Projects/Apache-Pig/Chic...

Chicago Crime Data Analysis

FINISHED

CHICAGO DATA PORTAL

Analysis of crime incidents in Chicago city area

The purpose of this notebook is to analyze the crime incidents that have occurred in the city of **Chicago**

The dataset has been obtained from the following source: Chicago (https://data.cityofchicago.org)

The Objective

The objective of this dataset to analyze the following:
* Overview
* Homicide Arrest Analysis
* Theft Analysis

Took 0 sec. Last updated by anonymous at April 20 2018, 9:29:57 AM. (outdated)

FINISHED

```
%sh

# steps to download the dataset
# create the local download directory if does not exist
mkdir -p ~/data-downloads
rm ~/data-downloads/*.csv

# create the required HDFS directory
hdfs dfs -mkdir -p /user/cloudera/data/raw/crime/chicago
```

```
# download the data from Chicago city data website
wget -O ~/data-downloads/crime_chicago.csv https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD

# copy the file to HDFS
hdfs dfs -put -f ~/data-downloads/crime_chicago.csv /user/cloudera/data/raw/crime/chicago/

# delete local data files
rm ~/data-downloads/*.csv
```

FINISHED

```
%sh

# steps to prepare the dataset for analysis
# download the piggybank jar from maven repository
cd ~/data-downloads/
wget http://central.maven.org/maven2/org/apache/pig/piggybank/0.17.0/piggybank-0.17.0.jar

# upload the piggybank jar to the HDFS jars directory
hdfs dfs -put -f ~/data-downloads/piggybank-0.17.0.jar /user/cloudera/jars/
```

FINISHED

```
%sh

# delete staging area
hdfs dfs -rm -r -skipTrash /user/cloudera/data/staging/crime/chicago
```

FINISHED

```
%pig

--register the piggybank jar
register hdfs://quickstart.cloudera:8020/user/cloudera/jars/piggybank-0.17.0.jar

--read contents from the chicago dataset on HDFS
raw_data_chicago = LOAD '/user/cloudera/data/raw/crime/chicago' USING org.apache.pig.piggybank.storage.CSVLoader() AS (id: chararray, case
    chararray, arrest: chararray, domestic: chararray, beat: chararray, district: chararray, ward: chararray, community_area: chararray, f
    longitude: chararray, location: chararray);

--remove the header record from the dataset
raw_data_chicago_headless = FILTER raw_data_chicago BY id != 'ID';

-- select only the required columns
data_chicago = FOREACH raw_data_chicago_headless GENERATE date, block, primary_type, description, location_description, arrest, domestic,

-- store the final data into another file
STORE data_chicago INTO '/user/cloudera/data/staging/crime/chicago' USING PigStorage(';');
```

Took 3 min 7 sec. Last updated by anonymous at April 20 2018, 9:24:19 AM. (outdated)

Analysis #1: Overview

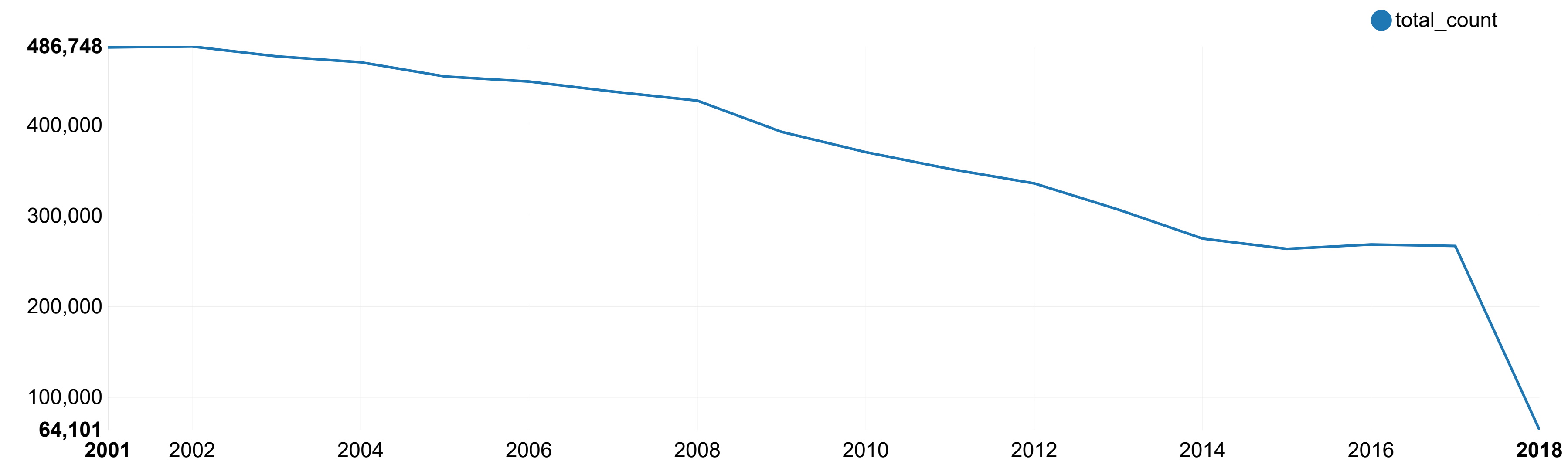
FINISHED

Took 0 sec. Last updated by anonymous at April 20 2018, 11:08:11 AM. (outdated)

Number of Incidents across the years

FINISHED

settings ▾



Took 3 min 58 sec. Last updated by anonymous at April 20 2018, 10:58:03 AM. (outdated)

Top N Primary Cause of Incidents

FINISHED

top_n

5

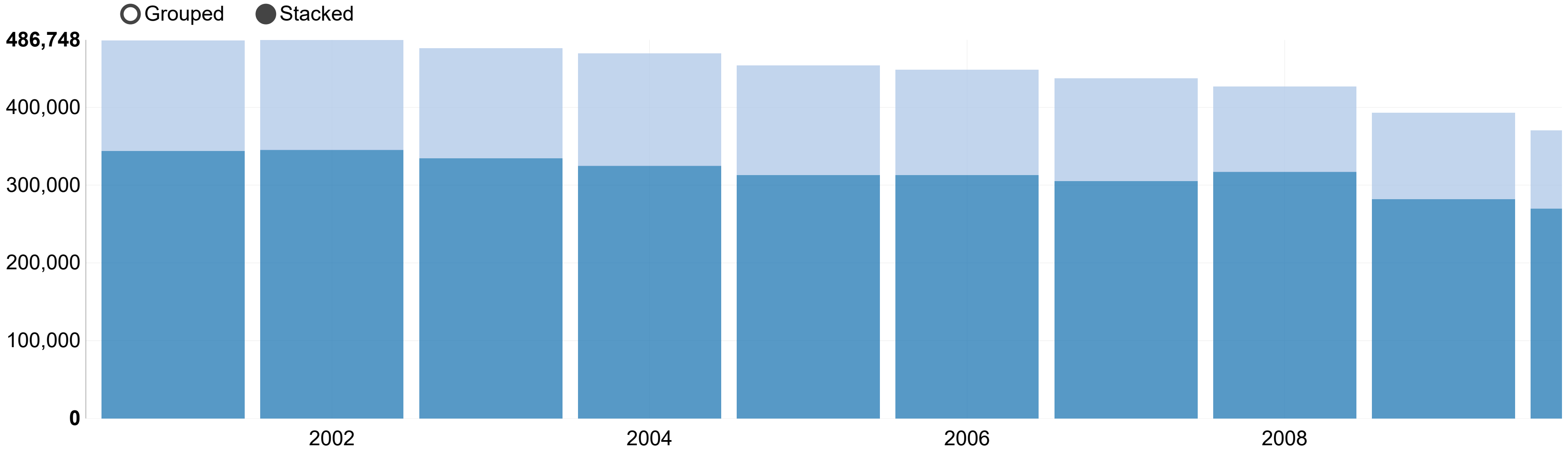
primary_type ▼	total_count ▼
THEFT	1376721
BATTERY	1200595
CRIMINAL DAMAGE	753842
NARCOTICS	702985
OTHER OFFENSE	408026

Took 5 min 30 sec. Last updated by anonymous at April 20 2018, 10:36:33 AM. (outdated)

Comparison of Arrest to Non-Arrest Incidents

FINISHED

settings ▾



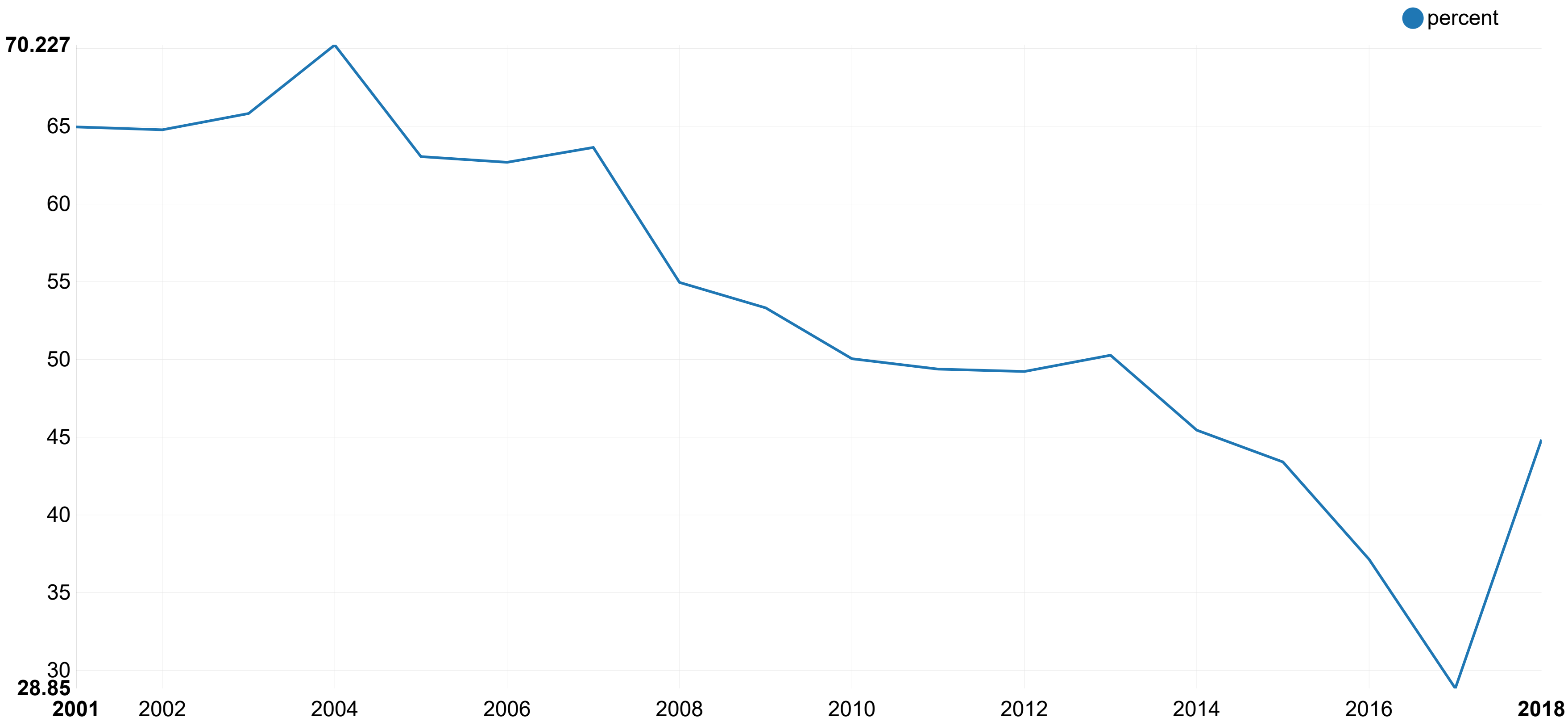
Took 3 min 53 sec. Last updated by anonymous at April 20 2018, 9:28:19 AM. (outdated)

Analysis #2: Homicide Arrest Analysis

FINISHED

Took 0 sec. Last updated by anonymous at April 20 2018, 11:08:36 AM. (outdated) settings

FINISHED



Took 4 min 20 sec. Last updated by anonymous at April 19 2018, 6:41:49 PM. (outdated)

FINISHED

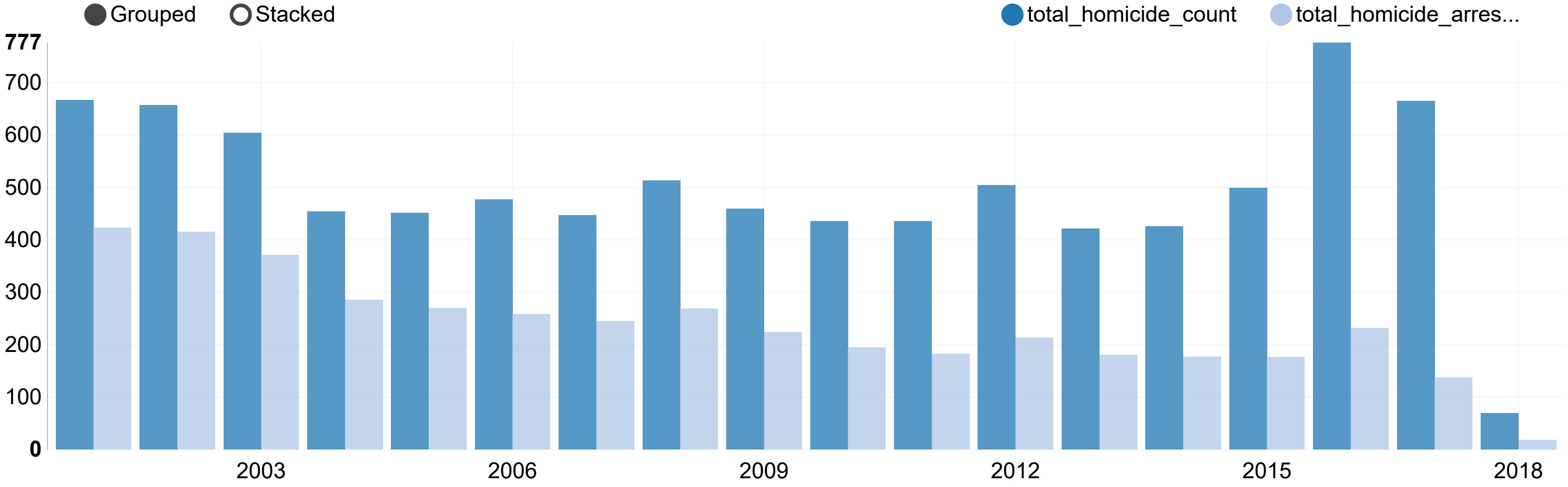
In the first chart, we look at the average percentage of arrests with respect to homicide cases over time and it looks like the percentage of arrests have decreased with respect to homicide cases over the years.

Now let's look at the number of incidents to understand the trend over years and see if we can derive any insights out of that.

Took 1 sec. Last updated by anonymous at April 19 2018, 11:16:35 PM. (outdated)

FINISHED

settings ▾



Took 3 min 34 sec. Last updated by anonymous at April 19 2018, 11:20:17 PM. (outdated)

FINISHED

Looking at the bar chart, the number of homicide cases show a high trend over the last couple of years. Probably the length of the cases for arrests matter.

We cannot say for sure as we do not have further data to analyze with respect to cycle time for each of these cases.

Took 0 sec. Last updated by anonymous at April 19 2018, 11:21:52 PM. (outdated)

FINISHED

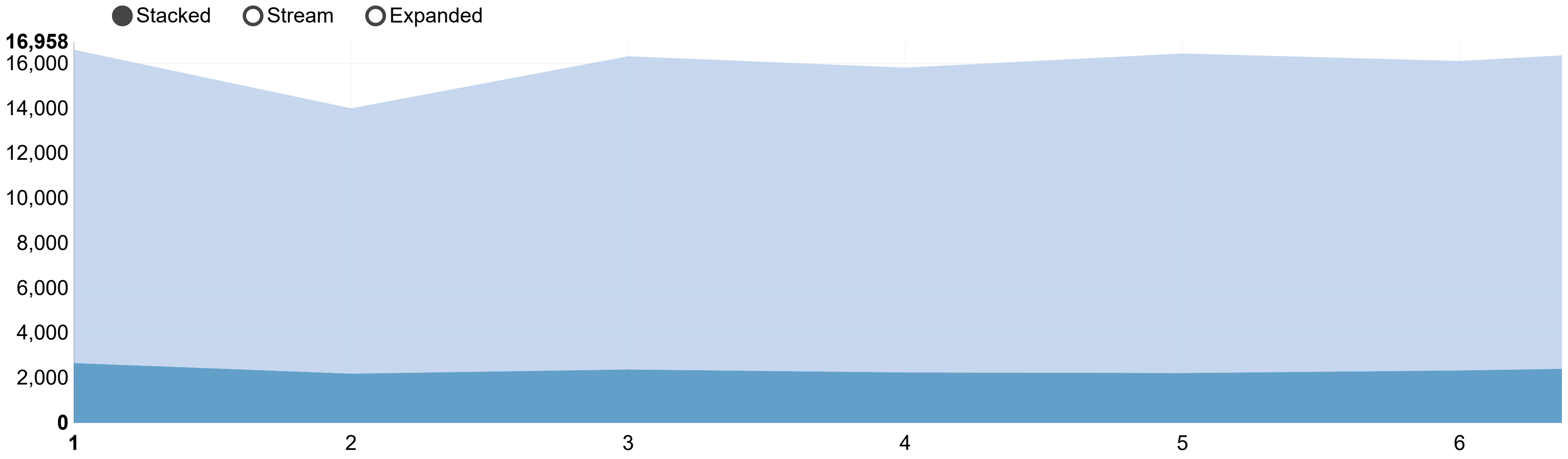
Analysis #3: Theft Analysis

Took 0 sec. Last updated by anonymous at April 20 2018, 11:09:00 AM. (outdated)

FINISHED

Number of thefts by month (all years combined)

settings ▾



Took 2 min 16 sec. Last updated by anonymous at April 19 2018, 6:54:19 PM. (outdated)

Thank you for going through this analysis!!

FINISHED

Took 0 sec. Last updated by anonymous at April 20 2018, 10:56:40 AM. (outdated)

```
%md
```

READY