Project 1 - USA Crime Analysis

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.

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

crime\_data = load '/user/ec2-user/pigdata/ usacrime/Crimes\_-\_2001\_to\_present.csv' using PigStorage(',') as (ID:INT,Case\_Number:chararray,Date:chararray,Block:chararray,IUCR:chararray,Primary\_Type:chararray,Description:chararray,Location\_Description:chararray,Arrest:chararray,Domestic:chararray,Beat:chararray,District:chararray,Ward:chararray,Community\_Area:chararray,FBICode:chararray,XCoordinate:chararray,YCoordinate:chararray,Year:INT,Updated\_On:chararray,Latitude:chararray,Longitude:chararray,Location:chararray);

Problem Statement

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

1. group cases by fbi code

**case\_fbi\_group = group crime\_data by FBICode;**

2.find COUNT for each group

**number\_od\_cases\_per\_fbicode = foreach case\_fbi\_group generate group as fbicode, COUNT(crime\_data) as numberofcases;**

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

**number\_of\_cases\_fbicode\_32 = FILTER number\_of\_cases\_per\_fbicode by fbicode=='32';**

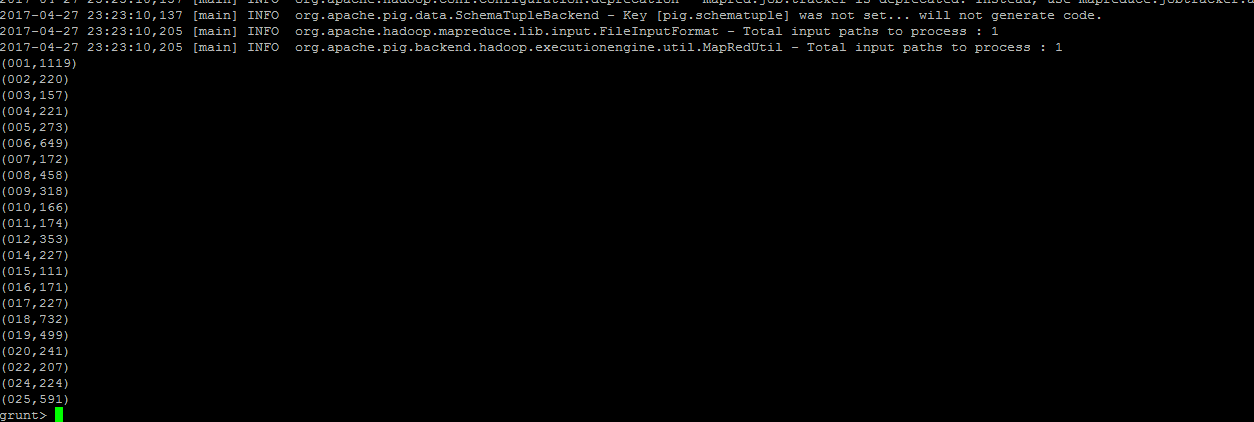
**3.Write a mapreduce/pig program to calculate the number of arrests in theft district wise.**

cases\_in\_theft = FILTER crime\_data by Primary\_Type=='THEFT';

arrest\_in\_theft = FILTER cases\_in\_theft by Arrest matches '(?i).\*TRUE.\*';

district\_wise\_theft\_arrest = group arrest\_in\_theft by District;

**number\_of\_district\_wise\_arrest\_in\_theft = foreach district\_wise\_theft\_arrest generate group as district, COUNT(arrest\_in\_theft) as numberOfArrest;**



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

total\_arrests = FILTER crime\_data by Arrest matches '(?i).\*TRUE.\*';

date\_of\_arrests = FOREACH total\_arrests GENERATE Date;

clean the date field

a = foreach date\_of\_arrests generate REGEX\_EXTRACT\_ALL(Date,'(\\d+)/(\\d+)/(\\d+).\*') as date:(month, day, year);

construct date format

date\_concat = foreach a generate CONCAT(date.month,'/',date.day,'/',date.year) as dateconcat;

convert to datetime

todate\_data = foreach date\_concat generate ToDate(dateconcat,'MM/dd/yyyy') as (date\_time:DateTime);

filter the date range, 2014 oct to 2015 oct

filtered = filter todate\_data by date\_time>=ToDate('10/01/2014','MM/dd/yyyy') AND date\_time<ToDate('10/01/2015','MM/dd/yyyy');

find the count by group all

group\_filtered\_date = GROUP filtered ALL;

number\_of\_arrests = foreach group\_filtered\_date generate COUNT(filtered) as numberOfArrests;

dump number\_of\_arrests

(63173)

