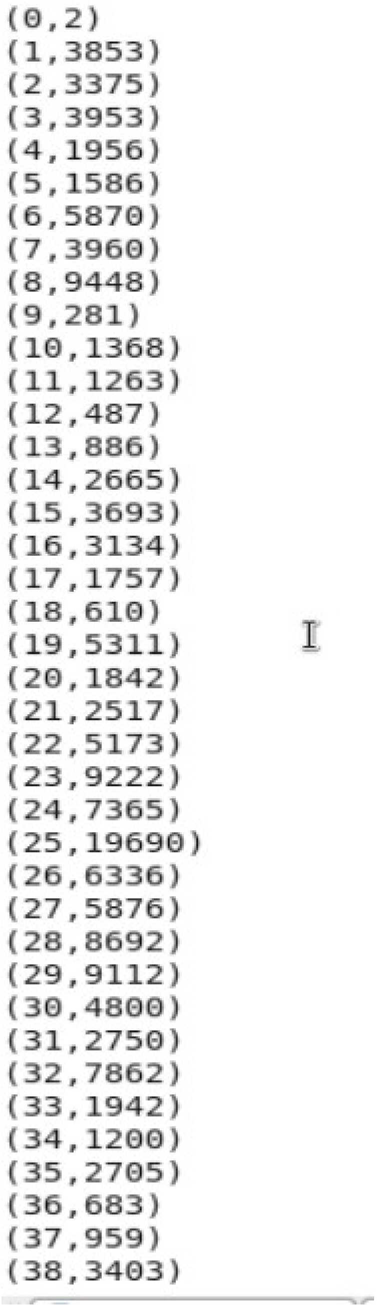
- Dumped the data for display.

  dump noCases_fbiCode;

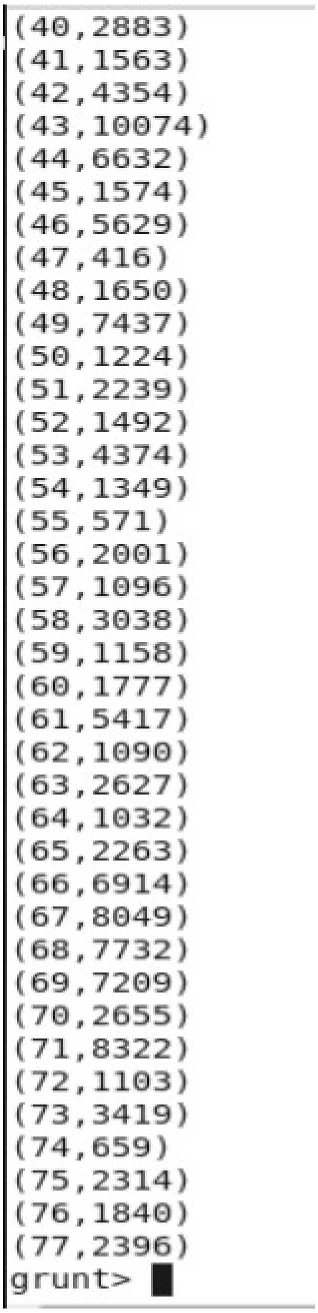


```
grunt> dump noCases_fbiCode;
```

Output -



```
(0,2)
(1,3853)
(2,3375)
(3,3953)
(4,1956)
(5,1586)
(6,5870)
(7,3960)
(8,9448)
(9,281)
(10,1368)
(11,1263)
(12,487)
(13,886)
(14,2665)
(15,3693)
(16,3134)
(17,1757)
(18,610)
(19,5311)
(20,1842)
(21,2517)
(22,5173)
(23,9222)
(24,7365)
(25,19690)
(26,6336)
(27,5876)
(28,8692)
(29,9112)
(30,4800)
(31,2750)
(32,7862)
(33,1942)
(34,1200)
(35,2705)
(36,683)
(37,959)
(38,3403)
```



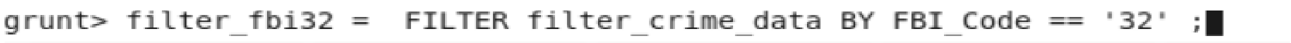
```
(40,2883)
(41,1563)
(42,4354)
(43,10074)
(44,6632)
(45,1574)
(46,5629)
(47,416)
(48,1650)
(49,7437)
(50,1224)
(51,2239)
(52,1492)
(53,4374)
(54,1349)
(55,571)
(56,2001)
(57,1096)
(58,3038)
(59,1158)
(60,1777)
(61,5417)
(62,1090)
(63,2627)
(64,1032)
(65,2263)
(66,6914)
(67,8049)
(68,7732)
(69,7209)
(70,2655)
(71,8322)
(72,1103)
(73,3419)
(74,659)
(75,2314)
(76,1840)
(77,2396)
grunt> 
```

For each fbi code , there is the no of cases investigated under it is diplayed.

2. To calculate the number of cases investigated under FBI code 32.

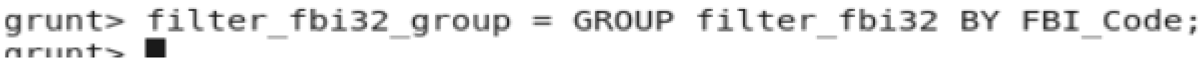
Ans -

- From the null filtered crime data(filter_crime_data) , we again filtered where fbi code = 32.
  filter_fbi32 = FILTER filter_crime_data BY FBI_Code == '32';



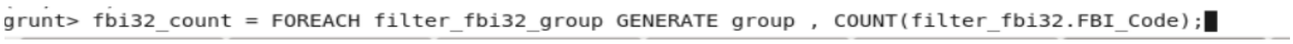
```
grunt> filter_fbi32 =  FILTER filter_crime_data BY FBI_Code == '32' ;
```

- We grouped the data by FBI_code .
  filter_fbi32_group = GROUP filter_fbi32 BY FBI_Code;



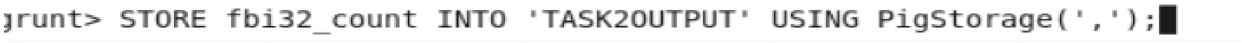
```
grunt> filter_fbi32_group = GROUP filter_fbi32 BY FBI_Code;
grunt>
```

- For fbi code = 32 , the no of cases invsetigated is counted
  fbi32_count = FOREACH filter_fbi32_group GENERATE group, COUNT(filter_fbi32.FBI_Code);



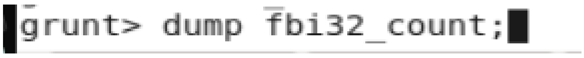
```
grunt> fbi32_count = FOREACH filter_fbi32_group GENERATE group , COUNT(filter_fbi32.FBI_Code);
```

- Store the final result into a folder
  STORE fib32_count INTO 'TASK2OUTPUT' USING PigStorage(',');



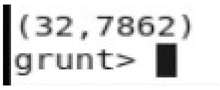
```
grunt> STORE fbi32_count INTO 'TASK2OUTPUT' USING PigStorage(',');
```

- Dumped the result into screen.
  dump fib32_count;



```
grunt> dump fbi32_count;
```

Output -



```
(32,7862)
grunt>
```

3. To calculate the number of arrests in theft district wise.

Ans -

- Filtered the crime data which was already filtered for null where crime type is THEFT.

  theftArrest_data = FILTER filter_crime_data By type == 'THEFT';

```
grunt> theftArrest_data =  FILTER filter_crime_data By type == 'THEFT' ;
```

- Grouped the filtered data district wise.
  theftArrest_district = GROUP theftArrest_data BY district;

```
grunt> theftArrest_district =  GROUP theftArrest_data BY district;
```

- For each group district , no of tuples in it is flattened and then counted for the no cases in each district.
  theftArrest_district_count = FOREACH theftArrest_district GENERATE FLATTEN(group),COUNT(theftArrest_data.district);

```
grunt> theftArrest_district_count =  FOREACH theftArrest_district GENERATE FLATTEN(group),COUNT(theftArrest_data.district);
```

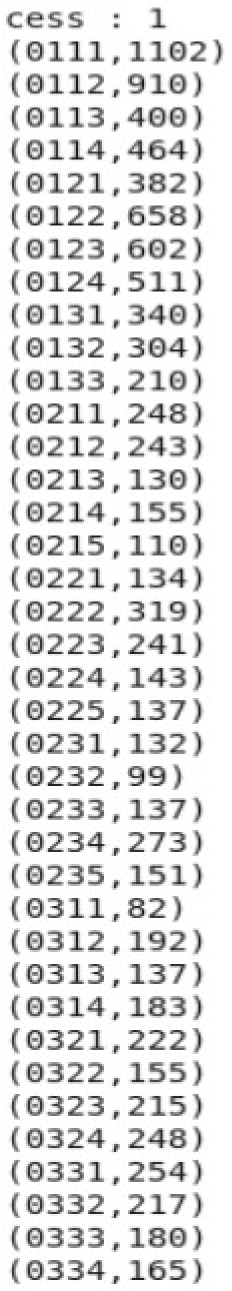
- Store the result that is district and its no of cases in a folder.
  STORE theftArrest_district_count INTO 'TASK3OUTPUT' USING PigStorage(',');

```
grunt> STORE theftArrest_district_count INTO 'TASK3OUTPUT' USING PigStorage(',');
```

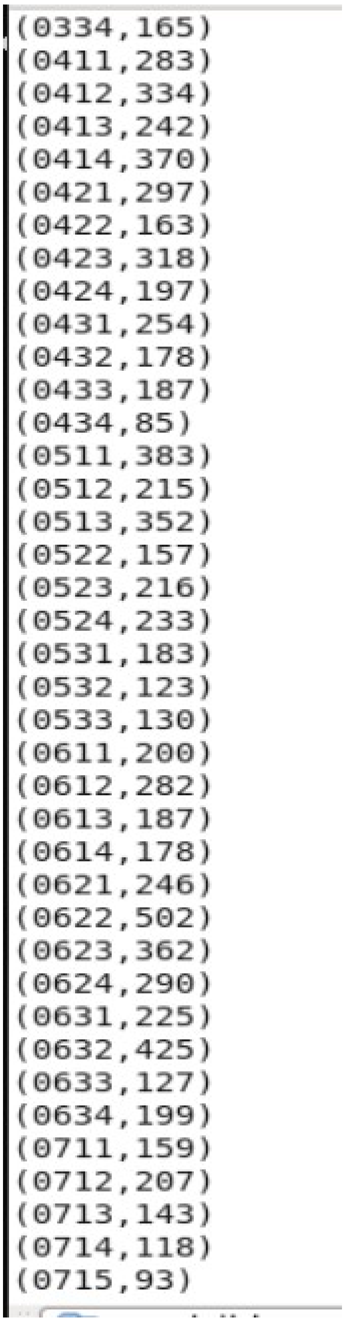
- Dumped the result on to screen.
  dump theftArrest_district_count;

```
grunt> dump theftArrest_district_count;
```

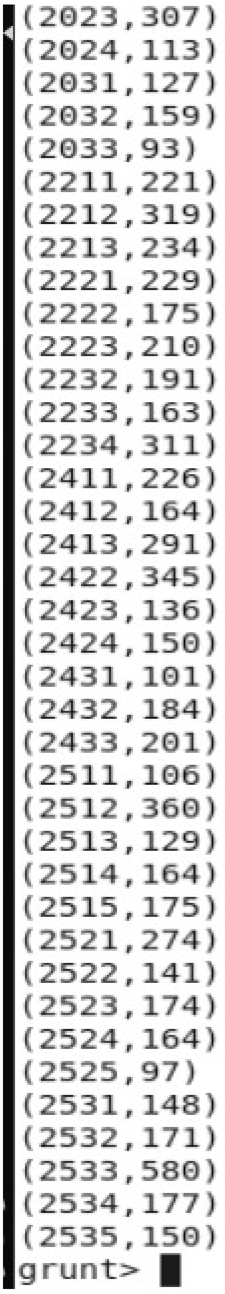
Output -



```
cess : 1
(0111,1102)
(0112,910)
(0113,400)
(0114,464)
(0121,382)
(0122,658)
(0123,602)
(0124,511)
(0131,340)
(0132,304)
(0133,210)
(0211,248)
(0212,243)
(0213,130)
(0214,155)
(0215,110)
(0221,134)
(0222,319)
(0223,241)
(0224,143)
(0225,137)
(0231,132)
(0232,99)
(0233,137)
(0234,273)
(0235,151)
(0311,82)
(0312,192)
(0313,137)
(0314,183)
(0321,222)
(0322,155)
(0323,215)
(0324,248)
(0331,254)
(0332,217)
(0333,180)
(0334,165)
```



```
(0334,165)
(0411,283)
(0412,334)
(0413,242)
(0414,370)
(0421,297)
(0422,163)
(0423,318)
(0424,197)
(0431,254)
(0432,178)
(0433,187)
(0434,85)
(0511,383)
(0512,215)
(0513,352)
(0522,157)
(0523,216)
(0524,233)
(0531,183)
(0532,123)
(0533,130)
(0611,200)
(0612,282)
(0613,187)
(0614,178)
(0621,246)
(0622,502)
(0623,362)
(0624,290)
(0631,225)
(0632,425)
(0633,127)
(0634,199)
(0711,159)
(0712,207)
(0713,143)
(0714,118)
(0715,93)
```



```
(2023,307)
(2024,113)
(2031,127)
(2032,159)
(2033,93)
(2211,221)
(2212,319)
(2213,234)
(2221,229)
(2222,175)
(2223,210)
(2232,191)
(2233,163)
(2234,311)
(2411,226)
(2412,164)
(2413,291)
(2422,345)
(2423,136)
(2424,150)
(2431,101)
(2432,184)
(2433,201)
(2511,106)
(2512,360)
(2513,129)
(2514,164)
(2515,175)
(2521,274)
(2522,141)
(2523,174)
(2524,164)
(2525,97)
(2531,148)
(2532,171)
(2533,580)
(2534,177)
(2535,150)
grunt>
```

4. To calculate the number of arrests done between October 2014 and October 2015.

Ans -

- Formatted the date of the filtered crime data.
  dateformatted = FOREACH filter_crime_data GENERATE ToDate(SUBSTRING(date,0,10),'MM/dd/YYYY') as (dt:datetime);

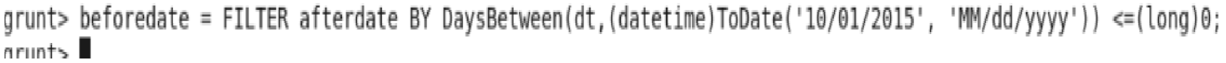
```
grunt> dateformatted = FOREACH filter_crime_data GENERATE ToDate(SUBSTRING(date,0,10),'MM/dd/YYYY') as (dt:datetime);
```

- Filter the data where date is after 10/01/2014.

afterdate = FILTER dateformatted BY DaysBetween(dt,(datetime)ToDate('10/01/2014', 'MM/dd/yyyy')) >=(long)0;

```
grunt> afterdate = FILTER dateformatted BY DaysBetween(dt,(datetime)ToDate('10/01/2014', 'MM/dd/yyyy')) >=(long)0;
```

- Filter the data where date is before 10/01/2015.
  beforedate = FILTER afterdate BY DaysBetween(dt,(datetime)ToDate('10/01/2015', 'MM/dd/yyyy')) <=(long)0;



```
grunt> beforedate = FILTER afterdate BY DaysBetween(dt,(datetime)ToDate('10/01/2015', 'MM/dd/yyyy')) <=(long)0;
grunt>
```

- Then grouped the filetred data .
  grouped = GROUP beforedate ALL;

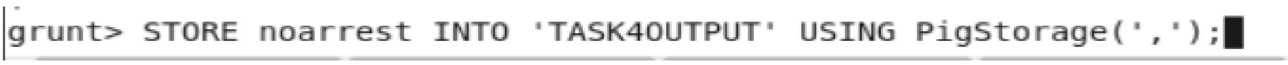
```
grunt> grouped = GROUP beforedate ALL;
```

- For the group , counted the no of tuples in it using count function.

  noarrest = FOREACH grouped GENERATE COUNT(beforedate);

```
grunt> noarrest = FOREACH grouped GENERATE COUNT(beforedate);
```

- Store the result into a folder
  STORE noarrest INTO 'TASK4OUTPUT' USING PigStorage(',');

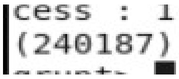


```
grunt> STORE noarrest INTO 'TASK4OUTPUT' USING PigStorage(',');
```

- Dumped the result onto screen.

  dump noarrest;

Output -



```
cess : 1
(240187)
grunt>
```