

ACADGILD

# BIG DATA AND HADOOP TRAINING

---

*Project 1.1 - USA Crime Analysis*

---

**Ankita Jaiswal**

**11/24/2017**

## **TABLE OF CONTENTS**

- Introduction
- Data Files
- Crime Sample Dataset
- Dataset Description
- Problem Statement
- Output

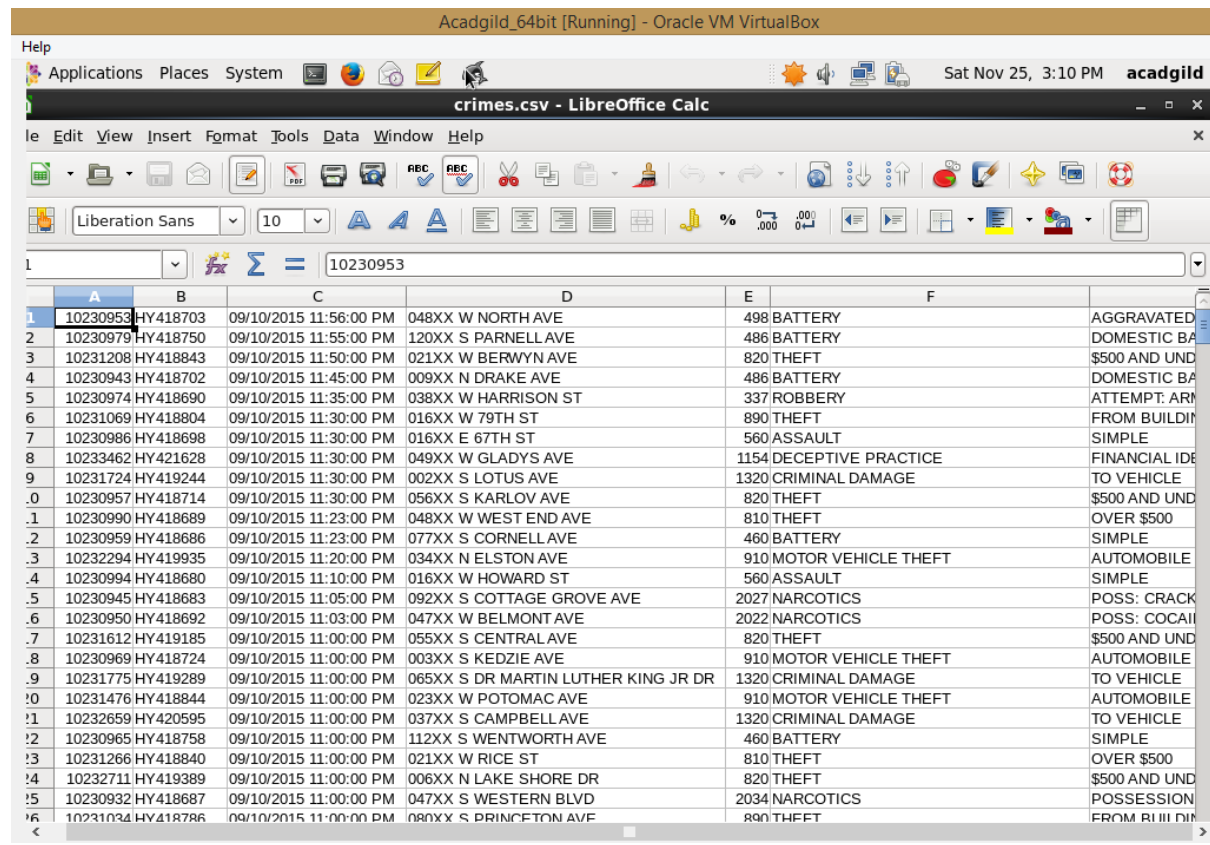
# Introduction

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.

## Data Files

<https://drive.google.com/file/d/0B1QaXx7tpw3SaUJHOHBZclBXWG8/view?usp=sharing>

## Crime Sample Dataset



	A	B	C	D	E	F
1	10230953	HY418703	09/10/2015 11:56:00 PM	048XX W NORTH AVE	498 BATTERY	AGGRAVATED
2	10230979	HY418750	09/10/2015 11:55:00 PM	120XX S PARNELL AVE	486 BATTERY	DOMESTIC BA
3	10231208	HY418843	09/10/2015 11:50:00 PM	021XX W BERWYN AVE	820 THEFT	\$500 AND UND
4	10230943	HY418702	09/10/2015 11:45:00 PM	009XX N DRAKE AVE	486 BATTERY	DOMESTIC BA
5	10230974	HY418690	09/10/2015 11:35:00 PM	038XX W HARRISON ST	337 ROBBERY	ATTEMPT: ARM
6	10231069	HY418804	09/10/2015 11:30:00 PM	016XX W 79TH ST	890 THEFT	FROM BUILDIN
7	10230986	HY418698	09/10/2015 11:30:00 PM	016XX E 67TH ST	560 ASSAULT	SIMPLE
8	10233462	HY421628	09/10/2015 11:30:00 PM	049XX W GLADYS AVE	1154 DECEPTIVE PRACTICE	FINANCIAL IDE
9	10231724	HY419244	09/10/2015 11:30:00 PM	002XX S LOTUS AVE	1320 CRIMINAL DAMAGE	TO VEHICLE
0	10230957	HY418714	09/10/2015 11:30:00 PM	056XX S KARLOV AVE	820 THEFT	\$500 AND UND
1	10230990	HY418689	09/10/2015 11:23:00 PM	048XX W WEST END AVE	810 THEFT	OVER \$500
2	10230959	HY418686	09/10/2015 11:23:00 PM	077XX S CORNELL AVE	460 BATTERY	SIMPLE
3	10232294	HY419935	09/10/2015 11:20:00 PM	034XX N ELSTON AVE	910 MOTOR VEHICLE THEFT	AUTOMOBILE
4	10230994	HY418680	09/10/2015 11:10:00 PM	016XX W HOWARD ST	560 ASSAULT	SIMPLE
5	10230945	HY418683	09/10/2015 11:05:00 PM	092XX S COTTAGE GROVE AVE	2027 NARCOTICS	POSS: CRACK
6	10230950	HY418692	09/10/2015 11:03:00 PM	047XX W BELMONT AVE	2022 NARCOTICS	POSS: COCAI
7	10231612	HY419185	09/10/2015 11:00:00 PM	055XX S CENTRAL AVE	820 THEFT	\$500 AND UND
8	10230969	HY418724	09/10/2015 11:00:00 PM	003XX S KEDZIE AVE	910 MOTOR VEHICLE THEFT	AUTOMOBILE
9	10231775	HY419289	09/10/2015 11:00:00 PM	065XX S DR MARTIN LUTHER KING JR DR	1320 CRIMINAL DAMAGE	TO VEHICLE
0	10231476	HY418844	09/10/2015 11:00:00 PM	023XX W POTOMAC AVE	910 MOTOR VEHICLE THEFT	AUTOMOBILE
1	10232659	HY420595	09/10/2015 11:00:00 PM	037XX S CAMPBELL AVE	1320 CRIMINAL DAMAGE	TO VEHICLE
2	10230965	HY418758	09/10/2015 11:00:00 PM	112XX S WENTWORTH AVE	460 BATTERY	SIMPLE
3	10231266	HY418840	09/10/2015 11:00:00 PM	021XX W RICE ST	810 THEFT	OVER \$500
4	10232711	HY419389	09/10/2015 11:00:00 PM	006XX N LAKE SHORE DR	820 THEFT	\$500 AND UND
5	10230932	HY418687	09/10/2015 11:00:00 PM	047XX S WESTERN BLVD	2034 NARCOTICS	POSSESSION
6	10231034	HY418786	09/10/2015 11:00:00 PM	080XX S PRINCETON AVE	890 THEFT	FROM BUILDIN

## Dataset Description

The columns present in the dataset:

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

# Problem Statement

- Pig program to calculate the number of cases investigated under each FBI code

## Loading the crime dataset

```
grunt> crime_dataset = LOAD '/home/acadgild/ankita/Project12_1/crimes.csv'
>> USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
>> 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:int,
>> district:int,
>> ward:int,community_area:int,fbi_code:chararray,x_coordinate:int,y_coordinate:int,year:int,
>> updated_on:chararray,latitude:float,longitude:float,location:chararray);
2017-11-25 15:37:54,692 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.persist.jobs
hours is deprecated. Instead, use mapreduce.jobtracker.persist.jobstatus.hours
2017-11-25 15:37:54,697 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.heartbeats.in.second is
ated. Instead, use mapreduce.jobtracker.heartbeats.in.second
```

## Grouping the fbi\_code

```
grunt> fbi_code_group = GROUP crime_dataset BY fbi_code;
```

## Calculating case number for each fbi\_code

```
grunt> cases_number = FOREACH fbi_code_group GENERATE FLATTEN(group) AS fbi_code, COUNT($1);
```

## Showing the the number of cases investigated under each FBI code

```
grunt> dump cases_number;
```

## Output:

```
cated. Instea
2017-11-25 15
2017-11-25 15
2017-11-25 15
cess : 1
(02,1502)
(03,10596)
(05,14842)
(06,64329)
(07,11105)
(09,445)
(10,1551)
(11,13757)
(12,27)
(13,57)
(14,31301)
(15,3694)
(16,1787)
(17,1126)
(18,25207)
(19,434)
(20,1267)
(22,371)
(24,4046)
(26,29474)
(01A,533)
(01B,6)
(04A,4994)
(04B,7710)
(08A,14167)
(08B,46938)
(,1)
```

- Pig program to calculate the number of cases investigated under FBI code 32

## Loading the crime dataset

```
grunt> crime_dataset = LOAD '/home/acadgild/ankita/Project12_1/crimes.csv'
>> USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
>> 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:int,
>> district:int,
>> ward:int,community_area:int,fbi_code:chararray,x_coordinate:int,y_coordinate:int,year:int,
>> updated_on:chararray,latitude:float,longitude:float,location:chararray);
2017-11-25 16:05:19,039 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.cou
cated. Instead, use mapreduce.job.counters.max
```

## Generating case\_number , fbi\_code from crime dataset

```
|grunt> fbi = FOREACH crime_dataset GENERATE case_number,fbi_code;
```

## Filtering fbi\_code =32 from alias fbi

```
|grunt> filter_fbi = FILTER fbi BY fbi_code == '32';
```

## Grouping the filter\_fbi

```
|grunt> group_filter_fbi = GROUP filter_fbi BY fbi_code;
```

## Calculating number of cases for fbi code 32

```
|grunt> case_count = FOREACH group_filter_fbi GENERATE group, COUNT(filter_fbi.fbi_code);
```

## Showing number of cases for fbi code 32

```
|grunt> dump case_count;
```

Output:

```
Acadgild_64bit [Running] - Oracle VM VirtualBox
Help
Input(s):
Successfully read 291267 records from: "/home/acadgild/ankita/Project12_1/crimes.csv"

Output(s):
Successfully stored 0 records in: "file:/tmp/temp-481921214/tmp-261862494"

Counters:
Total records written : 0
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1889679833_0002

2017-11-25 16:36:00,343 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-25 16:36:00,355 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-25 16:36:00,357 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-25 16:36:00,420 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encount
ered Warning ACCESSING_NON_EXISTENT_FIELD 2 time(s).
2017-11-25 16:36:00,420 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-11-25 16:36:00,437 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-11-25 16:36:00,446 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-11-25 16:36:00,446 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-11-25 16:36:00,446 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-25 16:36:00,598 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-25 16:36:00,598 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
grunt> |
```

- Pig program to calculate the number of arrests in theft district wise

#### Loading crime dataset

```
grunt> crime_dataset = LOAD '/home/acadgild/ankita/Project12_1/crimes.csv'
>> USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
>> 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:int,
>> district:int,
>> ward:int,community_area:int,fbi_code:chararray,x_coordinate:int,y_coordinate:int,year:int,
>> updated_on:chararray,latitude:float,longitude:float,location:chararray);
2017-11-25 16:05:19,039 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.cou
cated. Instead, use mapreduce.job.counters.max
```

#### Filtering theft from primary type

```
grunt> theft = FILTER crime_dataset BY primary_type == 'THEFT';
```

#### Generating arrest, district from alias theft

```
grunt> t = FOREACH theft GENERATE arrest,district;
```

#### Filtering alias t

```
grunt> u = FILTER t BY arrest == 'true';
```

#### Grouping alias u by district

```
grunt> district_group = GROUP u BY district;
```

#### Calculating the number of arrests in theft district wise

```
grunt> arrest_district = FOREACH district_group GENERATE group, COUNT(u.district);
```

#### Showing calculated the number of arrests in theft district wise

```
grunt> dump arrest_district;
```

Output:

```
Input(s):
Successfully read 291267 records from: "/home/acadgild/ankita/Project12_1/crimes.csv"
```

```
Output(s):
Successfully stored 22 records in: "file:/tmp/temp-89386445/tmp-1078916974"
```

```
Counters:
Total records written : 22
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

```
Job DAG:
job_local1865092730_0002
```

```
2017-11-25 17:14:56,333 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot
2017-11-25
cess : 1
(1,1124)
(2,227)
(3,162)
(4,230)
(5,286)
(6,652)
(7,176)
(8,471)
(9,320)
(10,170)
(11,178)
(12,360)
(14,228)
(15,115)
(16,177)
(17,237)
(18,734)
(19,501)
(20,244)
(22,220)
(24,226)
(25,596)
grunt> █
```

- Pig program to calculate the number of arrests done between October 2014 and October 2015

### Loading crime dataset

```
grunt> crime_dataset = LOAD '/home/acadgild/ankita/Project12_1/crimes.csv'
>> USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
>> 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:int,
>> district:int,
>> ward:int,community_area:int,fbi_code:chararray,x_coordinate:int,y_coordinate:int,year:int,
>> updated_on:chararray,latitude:float,longitude:float,location:chararray);
2017-11-25 16:05:19,039 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.cou
cated. Instead, use mapreduce.job.counters.max
```

---

Fetching date, arrest from crime dataset

```
grunt> a = FOREACH crime_dataset GENERATE date as dt ,arrest as at;
```

Filtering date not null and arrest true

```
grunt> b = FILTER a BY (dt IS NOT NULL) AND (at == 'true');
```

Changing date format

```
grunt> c = FOREACH b GENERATE ToDate(SUBSTRING(dt,0,19), 'MM/DD/YYYY hh:mm:ss') as dte,at;
```

Fetching month year and arrest

```
grunt> d = FOREACH c GENERATE GetMonth(dte) as month,GetYear(dte) as year,at;
```

Fetching and grouping the alias d between October 2014 and October 2015

```
grunt> e = FILTER d BY (month > 9 AND year == 2014) OR (month < 11 AND year == 2015);
grunt> group_arrest = GROUP e ALL;
```

Calculating number of arrest between October 2014 and October 2015

```
grunt> count_arrest = FOREACH group_arrest GENERATE COUNT(e.at) as NumberofArrest;
```

Showing the number of arrest records for October 2014 and October 2015

```
grunt> dump count_arrest;
```

Output:

```
Output(s):
Successfully stored 1 records in: "file:/tmp/tmp-89386445/tmp-1008839855"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1576982156_0003

2017-11-25 19:36:14,033 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetric
e=JobTracker, sessionId= - already initialized
2017-11-25 19:36:14,048 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetric
e=JobTracker, sessionId= - already initialized
2017-11-25 19:36:14,052 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetric
e=JobTracker, sessionId= - already initialized
2017-11-25 19:36:14,125 [main] WARN org.apache.pig.backend.hadoop.executic
ered Warning ACCESSING NON EXISTENT FIELD 2 time(s).
2017-11-25 19:36:14,126 [main] INFO org.apache.pig.backend.hadoop.executic
!
2017-11-25 19:36:14,130 [main] INFO org.apache.hadoop.conf.Configuration.c
Instead, use dfs.bytes-per-checksum
2017-11-25 19:36:14,131 [main] INFO org.apache.hadoop.conf.Configuration.c
d, use fs.defaultFS
2017-11-25 19:36:14,133 [main] INFO org.apache.hadoop.conf.Configuration.c
cated. Instead, use mapreduce.job.counters.max
2017-11-25 19:36:14,133 [main] WARN org.apache.pig.data.SchemaTupleBackenc
2017-11-25 19:36:14,250 [main] INFO org.apache.hadoop.mapreduce.lib.input.
2017-11-25 19:36:14,250 [main] INFO org.apache.pig.backend.hadoop.executic
cess : 1
(47174)
grunt>
```

[(no subject) - ankitaja... acadgild@localhost: ~