Term Project Tutorial

CIS 5200 -1

Group 5

Era Kajal Singh, Neha Gupta, Tanmai Aurangabadkar,Ying Ying Lai

Crime Analysis in 3 Biggest Metropolitan Areas in USA using Hive in IBM BigInsights

Objective:

In this lab, you will analyze and visualize Crime Report Data. Thus,

- How to use Google Apis for files managed in google drive
- How to load data into hadoop systems powered by Bluemix BigInsights
- How to use Pig for transforming and enriching the data
- How to analyze the transformed data using HiveQL.
- How to present the analysis using powerful visualization tool like Tableau .

Introduction:

This paper aims to research the crime situation of the main cities in United States, for example, Los Angeles, Chicago, and New York City. The analysis of crime data can help us in deriving the relationships between different types of crime with respect to time of their occurrence, and location where crimes occurred. We also can derive insights about which part of these cities are relatively safer place to reside in. This analysis can also help in extracting information like what time is best to stay home due to safety issue and what location to avoid within the city to reduce risk.

For this analysis we are using datasets provided through open data portal for the respective city. Local governments are sharing variety of their data through these Open Data portals. These portal house the data related to budget, environment, transportation and public safety. In our analysis we are focusing on crime data available under "Public Safety" category.

Since, we are focusing on 3 different metropolitan areas in United States, there is a variety in data itself, hence we will focus on the features common across all the datasets for coherent analysis across datasets which is also aligned to our goals for doing such analysis. Further we have chosen zipcodes as sub-region boundary for aggregation in a metropolitan area, since zipcodes are standard area boundaries across united states. Here are few things we will learn through this tutorials.

- Extract (E in ETL process) data from city specific Open Data portals
- Transform(T in ETL process) data to enrich data with zipcodes from the location information available in the data using Pig scripts and UDFs.
- Load(L in ETL process) data into hadoop systems using HDFS and Hive commands
- Analyze data to get top 10 crimes in the entire metropolitan area. Using HiveQL
- Analyze data to get top 10 most crime inflicted zipcodes in a particular metropolitan area. Using HiveQL
- How many times top 10 crimes happened in all zipcodes. Using HiveQL
- How many times top 10 crimes happened in all zipcodes by day of a week. Using HiveQL
- How many times top 10 crimes happened in all zipcodes by hour of a day. Using HiveQL
- Top 10 crimes per year. Using HiveQL
- Transfer data using WINSCP (when using windows) and SCP commands(when using mac or bash)
- Visualize data in Tableau

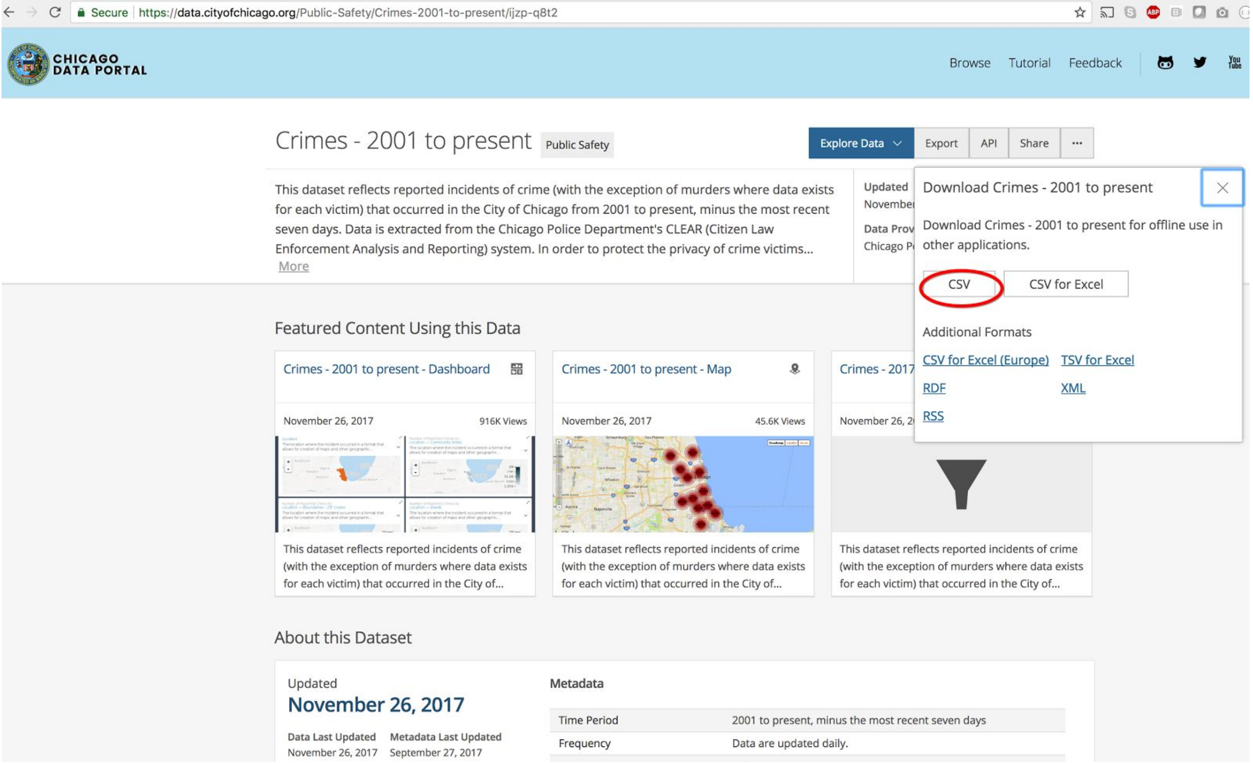
Prerequisites:

This analysis requires basic understanding of few basic bash commands as follows:

1. zip/unzip utility: To compress the data for reducing the time to download or upload the data.
2. curl: This command is needed to download the data on the remote linux systems from where it can be loaded into HDFS
3. scp: For extracting the processed data back on the local systems from the remote system after it is processed through hadoop and hive
4. winscp: For extracting the processed data back on the local systems from the remote system after it is processed, when our local system is windows and not mac or unix systems
5. Understanding of hdfs Commands: for loading data into hdfs
6. Pig version 0.16 or later. (Bluemix Analytics Engine)
7. Understanding of pig commands: for transforming/enriching the data
8. Understanding of hive commands: for running analysis queries
9. Tableau must be installed for visualizations

Step 1: Extracting the data:

1. Download the data locally from the Open Data portal of respective city as follows:



Secure | https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
CHICAGO DATA PORTAL
Browse
Tutorial
Feedback
Explore Data
Export
API
Share
···

Crimes - 2001 to present Public Safety

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims... More

Updated
Novembe

Data Prov
Chicago P

Download Crimes - 2001 to present

Download Crimes - 2001 to present for offline use in other applications.

[✓ CSV] CSV for Excel

Additional Formats

CSV for Excel (Europe) TSV for Excel
RDF XML
RSS

Featured Content Using this Data

Crimes - 2001 to present - Dashboard
November 26, 2017 916K Views
This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of...

Crimes - 2001 to present - Map
November 26, 2017 45.6K Views
This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of...

Crimes - 2017
November 26, 2
This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of...

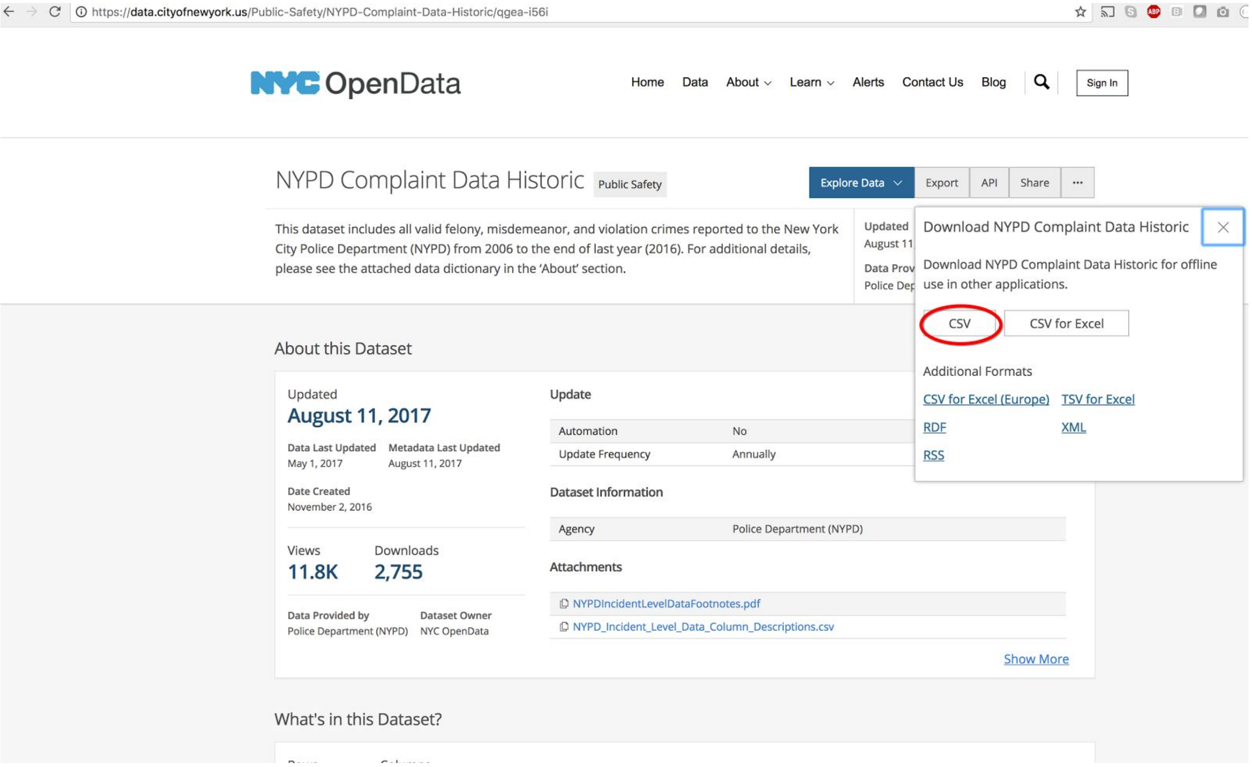
About this Dataset

Updated
November 26, 2017

Data Last Updated November 26, 2017
Metadata Last Updated September 27, 2017

Metadata

Time Period	2001 to present, minus the most recent seven days
Frequency	Data are updated daily.



NYPD Complaint Data Historic Public Safety

This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of last year (2016). For additional details, please see the attached data dictionary in the 'About' section.

Updated
August 11

Data Prov
Police Dep

Download NYPD Complaint Data Historic

Download NYPD Complaint Data Historic for offline use in other applications.

[✓ CSV] CSV for Excel

Additional Formats

CSV for Excel (Europe) TSV for Excel
RDF XML
RSS

About this Dataset

Updated
August 11, 2017

Data Last Updated May 1, 2017
Metadata Last Updated August 11, 2017

Date Created November 2, 2016

Views 11.8K
Downloads 2,755

Data Provided by Police Department (NYPD)
Dataset Owner NYC OpenData

Update

Automation	No
Update Frequency	Annually

Dataset Information

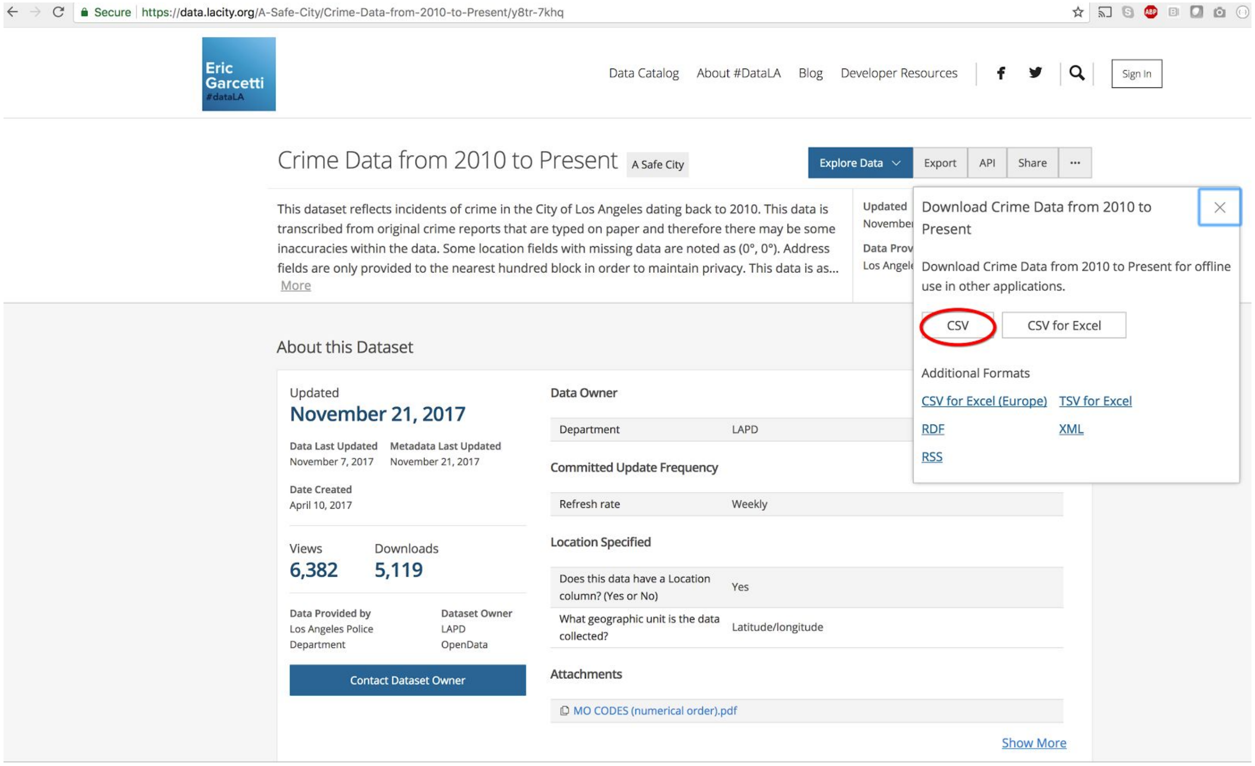
Agency	Police Department (NYPD)

Attachments

NYPDIncidentLevelDataFootnotes.pdf
NYPD_Incident_Level_Data_Column_Descriptions.csv

Show More

What's in this Dataset?



Secure | https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq

Eric Garcetti #dataLA

Data Catalog About #DataLA Blog Developer Resources Sign In

Crime Data from 2010 to Present A Safe City

Explore Data Export API Share ···

This dataset reflects incidents of crime in the City of Los Angeles dating back to 2010. This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data. Some location fields with missing data are noted as (0°, 0°). Address fields are only provided to the nearest hundred block in order to maintain privacy. This data is as... More

Download Crime Data from 2010 to Present

Download Crime Data from 2010 to Present for offline use in other applications.

CSV CSV for Excel

Additional Formats

CSV for Excel (Europe) TSV for Excel

RDF XML

RSS

About this Dataset

Updated
November 21, 2017

Data Last Updated
November 7, 2017

Metadata Last Updated
November 21, 2017

Date Created
April 10, 2017

Views 6,382

Downloads 5,119

Data Provided by
Los Angeles Police Department

Dataset Owner
LAPD OpenData

Contact Dataset Owner

Data Owner

Department	LAPD

Committed Update Frequency

Refresh rate	Weekly

Location Specified

Does this data have a Location column? (Yes or No)	Yes
What geographic unit is the data collected?	Latitude/longitude

Attachments

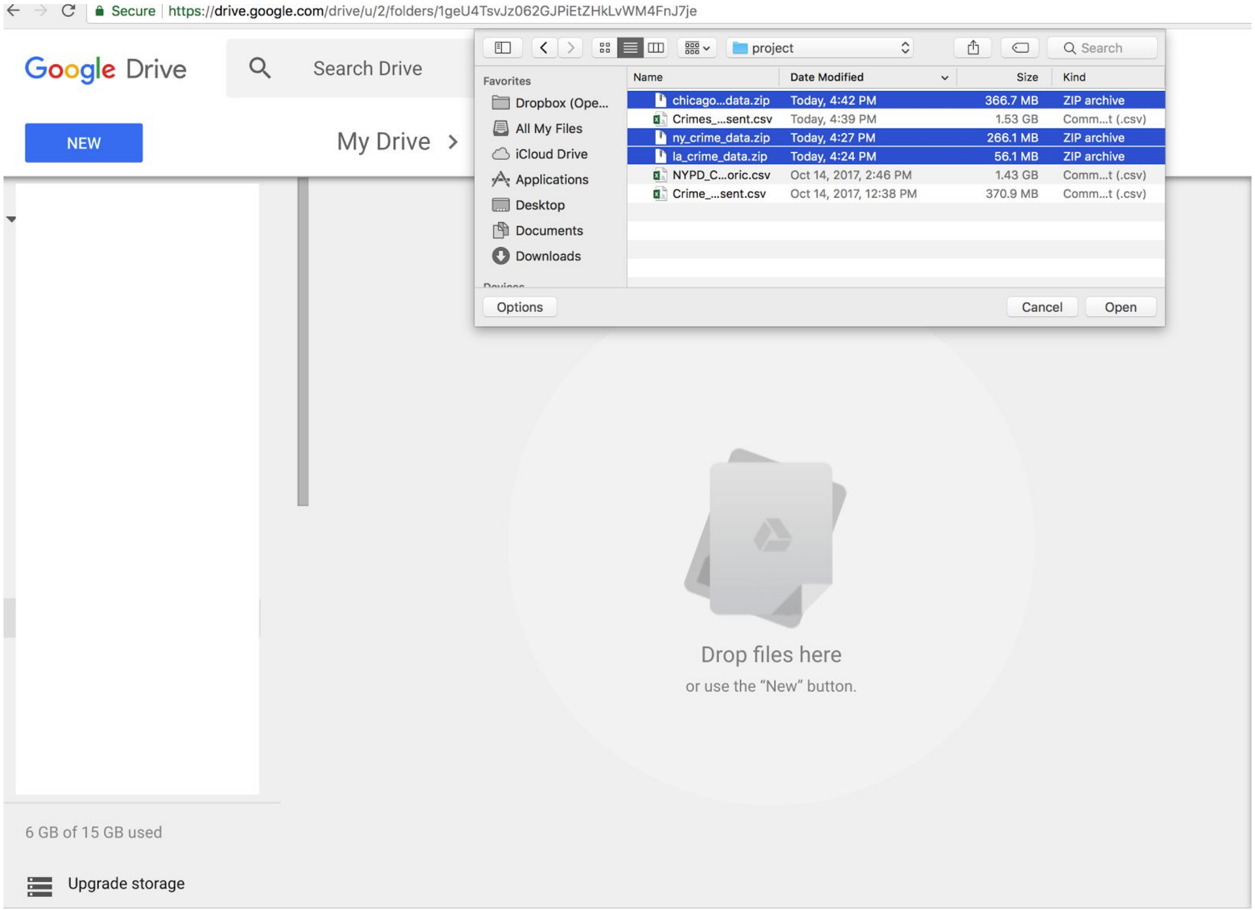
MO CODES (numerical order).pdf

Show More

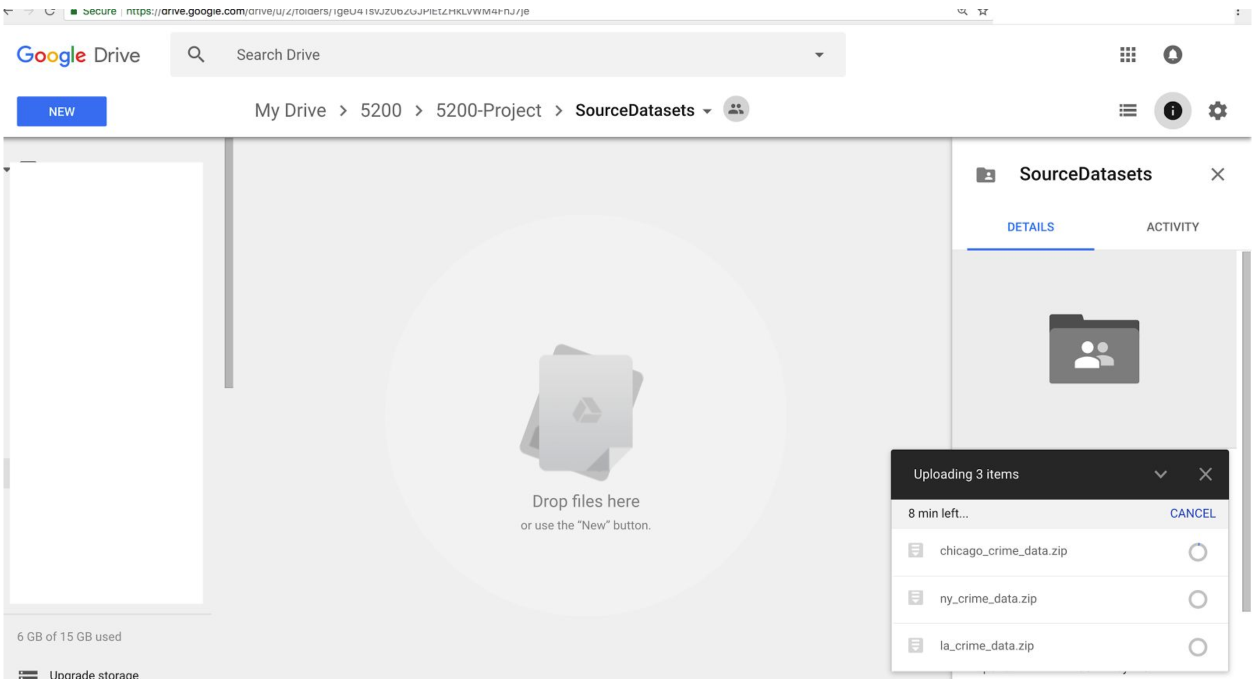
2. Zip the data using either zip utility on windows or using commands on linux bash as follows:
   - zip la_crime_data.zip Crime_Data_from_2010_to_Present.csv
   - zip ny_crime_data.zip NYPD_Complaint_Data_Historic.csv
   - zip chicago_crime_data.zip Crimes_-_2001_to_present.csv

3. Move the zipped file to the google drive as follows:

(continued on next page)



Name	Date Modified	Size	Kind
chicago...data.zip	Today, 4:42 PM	366.7 MB	ZIP archive
Crimes_...sent.csv	Today, 4:39 PM	1.53 GB	Comm...t (.csv)
ny_crime_data.zip	Today, 4:27 PM	266.1 MB	ZIP archive
la_crime_data.zip	Today, 4:24 PM	56.1 MB	ZIP archive
NYPD_C...oric.csv	Oct 14, 2017, 2:46 PM	1.43 GB	Comm...t (.csv)
Crime_...sent.csv	Oct 14, 2017, 12:38 PM	370.9 MB	Comm...t (.csv)



Uploading 3 items

8 min left...

chicago_crime_data.zip

ny_crime_data.zip

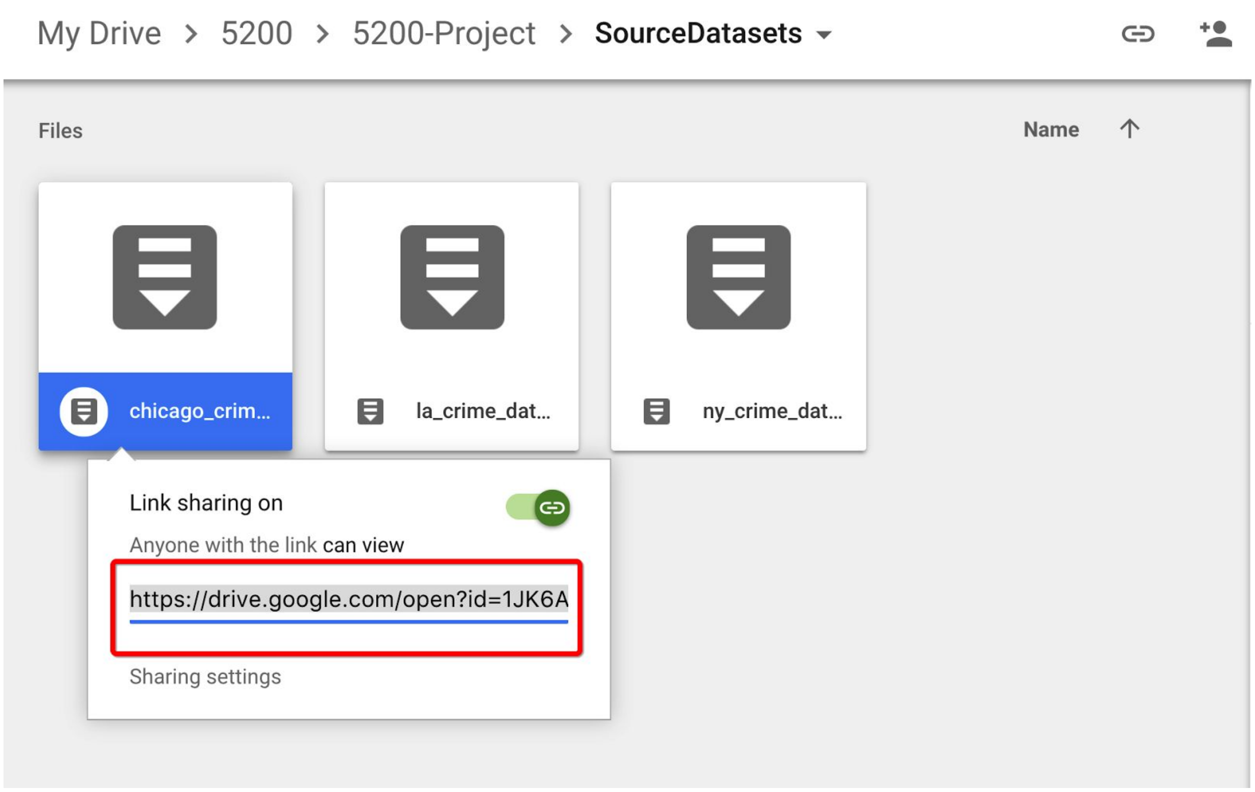
la_crime_data.zip

Step 2 Transform the Data:

While exploring the data after downloading, we realized that all the datasets had locations of crime in “latitude:longitude” format and were missing zipcode information. For our analysis it was critical to have zipcode information in the data. Hence, we wrote a pig udf and used pig scripts to enrich the data with zipcode information.

1. **Download the zip file on the remote machines :**

   a. Get the shareable links to all the raw files uploaded as follows



Link sharing on

Anyone with the link can view

https://drive.google.com/open?id=1JK6A

b. Extract the file id from the shareable link and use it in the step c to replace the highlighted id with red in the links.

c. Google uses different set of APIs for files smaller than 200 MB and files greater than 200 MB.

Hence for LA zipped data we can download it on remote machine as follows:

LA (zip file of raw data is less than 200 MB):

d. wget -O la_crime_data.zip "https://drive.google.com/uc?export=download&id=15UzDOgq1f_qjv3nEblgrnSIUn9mQyK5w"

For NY and Chicago zipped datasets we have to use curl instead as follows:

<u>NY (zip file of raw data is greater than 200 MB):</u>

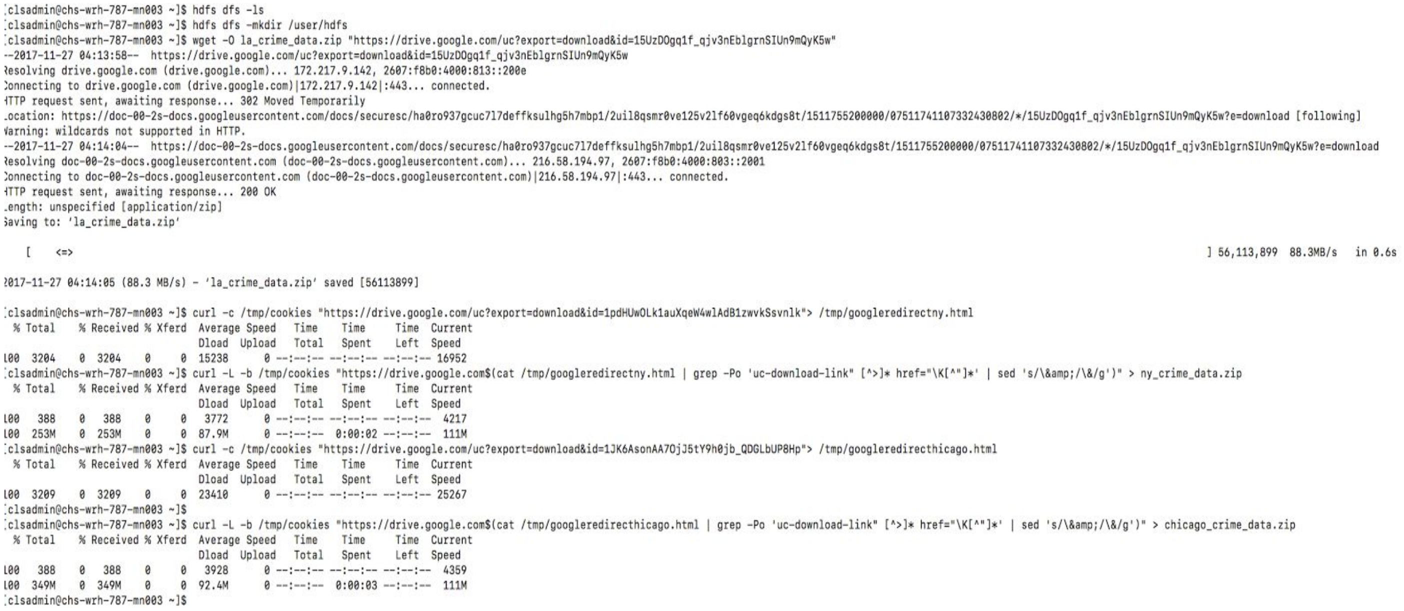
e. curl -c /tmp/cookies "https://drive.google.com/uc?export=download&id=1pdHUwOLk1auXqeW4wlAdB1zwvkSsvnlk"> /tmp/googleredirectny.html

f. curl -L -b /tmp/cookies "https://drive.google.com$(cat /tmp/googleredirectny.html | grep -Po 'uc-download-link" [^>]* href="\K[^"]*' | sed 's/\&amp;/\&/g')" > ny_crime_data.zip

<u>Chicago(zip file of raw data is greater than 200 MB):</u>

g. curl -c /tmp/cookies "https://drive.google.com/uc?export=download&id=1JK6AsonAA7OjJ5tY9h0jb_QDGLbUP8Hp"> /tmp/googleredirecthicago.html

h. curl -L -b /tmp/cookies "https://drive.google.com$(cat /tmp/googleredirecthicago.html | grep -Po 'uc-download-link" [^>]* href="\K[^"]*' | sed 's/\&amp;/\&/g')" > chicago_crime_data.zip

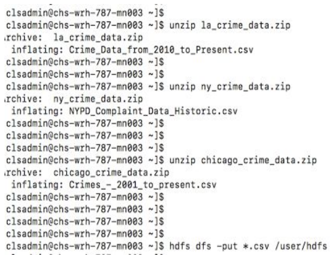


```
[clsadmin@chs-wrh-787-mn003 ~]$ hdfs dfs -ls
[clsadmin@chs-wrh-787-mn003 ~]$ hdfs dfs -mkdir /user/hdfs
[clsadmin@chs-wrh-787-mn003 ~]$ wget -O la_crime_data.zip "https://drive.google.com/uc?export=download&id=15UzDOgq1f_qjv3nEblgrnSIUn9mQyK5w"
--2017-11-27 04:13:58--  https://drive.google.com/uc?export=download&id=15UzDOgq1f_qjv3nEblgrnSIUn9mQyK5w
Resolving drive.google.com (drive.google.com)... 172.217.9.142, 2607:f8b0:4000:813::200e
Connecting to drive.google.com (drive.google.com)|172.217.9.142|:443... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily
Location: https://doc-00-2s-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/2uil8qsmr0ve125v2lf60vgeq6kdgs8t/1511755200000/07511741107332430802/*/15UzDOgq1f_qjv3nEblgrnSIUn9mQyK5w?e=download [following]
Warning: wildcards not supported in HTTP.
--2017-11-27 04:14:04--  https://doc-00-2s-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/2uil8qsmr0ve125v2lf60vgeq6kdgs8t/1511755200000/07511741107332430802/*/15UzDOgq1f_qjv3nEblgrnSIUn9mQyK5w?e=download
Resolving doc-00-2s-docs.googleusercontent.com (doc-00-2s-docs.googleusercontent.com)... 216.58.194.97, 2607:f8b0:4000:803::2001
Connecting to doc-00-2s-docs.googleusercontent.com (doc-00-2s-docs.googleusercontent.com)|216.58.194.97|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [application/zip]
Saving to: 'la_crime_data.zip'

    [    <=>                                                                                                                                         ] 56,113,899  88.3MB/s   in 0.6s

2017-11-27 04:14:05 (88.3 MB/s) - 'la_crime_data.zip' saved [56113899]

[clsadmin@chs-wrh-787-mn003 ~]$ curl -c /tmp/cookies "https://drive.google.com/uc?export=download&id=1pdHUwOLk1auXqeW4wlAdB1zwvkSsvnlk"> /tmp/googleredirectny.html
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  3204    0  3204    0     0  15238      0 --:--:-- --:--:-- --:--:-- 16952
[clsadmin@chs-wrh-787-mn003 ~]$ curl -L -b /tmp/cookies "https://drive.google.com$(cat /tmp/googleredirectny.html | grep -Po 'uc-download-link" [^>]* href="\K[^"]*' | sed 's/\&amp;/\&/g')" > ny_crime_data.zip
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   388    0   388    0     0   3772      0 --:--:-- --:--:-- --:--:--  4217
100  253M    0  253M    0     0  87.9M      0 --:--:--  0:00:02 --:--:--  111M
[clsadmin@chs-wrh-787-mn003 ~]$ curl -c /tmp/cookies "https://drive.google.com/uc?export=download&id=1JK6AsonAA7OjJ5tY9h0jb_QDGLbUP8Hp"> /tmp/googleredirecthicago.html
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  3209    0  3209    0     0  23410      0 --:--:-- --:--:-- --:--:-- 25267
[clsadmin@chs-wrh-787-mn003 ~]$
[clsadmin@chs-wrh-787-mn003 ~]$ curl -L -b /tmp/cookies "https://drive.google.com$(cat /tmp/googleredirecthicago.html | grep -Po 'uc-download-link" [^>]* href="\K[^"]*' | sed 's/\&amp;/\&/g')" > chicago_crime_data.zip
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   388    0   388    0     0   3928      0 --:--:-- --:--:-- --:--:--  4359
100  349M    0  349M    0     0  92.4M      0 --:--:--  0:00:03 --:--:--  111M
[clsadmin@chs-wrh-787-mn003 ~]$
```

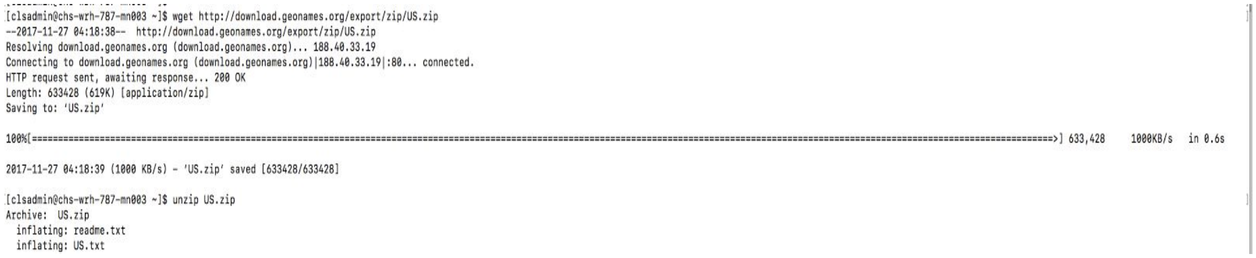
2. Load raw Data into Hdfs for transforming :

a. hdfs dfs -mkdir /user/hdfs
b. unzip la_crime_data.zip

c. unzip ny_crime_data.zip
d. unzip chicago_crime_data.zip
e. hdfs dfs -put *.csv /user/hdfs
f. wget http://download.geonames.org/export/zip/US.zip
g. unzip US.zip
h. hdfs dfs -put US.txt /user/hdfs/



```
clsadmin@chs-wrh-787-mn003 ~]$
clsadmin@chs-wrh-787-mn003 ~]$ unzip la_crime_data.zip
rchive:  la_crime_data.zip
 inflating: Crime_Data_from_2010_to_Present.csv
clsadmin@chs-wrh-787-mn003 ~]$
clsadmin@chs-wrh-787-mn003 ~]$
clsadmin@chs-wrh-787-mn003 ~]$ unzip ny_crime_data.zip
rchive:  ny_crime_data.zip
 inflating: NYPD_Complaint_Data_Historic.csv
clsadmin@chs-wrh-787-mn003 ~]$
clsadmin@chs-wrh-787-mn003 ~]$
clsadmin@chs-wrh-787-mn003 ~]$
clsadmin@chs-wrh-787-mn003 ~]$ unzip chicago_crime_data.zip
rchive:  chicago_crime_data.zip
 inflating: Crimes_-_2001_to_present.csv
clsadmin@chs-wrh-787-mn003 ~]$
clsadmin@chs-wrh-787-mn003 ~]$
clsadmin@chs-wrh-787-mn003 ~]$
clsadmin@chs-wrh-787-mn003 ~]$
clsadmin@chs-wrh-787-mn003 ~]$ hdfs dfs -put *.csv /user/hdfs
```



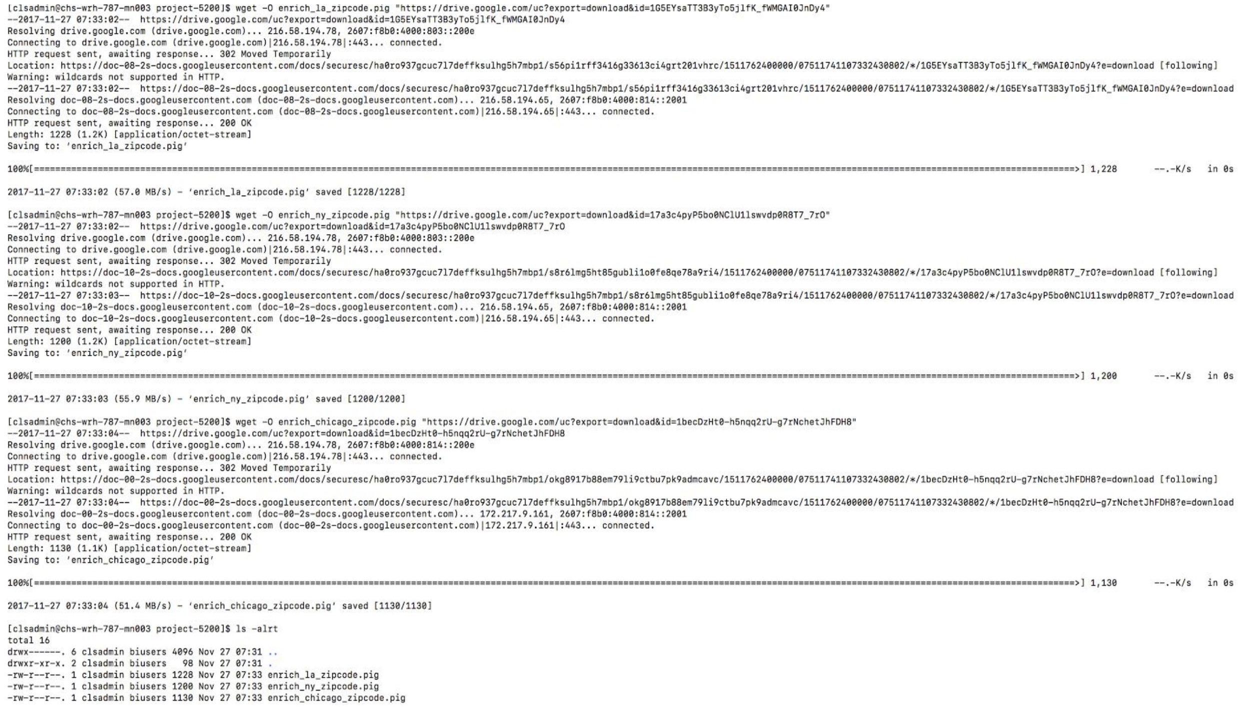
```
[clsadmin@chs-wrh-787-mn003 ~]$ wget http://download.geonames.org/export/zip/US.zip
--2017-11-27 04:18:38--  http://download.geonames.org/export/zip/US.zip
Resolving download.geonames.org (download.geonames.org)... 188.40.33.19
Connecting to download.geonames.org (download.geonames.org)|188.40.33.19|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 633428 (619K) [application/zip]
Saving to: 'US.zip'

100%[=====================================================================================================================================================>] 633,428     1000KB/s   in 0.6s

2017-11-27 04:18:39 (1000 KB/s) - 'US.zip' saved [633428/633428]

[clsadmin@chs-wrh-787-mn003 ~]$ unzip US.zip
Archive:  US.zip
  inflating: readme.txt
  inflating: US.txt
```

```
[clsadmin@chs-wrh-787-mn003 ~]$
[clsadmin@chs-wrh-787-mn003 ~]$ hdfs dfs -put US.txt /user/hdfs/
```

3. Run Transformations on the loaded data:

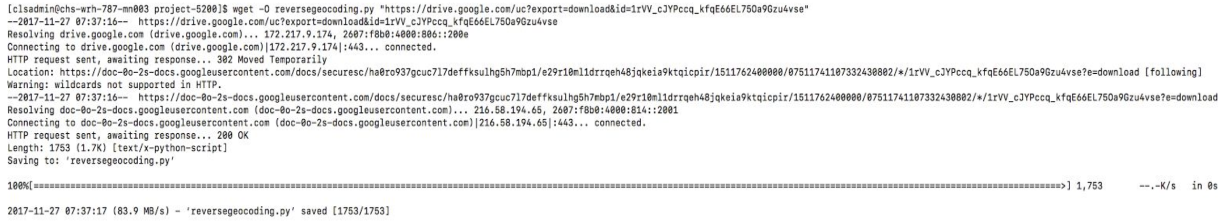
1. wget -O enrich_la_zipcode.pig "https://drive.google.com/uc?export=download&id=1G5EYsaTT3B3yTo5jlfK_fWMGAI0JnDy4"
2. wget -O enrich_ny_zipcode.pig "https://drive.google.com/uc?export=download&id=17a3c4pyP5bo0NClU1lswvdp0R8T7_7rO"
3. wget -O enrich_chicago_zipcode.pig "https://drive.google.com/uc?export=download&id=1becDzHt0-h5nqq2rU-g7rNchetJhFDH8"

(continued on next page)



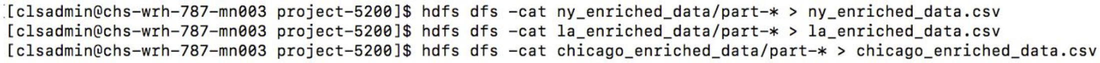
```
[clsadmin@chs-wrh-787-mn003 project-5200]$ wget -O enrich_la_zipcode.pig "https://drive.google.com/uc?export=download&id=1G5EYsaTT3B3yTo5jlfK_fWMGAI0JnDy4"
--2017-11-27 07:33:02--  https://drive.google.com/uc?export=download&id=1G5EYsaTT3B3yTo5jlfK_fWMGAI0JnDy4
Resolving drive.google.com (drive.google.com)... 216.58.194.78, 2607:f8b0:4000:803::200e
Connecting to drive.google.com (drive.google.com)|216.58.194.78|:443... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily
Location: https://doc-08-2s-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/s56pi1rff3416g33613ci4grt201vhrc/1511762400000/07511741107332430802/*/1G5EYsaTT3B3yTo5jlfK_fWMGAI0JnDy4?e=download [following]
Warning: wildcards not supported in HTTP.
--2017-11-27 07:33:02--  https://doc-08-2s-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/s56pi1rff3416g33613ci4grt201vhrc/1511762400000/07511741107332430802/*/1G5EYsaTT3B3yTo5jlfK_fWMGAI0JnDy4?e=download
Resolving doc-08-2s-docs.googleusercontent.com (doc-08-2s-docs.googleusercontent.com)... 216.58.194.65, 2607:f8b0:4000:814::2001
Connecting to doc-08-2s-docs.googleusercontent.com (doc-08-2s-docs.googleusercontent.com)|216.58.194.65|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1228 (1.2K) [application/octet-stream]
Saving to: 'enrich_la_zipcode.pig'

100%[==========================================================================================================================================================>] 1,228       --.-K/s   in 0s

2017-11-27 07:33:02 (57.0 MB/s) - 'enrich_la_zipcode.pig' saved [1228/1228]

[clsadmin@chs-wrh-787-mn003 project-5200]$ wget -O enrich_ny_zipcode.pig "https://drive.google.com/uc?export=download&id=17a3c4pyP5bo0NClU1lswvdp0R8T7_7rO"
--2017-11-27 07:33:02--  https://drive.google.com/uc?export=download&id=17a3c4pyP5bo0NClU1lswvdp0R8T7_7rO
Resolving drive.google.com (drive.google.com)... 216.58.194.78, 2607:f8b0:4000:803::200e
Connecting to drive.google.com (drive.google.com)|216.58.194.78|:443... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily
Location: https://doc-10-2s-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/s8r6lmg5ht85gubli1o0fe8qe78a9ri4/1511762400000/07511741107332430802/*/17a3c4pyP5bo0NClU1lswvdp0R8T7_7rO?e=download [following]
Warning: wildcards not supported in HTTP.
--2017-11-27 07:33:03--  https://doc-10-2s-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/s8r6lmg5ht85gubli1o0fe8qe78a9ri4/1511762400000/07511741107332430802/*/17a3c4pyP5bo0NClU1lswvdp0R8T7_7rO?e=download
Resolving doc-10-2s-docs.googleusercontent.com (doc-10-2s-docs.googleusercontent.com)... 216.58.194.65, 2607:f8b0:4000:814::2001
Connecting to doc-10-2s-docs.googleusercontent.com (doc-10-2s-docs.googleusercontent.com)|216.58.194.65|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1200 (1.2K) [application/octet-stream]
Saving to: 'enrich_ny_zipcode.pig'

100%[==========================================================================================================================================================>] 1,200       --.-K/s   in 0s

2017-11-27 07:33:03 (55.9 MB/s) - 'enrich_ny_zipcode.pig' saved [1200/1200]

[clsadmin@chs-wrh-787-mn003 project-5200]$ wget -O enrich_chicago_zipcode.pig "https://drive.google.com/uc?export=download&id=1becDzHt0-h5nqq2rU-g7rNchetJhFDH8"
--2017-11-27 07:33:04--  https://drive.google.com/uc?export=download&id=1becDzHt0-h5nqq2rU-g7rNchetJhFDH8
Resolving drive.google.com (drive.google.com)... 216.58.194.78, 2607:f8b0:4000:814::200e
Connecting to drive.google.com (drive.google.com)|216.58.194.78|:443... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily
Location: https://doc-00-2s-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/okg8917b88em79li9ctbu7pk9admcavc/1511762400000/07511741107332430802/*/1becDzHt0-h5nqq2rU-g7rNchetJhFDH8?e=download [following]
Warning: wildcards not supported in HTTP.
--2017-11-27 07:33:04--  https://doc-00-2s-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/okg8917b88em79li9ctbu7pk9admcavc/1511762400000/07511741107332430802/*/1becDzHt0-h5nqq2rU-g7rNchetJhFDH8?e=download
Resolving doc-00-2s-docs.googleusercontent.com (doc-00-2s-docs.googleusercontent.com)... 172.217.9.161, 2607:f8b0:4000:814::2001
Connecting to doc-00-2s-docs.googleusercontent.com (doc-00-2s-docs.googleusercontent.com)|172.217.9.161|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1130 (1.1K) [application/octet-stream]
Saving to: 'enrich_chicago_zipcode.pig'

100%[==========================================================================================================================================================>] 1,130       --.-K/s   in 0s

2017-11-27 07:33:04 (51.4 MB/s) - 'enrich_chicago_zipcode.pig' saved [1130/1130]

[clsadmin@chs-wrh-787-mn003 project-5200]$ ls -alrt
total 16
drwx------. 6 clsadmin biusers 4096 Nov 27 07:31 ..
drwxr-xr-x. 2 clsadmin biusers   98 Nov 27 07:31 .
-rw-r--r--. 1 clsadmin biusers 1228 Nov 27 07:33 enrich_la_zipcode.pig
-rw-r--r--. 1 clsadmin biusers 1200 Nov 27 07:33 enrich_ny_zipcode.pig
-rw-r--r--. 1 clsadmin biusers 1130 Nov 27 07:33 enrich_chicago_zipcode.pig
```

4. wget -O reversegeocoding.py "https://drive.google.com/uc?export=download&id=1rVV_cJYPccq_kfqE66EL75Oa9Gzu4vse"
5. hdfs dfs -put reversegeocoding.py /user/hdfs/



```
[clsadmin@chs-wrh-787-mn003 project-5200]$ wget -O reversegeocoding.py "https://drive.google.com/uc?export=download&id=1rVV_cJYPccq_kfqE66EL75Oa9Gzu4vse"
--2017-11-27 07:37:16--  https://drive.google.com/uc?export=download&id=1rVV_cJYPccq_kfqE66EL75Oa9Gzu4vse
Resolving drive.google.com (drive.google.com)... 172.217.9.174, 2607:f8b0:4000:806::200e
Connecting to drive.google.com (drive.google.com)|172.217.9.174|:443... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily
Location: https://doc-0o-2s-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/e29r10ml1drrqeh48jqkeia9ktqicpir/1511762400000/07511741107332430802/*/1rVV_cJYPccq_kfqE66EL75Oa9Gzu4vse?e=download [following]
Warning: wildcards not supported in HTTP.
--2017-11-27 07:37:16--  https://doc-0o-2s-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/e29r10ml1drrqeh48jqkeia9ktqicpir/1511762400000/07511741107332430802/*/1rVV_cJYPccq_kfqE66EL75Oa9Gzu4vse?e=download
Resolving doc-0o-2s-docs.googleusercontent.com (doc-0o-2s-docs.googleusercontent.com)... 216.58.194.65, 2607:f8b0:4000:814::2001
Connecting to doc-0o-2s-docs.googleusercontent.com (doc-0o-2s-docs.googleusercontent.com)|216.58.194.65|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1753 (1.7K) [text/x-python-script]
Saving to: 'reversegeocoding.py'

100%[==========================================================================================================================================================>] 1,753       --.-K/s   in 0s

2017-11-27 07:37:17 (83.9 MB/s) - 'reversegeocoding.py' saved [1753/1753]
```

6. pig enrich_la_zipcode.pig
7. pig enrich_ny_zipcode.pig
8. pig enrich_chicago_zipcode.pig
9. hdfs dfs -cat la_enriched_data/part-* > la_enriched_data.csv
10. hdfs dfs -cat ny_enriched_data/part-* > ny_enriched_data.csv
11. hdfs dfs -cat chicago_enriched_data/part-* > chicago_enriched_data.csv



```
[clsadmin@chs-wrh-787-mn003 project-5200]$ hdfs dfs -cat ny_enriched_data/part-* > ny_enriched_data.csv
[clsadmin@chs-wrh-787-mn003 project-5200]$ hdfs dfs -cat la_enriched_data/part-* > la_enriched_data.csv
[clsadmin@chs-wrh-787-mn003 project-5200]$ hdfs dfs -cat chicago_enriched_data/part-* > chicago_enriched_data.csv
```

12. We have zipped up files individually and we posted it on google drive to share across with other group mates.
13. La_enriched_data: https://drive.google.com/open?id=1pyMmySMP5MvaEBcuLQ3pA8CSyP_JTqcT
14. Chicago_enriched_data: https://drive.google.com/open?id=1IKFo8GFWu4gshICahcaQlwT7UcXVCCg2

15. NY_enriched_data: https://drive.google.com/open?id=1Pqk6x-W4MW4a0xVGOeLztVYkXnkiEauu

Step 3: Load the enriched data into hive and analyze:

Note: replace the username (ngupta8) with your respective username.

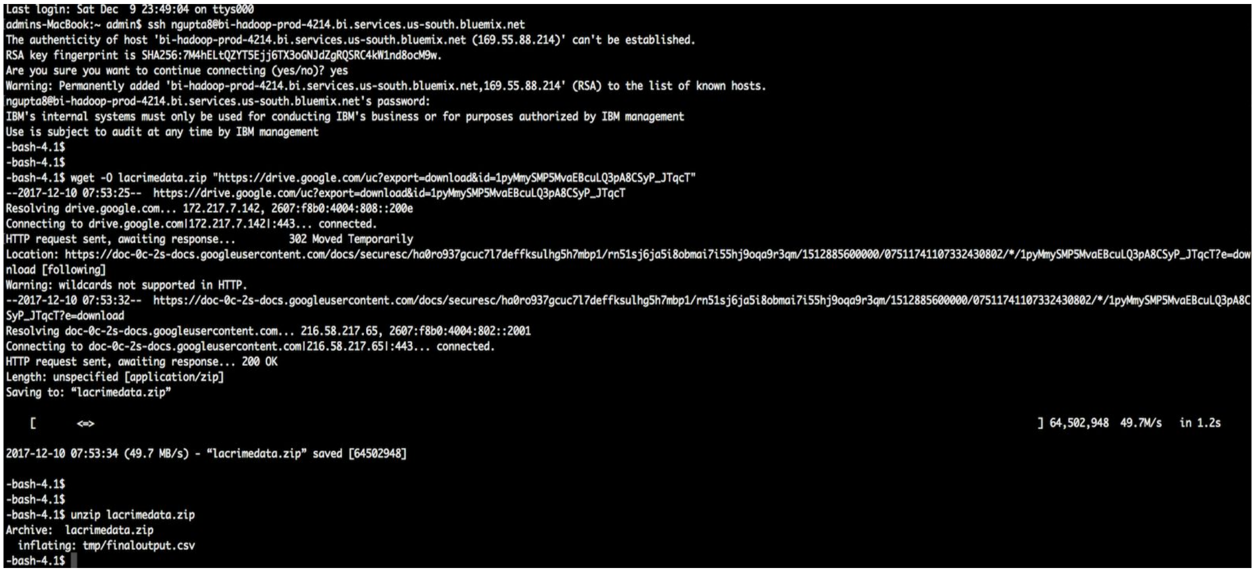
For Los Angeles City:

1. Download the data

```
wget -O lacrimedata.zip
"https://drive.google.com/uc?export=download&id=1pyMmySMP5MvaEBcuLQ3p
A8CSyP_JTqcT"
```

2. Unzip the file

```
unzip lacrimedata.zip
```



```
Last login: Sat Dec  9 23:49:04 on ttys000
admins-MacBook:~ admin$ ssh ngupta8@bi-hadoop-prod-4214.bi.services.us-south.bluemix.net
The authenticity of host 'bi-hadoop-prod-4214.bi.services.us-south.bluemix.net (169.55.88.214)' can't be established.
RSA key fingerprint is SHA256:7M4hELtQZYT5Ejj6TX3oGNJdZgRQSRC4kW1nd8ocM9w.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'bi-hadoop-prod-4214.bi.services.us-south.bluemix.net,169.55.88.214' (RSA) to the list of known hosts.
ngupta8@bi-hadoop-prod-4214.bi.services.us-south.bluemix.net's password:
IBM's internal systems must only be used for conducting IBM's business or for purposes authorized by IBM management
Use is subject to audit at any time by IBM management
-bash-4.1$
-bash-4.1$
-bash-4.1$ wget -O lacrimedata.zip "https://drive.google.com/uc?export=download&id=1pyMmySMP5MvaEBcuLQ3pA8CSyP_JTqcT"
--2017-12-10 07:53:25--  https://drive.google.com/uc?export=download&id=1pyMmySMP5MvaEBcuLQ3pA8CSyP_JTqcT
Resolving drive.google.com... 172.217.7.142, 2607:f8b0:4004:808::200e
Connecting to drive.google.com|172.217.7.142|:443... connected.
HTTP request sent, awaiting response...        302 Moved Temporarily
Location: https://doc-0c-2s-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/rn51sj6ja5i8obmai7i55hj9oqa9r3qm/1512885600000/07511741107332430802/*/1pyMmySMP5MvaEBcuLQ3pA8CSyP_JTqcT?e=dow
nload [following]
Warning: wildcards not supported in HTTP.
--2017-12-10 07:53:32--  https://doc-0c-2s-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/rn51sj6ja5i8obmai7i55hj9oqa9r3qm/1512885600000/07511741107332430802/*/1pyMmySMP5MvaEBcuLQ3pA8C
SyP_JTqcT?e=download
Resolving doc-0c-2s-docs.googleusercontent.com... 216.58.217.65, 2607:f8b0:4004:802::2001
Connecting to doc-0c-2s-docs.googleusercontent.com|216.58.217.65|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [application/zip]
Saving to: “lacrimedata.zip”

    [      <=>                                                                                                          ] 64,502,948  49.7M/s   in 1.2s

2017-12-10 07:53:34 (49.7 MB/s) - “lacrimedata.zip” saved [64502948]

-bash-4.1$
-bash-4.1$
-bash-4.1$ unzip lacrimedata.zip
Archive:  lacrimedata.zip
  inflating: tmp/finaloutput.csv
-bash-4.1$
```

3. Open hive terminal using: hive
4. Run Following commands on hive shell:
    a. Create the database named crime_data

```
create database if not exists crime_data;
```

    b. Check the created database

```
show databases;
```

c. Select the database

```
use crime_data;
```

d. Create the table la_crime_data

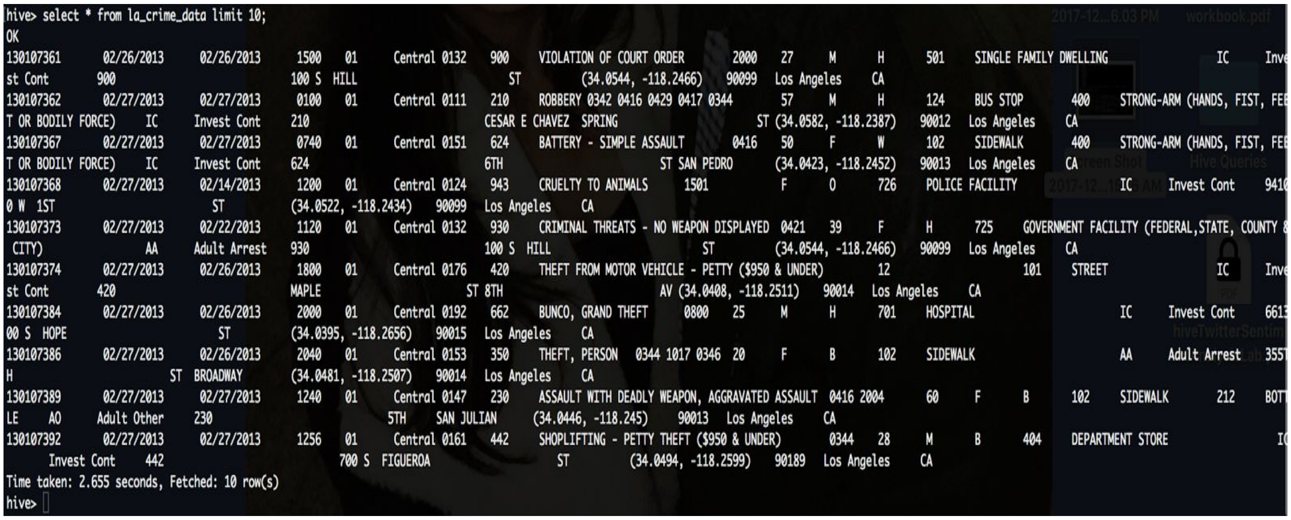
```
CREATE TABLE IF NOT EXISTS la_crime_data(
	dr_no STRING,
	date_rptd DATE,
	date_occ DATE,
	time_occ STRING,
	area_id STRING,
	area_name STRING,
	rpt_dist_no STRING,
	crm_cd STRING,
	crm_cd_desc STRING,
	mocodes STRING,
	vict_age STRING,
	vict_sex STRING,
	vict_descent STRING,
	premis_cd STRING,
	premis_desc STRING,
	weapon_used_cd STRING,
	weapon_desc STRING,
	status STRING,
	status_desc STRING,
	crm_cd_1 STRING,
	crm_cd_2 STRING,
	crm_cd_3 STRING,
	crm_cd_4 STRING,
	location STRING,
	cross_street STRING,
	location_1 STRING,
	zipcode STRING,
	city STRING,
	state STRING)ROW FORMAT
SERDE'org.apache.hadoop.hive.serde2.OpenCSVSerde'WITH SERDEPROPERTIES
("separatorChar" = ",","quoteChar" = "\"")TBLPROPERTIES
("skip.header.line.count"="1");
```

e. Load the data into the table la_crime_data

```
load data local inpath '/home/ngupta8/tmp/finaloutput.csv' into table
la_crime_data;
```

f. Check whether data is uploaded properly or not

```
select * from la_crime_data limit 10;
```



```
hive> select * from la_crime_data limit 10;
OK
130107361       02/26/2013      02/26/2013      1500    01      Central 0132    900     VIOLATION OF COURT ORDER        2000    27      M       H       501     SINGLE FAMILY DWELLING                  IC      Inve
st Cont         900                             100 S  HILL                             ST              (34.0544, -118.2466)    90099   Los Angeles     CA
130107362       02/27/2013      02/27/2013      0100    01      Central 0111    210     ROBBERY 0342 0416 0429 0417 0344    57      M       H       124     BUS STOP        400     STRONG-ARM (HANDS, FIST, FEE
T OR BODILY FORCE)      IC      Invest Cont     210                             CESAR E CHAVEZ  SPRING          ST (34.0582, -118.2387)     90012   Los Angeles     CA
130107367       02/27/2013      02/27/2013      0740    01      Central 0151    624     BATTERY - SIMPLE ASSAULT        0416    50      F       W       102     SIDEWALK        400     STRONG-ARM (HANDS, FIST, FEE
T OR BODILY FORCE)      IC      Invest Cont     624                             6TH                     ST SAN PEDRO    (34.0423, -118.2452)    90013   Los Angeles     CA
130107368       02/27/2013      02/14/2013      1200    01      Central 0124    943     CRUELTY TO ANIMALS      1501            F       O       726     POLICE FACILITY         IC      Invest Cont     941
0 W  1ST                ST              (34.0522, -118.2434)    90099   Los Angeles     CA
130107373       02/27/2013      02/22/2013      1120    01      Central 0132    930     CRIMINAL THREATS - NO WEAPON DISPLAYED  0421    39      F       H       725     GOVERNMENT FACILITY (FEDERAL,STATE, COUNTY &
 CITY)                  AA      Adult Arrest    930                             100 S  HILL                     ST              (34.0544, -118.2466)    90099   Los Angeles     CA
130107374       02/27/2013      02/26/2013      1800    01      Central 0176    420     THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)         12                      101     STREET                  IC      Inve
st Cont         420                             MAPLE                   ST 8TH                          AV (34.0408, -118.2511)     90014   Los Angeles     CA
130107384       02/27/2013      02/26/2013      2000    01      Central 0192    662     BUNCO, GRAND THEFT      0800    25      M       H       701     HOSPITAL                        IC      Invest Cont     661
00 S  HOPE              ST              (34.0395, -118.2656)    90015   Los Angeles     CA
130107386       02/27/2013      02/26/2013      2040    01      Central 0153    350     THEFT, PERSON   0344 1017 0346  20      F       B       102     SIDEWALK                        AA      Adult Arrest    355
H                       ST  BROADWAY    (34.0481, -118.2507)    90014   Los Angeles     CA
130107389       02/27/2013      02/27/2013      1240    01      Central 0147    230     ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT  0416 2004       60      F       B       102     SIDEWALK        212     BOTT
LE      AO      Adult Other     230                     5TH     SAN JULIAN      (34.0446, -118.245)     90013   Los Angeles     CA
130107392       02/27/2013      02/27/2013      1256    01      Central 0161    442     SHOPLIFTING - PETTY THEFT ($950 & UNDER)        0344    28      M       B       404     DEPARTMENT STORE                I
        Invest Cont     442                     700 S  FIGUEROA                 ST              (34.0494, -118.2599)    90189   Los Angeles     CA
Time taken: 2.655 seconds, Fetched: 10 row(s)
hive> 
```

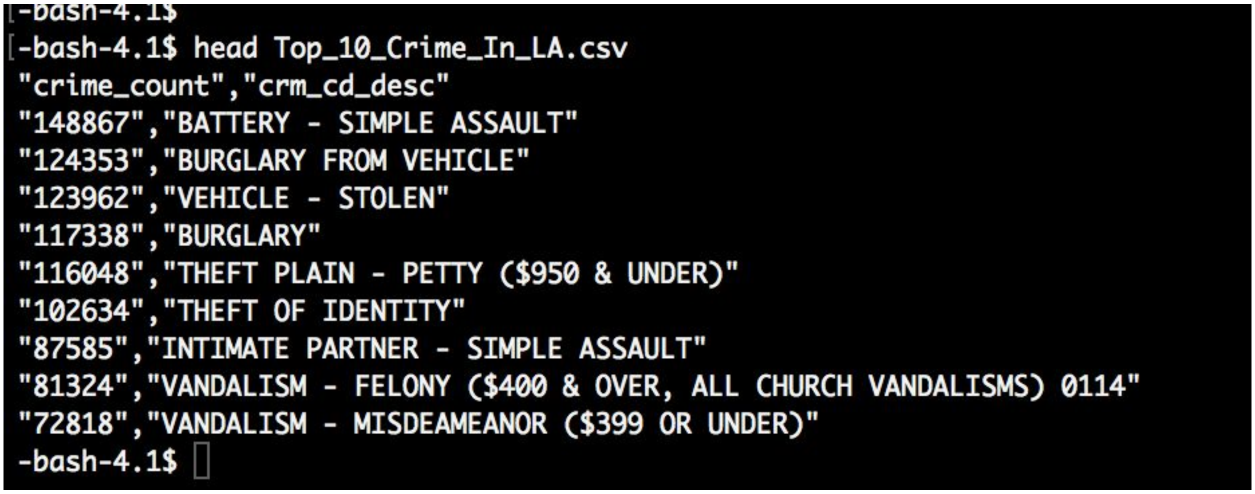
5. Run Following commands from the bash shell and your home directory for instance /home/ngupta8 (NOT from Hive shell).

Query -1: Top 10 crimes:

```
hive -e 'set hive.cli.print.header=true;use crime_data;Select
count(*) as crime_count,crm_cd_desc from la_crime_data group by
crm_cd_desc order by crime_count DESC limit 10;'| perl -lpe
's/"/\\"/g; s/^|$/"/g; s/\t/","/g' > Top_10_Crime_In_LA.csv;
```

Check the file Top_10_Crime_In_LA.csv:

```
head Top_10_Crime_In_LA.csv
```



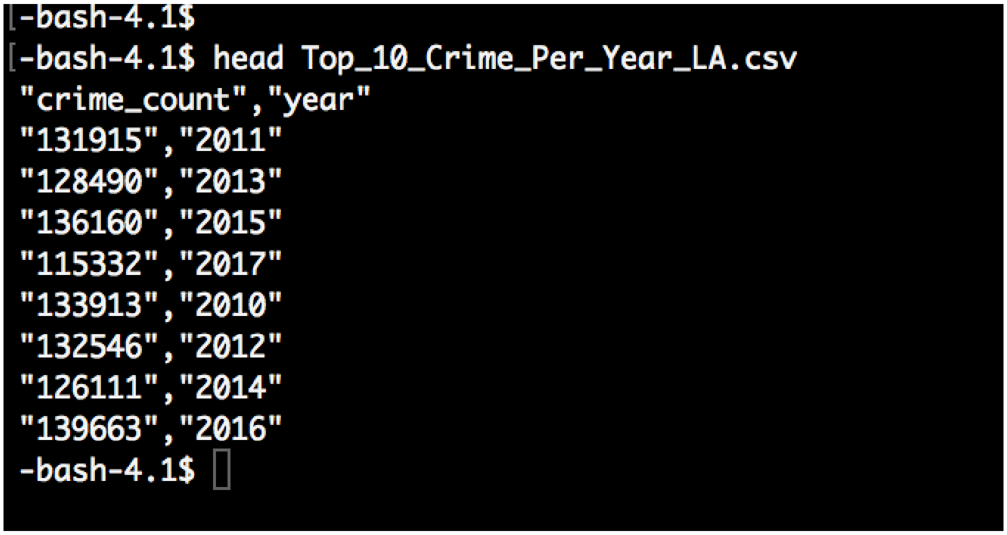
```
-bash-4.1$
-bash-4.1$ head Top_10_Crime_In_LA.csv
"crime_count","crm_cd_desc"
"148867","BATTERY - SIMPLE ASSAULT"
"124353","BURGLARY FROM VEHICLE"
"123962","VEHICLE - STOLEN"
"117338","BURGLARY"
"116048","THEFT PLAIN - PETTY ($950 & UNDER)"
"102634","THEFT OF IDENTITY"
"87585","INTIMATE PARTNER - SIMPLE ASSAULT"
"81324","VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS) 0114"
"72818","VANDALISM - MISDEAMEANOR ($399 OR UNDER)"
-bash-4.1$ 
```

Query -2: Top 10 crimes per year:

```
hive -e 'set hive.cli.print.header=true ; SET
hive.groupby.orderby.position.alias=true ; use crime_data; select
count(*) as
crime_count,from_unixtime(unix_timestamp(date_occ,"mm/dd/yyyy"),"Y")
as year from la_crime_data where crm_cd_desc is not NULL and
crm_cd_desc != "" AND crm_cd_desc IN(
"BATTERY - SIMPLE ASSAULT",
"BURGLARY FROM VEHICLE",
"VEHICLE - STOLEN",
"BURGLARY",
"THEFT PLAIN - PETTY ($950 & UNDER)",
"THEFT OF IDENTITY",
"INTIMATE PARTNER - SIMPLE ASSAULT",
"VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS) 0114",
"VANDALISM - MISDEAMEANOR ($399 OR UNDER)",
"ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT") group by 2'| perl
-lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g' >Top_10_Crime_Per_Year_LA.csv
;
```

Check the file Top_10_Crime_Per_Year_LA.csv:

```
head Top_10_Crime_Per_Year_LA.csv
```



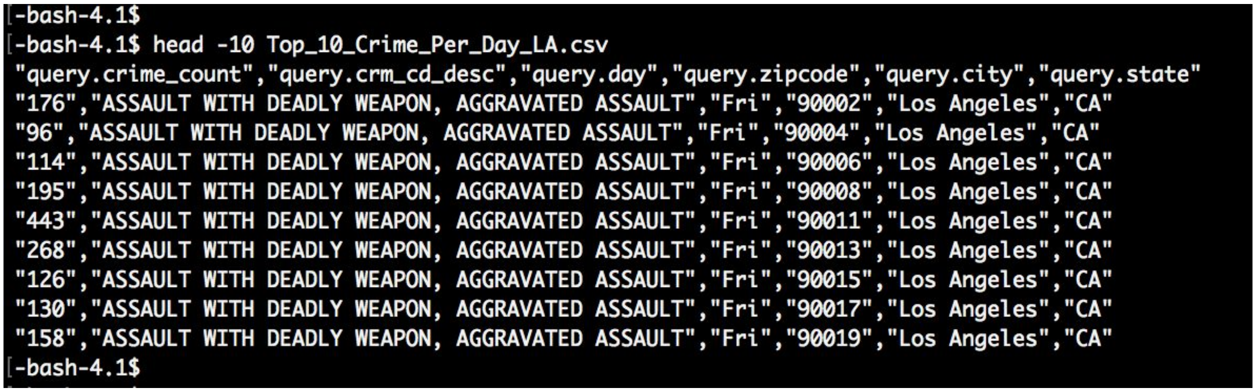
```
-bash-4.1$
-bash-4.1$ head Top_10_Crime_Per_Year_LA.csv
"crime_count","year"
"131915","2011"
"128490","2013"
"136160","2015"
"115332","2017"
"133913","2010"
"132546","2012"
"126111","2014"
"139663","2016"
-bash-4.1$
```

Query -3: Top 10 crimes per day:

```
hive -e 'set hive.cli.print.header=true ; SET
hive.groupby.orderby.position.alias=true ; use crime_data ;with query
as (Select count(*) as
crime_count,crm_cd_desc,from_unixtime(unix_timestamp(date_occ,"mm/dd/
yyyy"),"E") as day,zipcode,city,state from la_crime_data where
crm_cd_desc is not NULL and crm_cd_desc != ""  AND crm_cd_desc IN(
"BATTERY - SIMPLE ASSAULT",
"BURGLARY FROM VEHICLE",
"VEHICLE - STOLEN",
"BURGLARY",
"THEFT PLAIN - PETTY ($950 & UNDER)",
"THEFT OF IDENTITY",
"INTIMATE PARTNER - SIMPLE ASSAULT",
"VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS) 0114",
"VANDALISM - MISDEAMEANOR ($399 OR UNDER)",
"ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT") AND zipcode !=
"NONE" group by crm_cd_desc,3, zipcode,city,state)  select * from
query where day != "NULL";'| perl -lpe 's/"/\\"/g; s/^|$/"/g;
s/\t/","/g' >Top_10_Crime_Per_Day_LA.csv;
```

Check the file Top_10_Crime_Per_Day_LA.csv:

```
head -10 Top_10_Crime_Per_Day_LA.csv
```



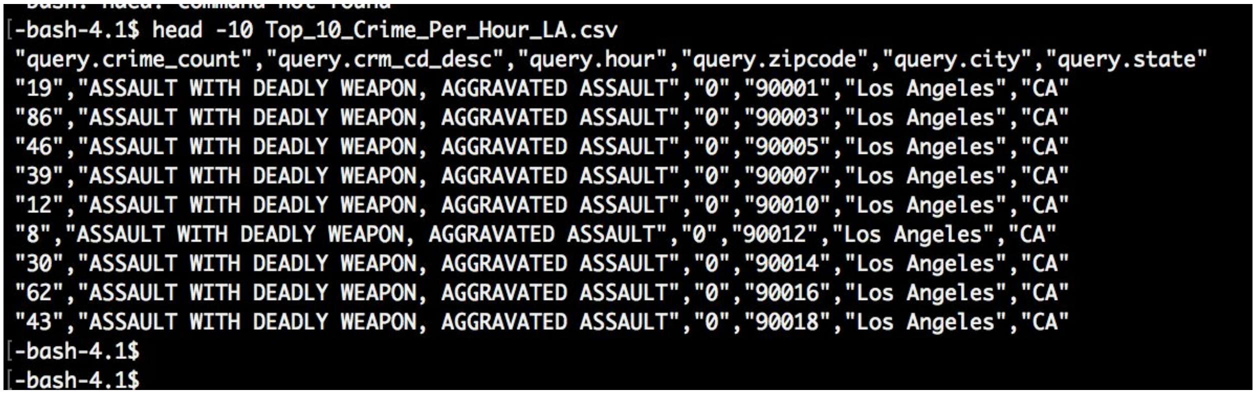
```
[-bash-4.1$
[-bash-4.1$ head -10 Top_10_Crime_Per_Day_LA.csv
"query.crime_count","query.crm_cd_desc","query.day","query.zipcode","query.city","query.state"
"176","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","Fri","90002","Los Angeles","CA"
"96","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","Fri","90004","Los Angeles","CA"
"114","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","Fri","90006","Los Angeles","CA"
"195","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","Fri","90008","Los Angeles","CA"
"443","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","Fri","90011","Los Angeles","CA"
"268","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","Fri","90013","Los Angeles","CA"
"126","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","Fri","90015","Los Angeles","CA"
"130","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","Fri","90017","Los Angeles","CA"
"158","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","Fri","90019","Los Angeles","CA"
[-bash-4.1$
```

Query -4: Top 10 crimes by hour:

```
hive -e 'set hive.cli.print.header=true ; SET
hive.groupby.orderby.position.alias=true ; use crime_data ; with
query as (Select count(*) as
crime_count,crm_cd_desc,CAST(round((time_occ/100),0) as INT) as
hour,zipcode,city,state from la_crime_data where crm_cd_desc is not
NULL and crm_cd_desc != "" AND crm_cd_desc IN(
"BATTERY - SIMPLE ASSAULT",
"BURGLARY FROM VEHICLE",
"VEHICLE - STOLEN",
"BURGLARY",
"THEFT PLAIN - PETTY ($950 & UNDER)",
"THEFT OF IDENTITY",
"INTIMATE PARTNER - SIMPLE ASSAULT",
"VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS) 0114",
"VANDALISM - MISDEAMEANOR ($399 OR UNDER)",
"ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT") AND zipcode !=
"NONE" group by crm_cd_desc,3,zipcode,city,state) select * from query
where hour is  not NULL;'| perl -lpe 's/"/\\"/g; s/^|$/"/g;
s/\t/","/g' >Top_10_Crime_Per_Hour_LA.csv;
```

Check the file Top_10_Crime_Per_Hour_LA.csv:

```
head -10 Top_10_Crime_Per_Hour_LA.csv
```



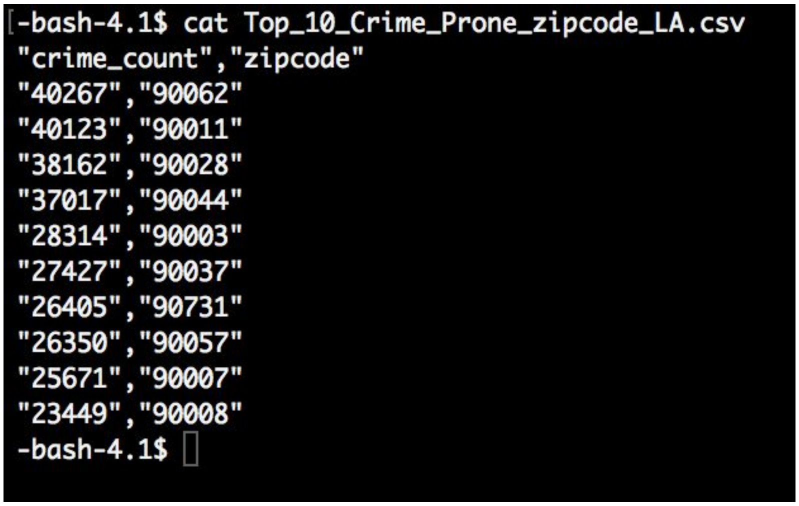
```
 bash: hadu: command not found
[-bash-4.1$ head -10 Top_10_Crime_Per_Hour_LA.csv
 "query.crime_count","query.crm_cd_desc","query.hour","query.zipcode","query.city","query.state"
 "19","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","0","90001","Los Angeles","CA"
 "86","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","0","90003","Los Angeles","CA"
 "46","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","0","90005","Los Angeles","CA"
 "39","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","0","90007","Los Angeles","CA"
 "12","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","0","90010","Los Angeles","CA"
 "8","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","0","90012","Los Angeles","CA"
 "30","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","0","90014","Los Angeles","CA"
 "62","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","0","90016","Los Angeles","CA"
 "43","ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT","0","90018","Los Angeles","CA"
[-bash-4.1$
[-bash-4.1$
```

Query -5: Top 10 crime prone zipcodes:

```
hive -e 'set hive.cli.print.header=true;use crime_data ;select
count(*) as crime_count ,zipcode from la_crime_data group by zipcode
order by crime_count DESC limit 10;'| perl -lpe 's/"/\\"/g;
s/^|$/"/g; s/\t/","/g' > Top_10_Crime_Prone_zipcode_LA.csv;
```

Check the file Top_10_Crime_Prone_zipcode_LA.csv:

```
cat Top_10_Crime_Prone_zipcode_LA.csv
```



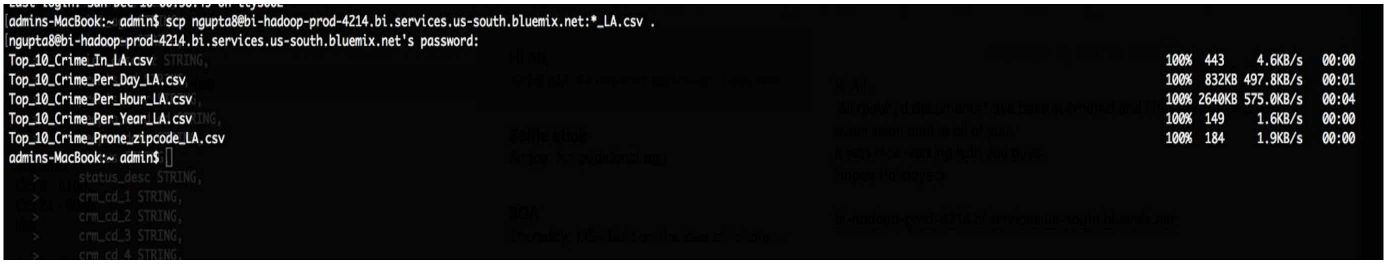
```
[-bash-4.1$ cat Top_10_Crime_Prone_zipcode_LA.csv
 "crime_count","zipcode"
 "40267","90062"
 "40123","90011"
 "38162","90028"
 "37017","90044"
 "28314","90003"
 "27427","90037"
 "26405","90731"
 "26350","90057"
 "25671","90007"
 "23449","90008"
 -bash-4.1$ ▯
```

Download the files on local system by running below command on local shell:

Note:

1. Change the username(ngupta8) and the SSH Host link.
2. For windows system use ftp client like winscp

```
scp ngupta8@bi-hadoop-prod-4214.bi.services.us-south.bluemix.net:*_LA.csv .
```



```
admins-MacBook:~ admin$ scp ngupta8@bi-hadoop-prod-4214.bi.services.us-south.bluemix.net:*_LA.csv .
ngupta8@bi-hadoop-prod-4214.bi.services.us-south.bluemix.net's password:
Top_10_Crime_In_LA.csv                                                   100%  443      4.6KB/s   00:00
Top_10_Crime_Per_Day_LA.csv                                              100%  832KB 497.8KB/s   00:01
Top_10_Crime_Per_Hour_LA.csv                                             100% 2640KB 575.0KB/s   00:04
Top_10_Crime_Per_Year_LA.csv                                             100%  149      1.6KB/s   00:00
Top_10_Crime_Prone_zipcode_LA.csv                                        100%  184      1.9KB/s   00:00
admins-MacBook:~ admin$ 
```

<u>For New York City:</u>

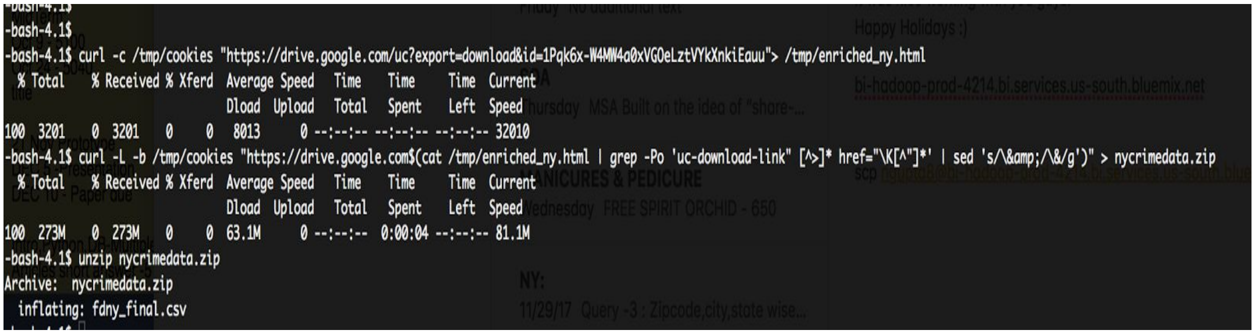
1. Download the data

```
curl -c /tmp/cookies
"https://drive.google.com/uc?export=download&id=1Pqk6x-W4MW4a0xVGOeLz
tVYkXnkiEauu"> /tmp/enriched_ny.html
```

```
curl -L -b /tmp/cookies "https://drive.google.com$(cat
/tmp/enriched_ny.html | grep -Po 'uc-download-link" [^>]*
href="\K[^"]*' | sed 's/\&amp;/\&/g')" > nycrimedata.zip
```

2. Unzip the file

```
unzip nycrimedata.zip
```



```
-bash-4.1$
-bash-4.1$ curl -c /tmp/cookies "https://drive.google.com/uc?export=download&id=1Pqk6x-W4MW4a0xVGOeLztVYkXnkiEauu"> /tmp/enriched_ny.html
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  3201    0  3201    0     0   8013      0 --:--:-- --:--:-- --:--:-- 32010
-bash-4.1$ curl -L -b /tmp/cookies "https://drive.google.com$(cat /tmp/enriched_ny.html | grep -Po 'uc-download-link" [^>]* href="\K[^"]*' | sed 's/\&amp;/\&/g')" > nycrimedata.zip
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  273M    0  273M    0     0  63.1M      0 --:--:--  0:00:04 --:--:-- 81.1M
-bash-4.1$ unzip nycrimedata.zip
Archive:  nycrimedata.zip
  inflating: fdny_final.csv
```

3. Open hive terminal using: hive
4. Run Following commands on hive shell :
    a. Create the database named cime_data

```
create database if not exists crime_data;
```

    b. Check the created database

```
show databases;
```

    c. Select the database.

```
use crime_data;
```

    d. Create the table ny_crime_data

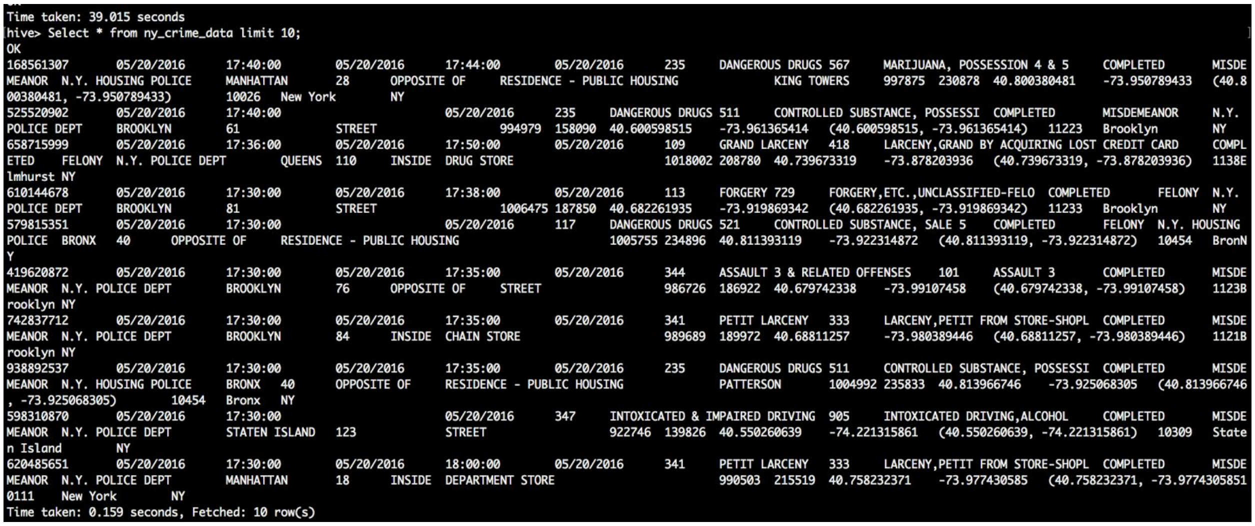
```
CREATE TABLE IF NOT EXISTS ny_crime_data(
     cmplnt_num DECIMAL,
     cmplnt_fr_dt DATE,
     cmplnt_fr_tm STRING,
     cmplnt_to_dt DATE,
     cmplnt_to_tm STRING,
     rpt_dt DATE,
     ky_cd DECIMAL,
     ofns_desc STRING,
     pd_cd DECIMAL,
     pd_desc STRING,
     crm_atpt_cptd_cd STRING,
     law_cat_cd STRING,
     juris_desc STRING,
     boro_nm STRING,
     addr_pct_cd DECIMAL,
     loc_of_occur_desc STRING,
     prem_typ_desc STRING,
     parks_nm STRING,
      hadevelopt STRING,
      x_coord_cd DECIMAL,
      y_coord_cd DECIMAL,
      latitude DECIMAL,
      longitude DECIMAL,
      lat_lon STRING,
    zipcode STRING,
    city STRING,
  state STRING) ROW FORMAT
SERDE'org.apache.hadoop.hive.serde2.OpenCSVSerde'WITH SERDEPROPERTIES
("separatorChar" = ",","quoteChar" = "\"")TBLPROPERTIES
("skip.header.line.count"="1");
```

e. Load the data into the table ny_crime_data

```
load data local inpath '/home/ngupta8/fdny_final.csv' into table
ny_crime_data;
```

f. Check whether data is uploaded properly or not.

```
Select * from ny_crime_data limit 10;
```



```
Time taken: 39.015 seconds
hive> Select * from ny_crime_data limit 10;
OK
168561307       05/20/2016      17:40:00        05/20/2016      17:44:00        05/20/2016      235     DANGEROUS DRUGS 567     MARIJUANA, POSSESSION 4 & 5     COMPLETED       MISDE
MEANOR  N.Y. HOUSING POLICE     MANHATTAN       28      OPPOSITE OF     RESIDENCE - PUBLIC HOUSING              KING TOWERS     997875  230878  40.800380481    -73.950789433   (40.8
00380481, -73.950789433)        10026   New York        NY
525520902       05/20/2016      17:40:00                        05/20/2016      235     DANGEROUS DRUGS 511     CONTROLLED SUBSTANCE, POSSESSI  COMPLETED       MISDEMEANOR     N.Y.
POLICE DEPT     BROOKLYN        61              STREET                  994979  158090  40.600598515    -73.961365414   (40.600598515, -73.961365414)   11223   Brooklyn        NY
658715999       05/20/2016      17:36:00        05/20/2016      17:50:00        05/20/2016      109     GRAND LARCENY   418     LARCENY,GRAND BY ACQUIRING LOST CREDIT CARD     COMPL
ETED    FELONY  N.Y. POLICE DEPT        QUEENS  110     INSIDE  DRUG STORE                      1018002 208780  40.739673319    -73.878203936   (40.739673319, -73.878203936)   1138E
lmhurst NY
610144678       05/20/2016      17:30:00        05/20/2016      17:38:00        05/20/2016      113     FORGERY 729     FORGERY,ETC.,UNCLASSIFIED-FELO  COMPLETED       FELONY  N.Y.
POLICE DEPT     BROOKLYN        81              STREET                  1006475 187850  40.682261935    -73.919869342   (40.682261935, -73.919869342)   11233   Brooklyn        NY
579815351       05/20/2016      17:30:00                        05/20/2016      117     DANGEROUS DRUGS 521     CONTROLLED SUBSTANCE, SALE 5    COMPLETED       FELONY  N.Y. HOUSING
POLICE  BRONX   40      OPPOSITE OF     RESIDENCE - PUBLIC HOUSING              1005755 234896  40.811393119    -73.922314872   (40.811393119, -73.922314872)   10454   BronN
Y
419620872       05/20/2016      17:30:00        05/20/2016      17:35:00        05/20/2016      344     ASSAULT 3 & RELATED OFFENSES    101     ASSAULT 3       COMPLETED       MISDE
MEANOR  N.Y. POLICE DEPT        BROOKLYN        76      OPPOSITE OF     STREET                  986726  186922  40.679742338    -73.99107458    (40.679742338, -73.99107458)    1123B
rooklyn NY
742837712       05/20/2016      17:30:00        05/20/2016      17:35:00        05/20/2016      341     PETIT LARCENY   333     LARCENY,PETIT FROM STORE-SHOPL  COMPLETED       MISDE
MEANOR  N.Y. POLICE DEPT        BROOKLYN        84      INSIDE  CHAIN STORE                     989689  189972  40.68811257     -73.980389446   (40.68811257, -73.980389446)    1121B
rooklyn NY
938892537       05/20/2016      17:30:00        05/20/2016      17:35:00        05/20/2016      235     DANGEROUS DRUGS 511     CONTROLLED SUBSTANCE, POSSESSI  COMPLETED       MISDE
MEANOR  N.Y. HOUSING POLICE     BRONX   40      OPPOSITE OF     RESIDENCE - PUBLIC HOUSING              PATTERSON       1004992 235833  40.813966746    -73.925068305   (40.813966746
, -73.925068305)        10454   Bronx   NY
598310870       05/20/2016      17:30:00                        05/20/2016      347     INTOXICATED & IMPAIRED DRIVING  905     INTOXICATED DRIVING,ALCOHOL     COMPLETED       MISDE
MEANOR  N.Y. POLICE DEPT        STATEN ISLAND   123             STREET                  922746  139826  40.550260639    -74.221315861   (40.550260639, -74.221315861)   10309   State
n Island        NY
620485651       05/20/2016      17:30:00        05/20/2016      18:00:00        05/20/2016      341     PETIT LARCENY   333     LARCENY,PETIT FROM STORE-SHOPL  COMPLETED       MISDE
MEANOR  N.Y. POLICE DEPT        MANHATTAN       18      INSIDE  DEPARTMENT STORE                        990503  215519  40.758232371    -73.977430585   (40.758232371, -73.9774305851
0111    New York        NY
Time taken: 0.159 seconds, Fetched: 10 row(s)
```

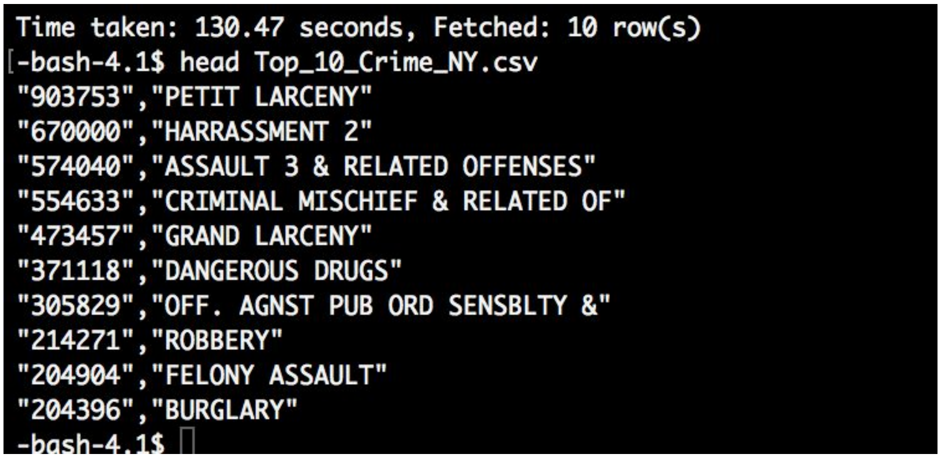
5 . Run Following commands from the bash shell and your home directory for instance /home/ngupta8 (NOT from Hive shell).

Query -1: Top 10 crimes:

```
hive -e "use crime_data ;Select count(*) as crime_count,ofns_desc
from ny_crime_data group by ofns_desc order by crime_count DESC limit
10;"| perl -lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g' >
Top_10_Crime_NY.csv;
```

Check the file Top_10_Crime_NY.csv:

```
head Top_10_Crime_NY.csv
```



```
Time taken: 130.47 seconds, Fetched: 10 row(s)
[-bash-4.1$ head Top_10_Crime_NY.csv
"903753","PETIT LARCENY"
"670000","HARRASSMENT 2"
"574040","ASSAULT 3 & RELATED OFFENSES"
"554633","CRIMINAL MISCHIEF & RELATED OF"
"473457","GRAND LARCENY"
"371118","DANGEROUS DRUGS"
"305829","OFF. AGNST PUB ORD SENSBLTY &"
"214271","ROBBERY"
"204904","FELONY ASSAULT"
"204396","BURGLARY"
-bash-4.1$ 
```

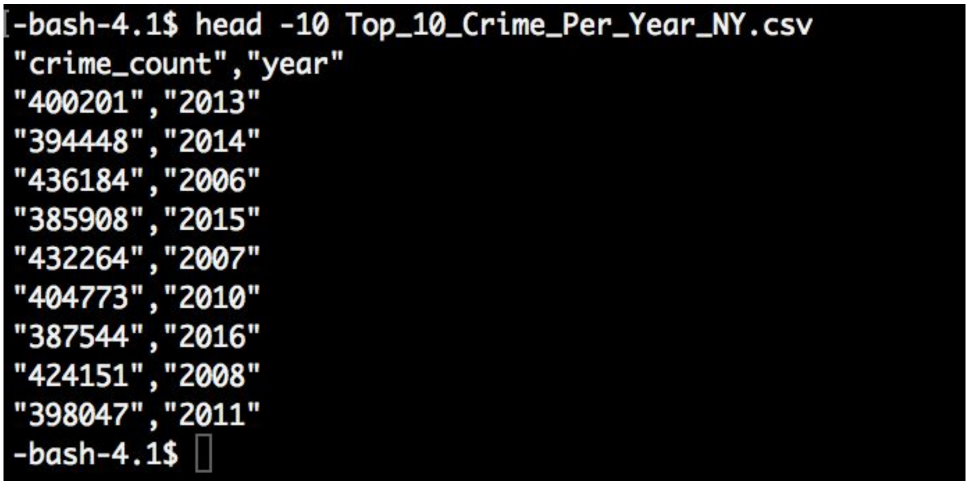
Query -2: Top 10 crimes per year:

```
hive -e 'set hive.cli.print.header=true ; SET
hive.groupby.orderby.position.alias=true ; use crime_data ;select
count(*) as
crime_count,from_unixtime(unix_timestamp(rpt_dt,"mm/dd/yyyy"),"Y") as
year from ny_crime_data where ofns_desc is not NULL and ofns_desc !=
"" AND ofns_desc IN(
"PETIT LARCENY",
"HARRASSMENT 2",
"ASSAULT 3 & RELATED OFFENSES",
"CRIMINAL MISCHIEF & RELATED OF",
"GRAND LARCENY",
"DANGEROUS DRUGS",
"OFF. AGNST PUB ORD SENSBLTY &",
"ROBBERY",
"FELONY ASSAULT",
"BURGLARY") group by 2'| perl -lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g'
```

```
>Top_10_Crime_Per_Year_NY.csv ;
```

Check the file Top_10_Crime_Per_Year_NY.csv:

```
head -10 Top_10_Crime_Per_Year_NY.csv
```



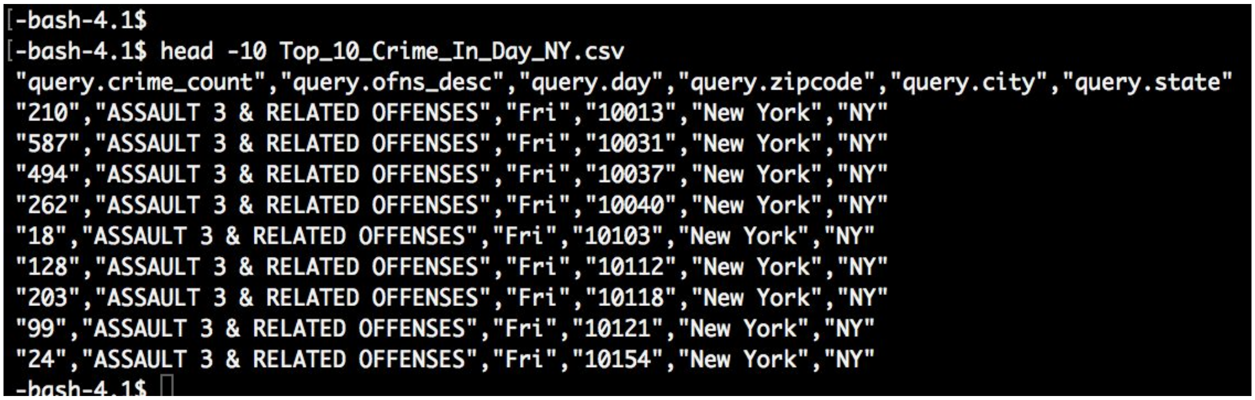
```
-bash-4.1$ head -10 Top_10_Crime_Per_Year_NY.csv
"crime_count","year"
"400201","2013"
"394448","2014"
"436184","2006"
"385908","2015"
"432264","2007"
"404773","2010"
"387544","2016"
"424151","2008"
"398047","2011"
-bash-4.1$ 
```

Query -3: TOP 10 crimes per day:

```
hive -e 'set hive.cli.print.header=true ; SET
hive.groupby.orderby.position.alias=true ; use crime_data ; with
query as (Select count(*) as
crime_count,ofns_desc,from_unixtime(unix_timestamp(cmplnt_fr_dt,"mm/d
d/yyyy"),"E") as day ,zipcode,city,state from ny_crime_data where
ofns_desc is not NULL and ofns_desc != ""  AND ofns_desc IN(
"PETIT LARCENY",
"HARRASSMENT 2",
"ASSAULT 3 & RELATED OFFENSES",
"CRIMINAL MISCHIEF & RELATED OF",
"GRAND LARCENY",
"DANGEROUS DRUGS",
"OFF. AGNST PUB ORD SENSBLTY &",
"ROBBERY",
"FELONY ASSAULT",
"BURGLARY") AND zipcode != "NONE" group by ofns_desc,3,
zipcode,city,state) select * from query where day != "NULL";'| perl
-lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g' >Top_10_Crime_In_Day_NY.csv;
```

Check the file Top_10_Crime_In_Day_NY.csv:

```
head -10 Top_10_Crime_In_Day_NY.csv
```



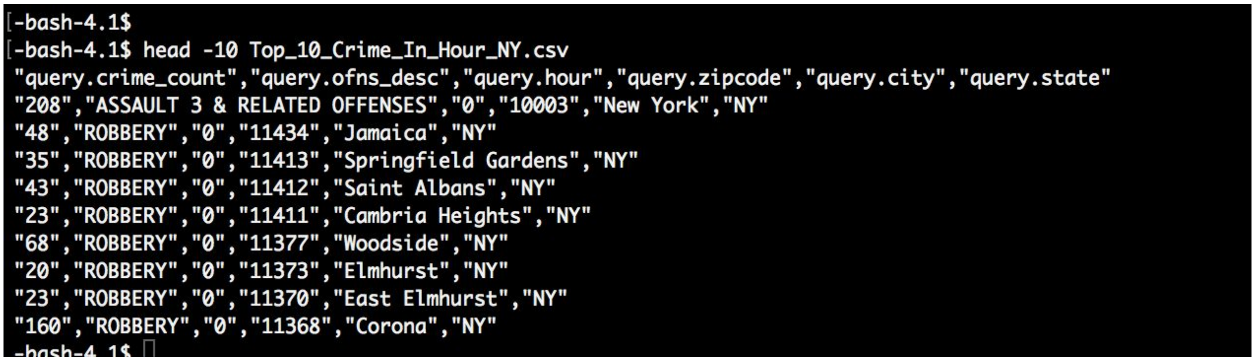
```
[-bash-4.1$
[-bash-4.1$ head -10 Top_10_Crime_In_Day_NY.csv
"query.crime_count","query.ofns_desc","query.day","query.zipcode","query.city","query.state"
"210","ASSAULT 3 & RELATED OFFENSES","Fri","10013","New York","NY"
"587","ASSAULT 3 & RELATED OFFENSES","Fri","10031","New York","NY"
"494","ASSAULT 3 & RELATED OFFENSES","Fri","10037","New York","NY"
"262","ASSAULT 3 & RELATED OFFENSES","Fri","10040","New York","NY"
"18","ASSAULT 3 & RELATED OFFENSES","Fri","10103","New York","NY"
"128","ASSAULT 3 & RELATED OFFENSES","Fri","10112","New York","NY"
"203","ASSAULT 3 & RELATED OFFENSES","Fri","10118","New York","NY"
"99","ASSAULT 3 & RELATED OFFENSES","Fri","10121","New York","NY"
"24","ASSAULT 3 & RELATED OFFENSES","Fri","10154","New York","NY"
-bash-4.1$
```

Query -4: Top 10 crimes per hour:

```
hive -e 'set hive.cli.print.header=true ; SET
hive.groupby.orderby.position.alias=true ; use crime_data ; with
query as (Select count(*) as crime_count,ofns_desc,hour(cmplnt_fr_tm)
as hour,zipcode,city,state from ny_crime_data where ofns_desc is not
NULL and ofns_desc != ""  AND ofns_desc IN(
"PETIT LARCENY",
"HARRASSMENT 2",
"ASSAULT 3 & RELATED OFFENSES",
"CRIMINAL MISCHIEF & RELATED OF",
"GRAND LARCENY",
"DANGEROUS DRUGS",
"OFF. AGNST PUB ORD SENSBLTY &",
"ROBBERY",
"FELONY ASSAULT",
"BURGLARY") AND zipcode != "NONE" group by ofns_desc,3,
zipcode,city,state) select * from query where hour is not NULL  order
by hour;'| perl -lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g'
>Top_10_Crime_In_Hour_NY.csv;
```

Check the file Top_10_Crime_In_Hour_NY.csv:

```
head -10 Top_10_Crime_In_Hour_NY.csv
```



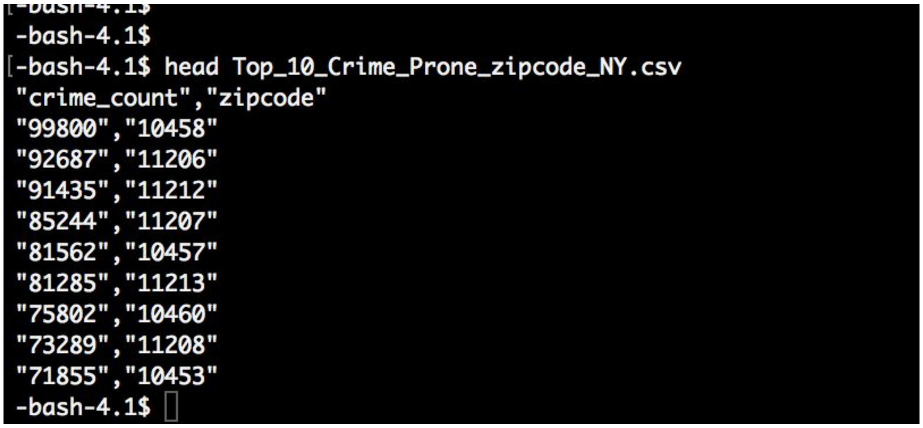
```
[-bash-4.1$
[-bash-4.1$ head -10 Top_10_Crime_In_Hour_NY.csv
"query.crime_count","query.ofns_desc","query.hour","query.zipcode","query.city","query.state"
"208","ASSAULT 3 & RELATED OFFENSES","0","10003","New York","NY"
"48","ROBBERY","0","11434","Jamaica","NY"
"35","ROBBERY","0","11413","Springfield Gardens","NY"
"43","ROBBERY","0","11412","Saint Albans","NY"
"23","ROBBERY","0","11411","Cambria Heights","NY"
"68","ROBBERY","0","11377","Woodside","NY"
"20","ROBBERY","0","11373","Elmhurst","NY"
"23","ROBBERY","0","11370","East Elmhurst","NY"
"160","ROBBERY","0","11368","Corona","NY"
-bash-4.1$
```

Query -5: Top 10 crime prone zipcodes:

```
hive -e 'set hive.cli.print.header=true;use crime_data ; select
count(*) as crime_count,zipcode from ny_crime_data where zipcode is
not NULL and zipcode !="NONE" and zipcode != "" group by zipcode
order by crime_count DESC limit 10;'| perl -lpe 's/"/\\"/g;
s/^|$/"/g; s/\t/","/g' > Top_10_Crime_Prone_zipcode_NY.csv;
```

Check the file Top_10_Crime_Prone_zipcode_NY.csv:

```
head Top_10_Crime_Prone_zipcode_NY.csv
```



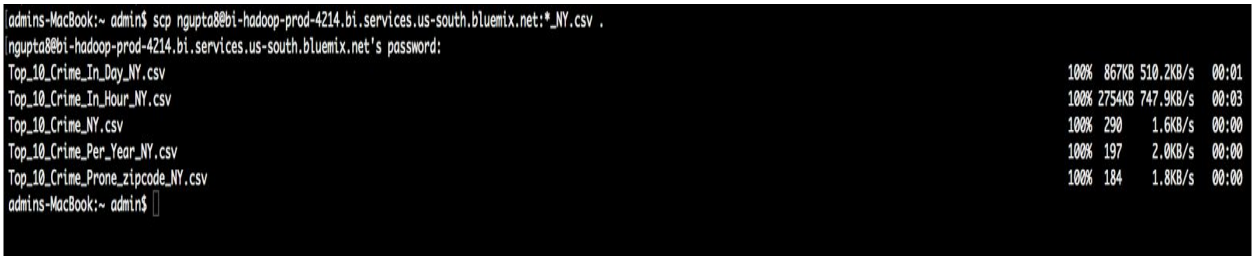
```
[-bash-4.1$
-bash-4.1$
[-bash-4.1$ head Top_10_Crime_Prone_zipcode_NY.csv
"crime_count","zipcode"
"99800","10458"
"92687","11206"
"91435","11212"
"85244","11207"
"81562","10457"
"81285","11213"
"75802","10460"
"73289","11208"
"71855","10453"
-bash-4.1$
```

Download processed data on local:

Note:

1. Change the username(ngupta8) and the SSH Host link.
2. For windows system use ftp client like winscp

```
scp ngupta8@bi-hadoop-prod-4214.bi.services.us-south.bluemix.net:*_NY.csv .
```



```
admins-MacBook:~ admin$ scp ngupta8@bi-hadoop-prod-4214.bi.services.us-south.bluemix.net:*_NY.csv .
ngupta8@bi-hadoop-prod-4214.bi.services.us-south.bluemix.net's password:
Top_10_Crime_In_Day_NY.csv                                                          100%  867KB 510.2KB/s   00:01
Top_10_Crime_In_Hour_NY.csv                                                         100% 2754KB 747.9KB/s   00:03
Top_10_Crime_NY.csv                                                                 100%  290     1.6KB/s   00:00
Top_10_Crime_Per_Year_NY.csv                                                        100%  197     2.0KB/s   00:00
Top_10_Crime_Prone_zipcode_NY.csv                                                   100%  184     1.8KB/s   00:00
admins-MacBook:~ admin$
```

<u>For Chicago City:</u>

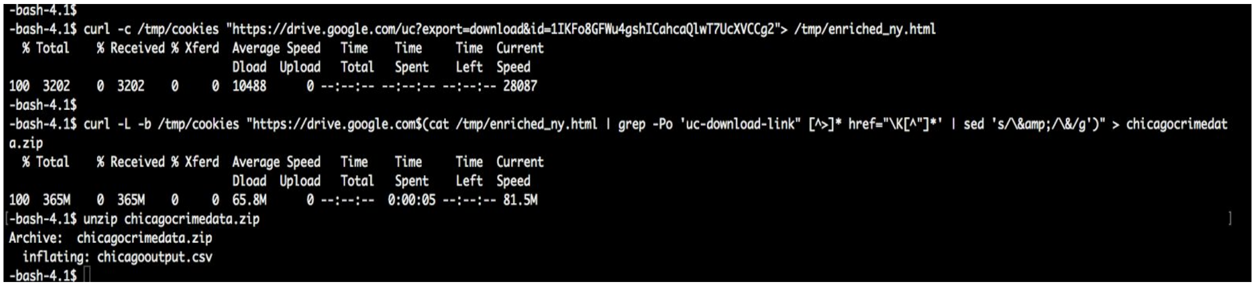
1. Download the data

```
curl -c /tmp/cookies
"https://drive.google.com/uc?export=download&id=1IKFo8GFWu4gshICahcaQ
lwT7UcXVCCg2"> /tmp/enriched_ny.html
```

```
curl -L -b /tmp/cookies "https://drive.google.com$(cat
/tmp/enriched_ny.html | grep -Po 'uc-download-link" [^>]*
href="\K[^"]*' | sed 's/\&amp;/\&/g')" > chicagocrimedata.zip
```

2. Unzip the file

```
unzip chicagocrimedata.zip
```



```
-bash-4.1$
-bash-4.1$ curl -c /tmp/cookies "https://drive.google.com/uc?export=download&id=1IKFo8GFWu4gshICahcaQlwT7UcXVCCg2"> /tmp/enriched_ny.html
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  3202    0  3202    0     0  10488      0 --:--:-- --:--:-- --:--:-- 28087
-bash-4.1$
-bash-4.1$ curl -L -b /tmp/cookies "https://drive.google.com$(cat /tmp/enriched_ny.html | grep -Po 'uc-download-link" [^>]* href="\K[^"]*' | sed 's/\&amp;/\&/g')" > chicagocrimedat
a.zip
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  365M    0  365M    0     0  65.8M      0 --:--:--  0:00:05 --:--:-- 81.5M
-bash-4.1$ unzip chicagocrimedata.zip
Archive:  chicagocrimedata.zip
  inflating: chicagooutput.csv
-bash-4.1$
```

3. Open hive terminal using: hive
4. Run Following commands on hive shell :
    a. Create the database named crime_data

```
create database if not exists crime_data;
```

    b. Check the created database

```
show databases;
```

    c. Select the database

```
use crime_data;
```

d. Create the table chicago_crime_data

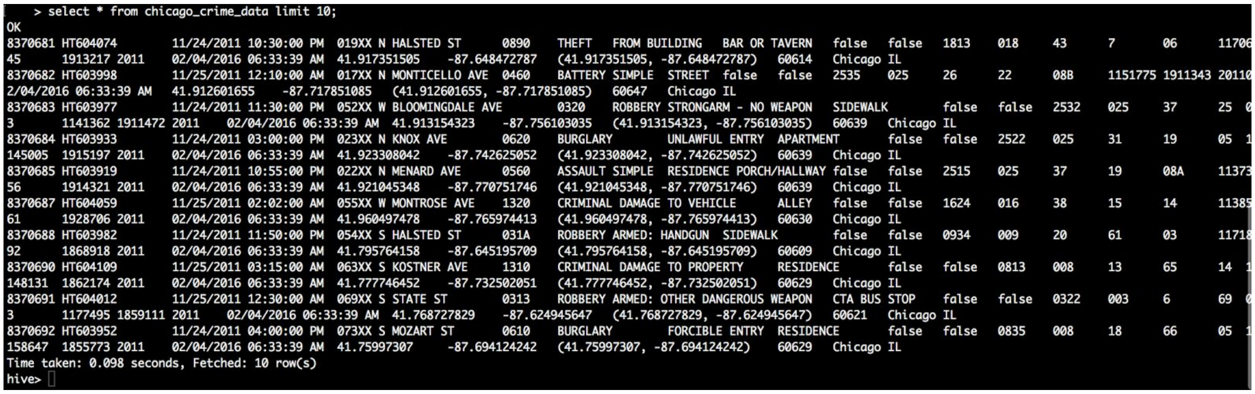
```
CREATE TABLE IF NOT EXISTS chicago_crime_data(ID INT,
Case_Number STRING,
Crime_date STRING,
Block STRING,
IUCR INT,
Primary_Type STRING,
Description STRING,
Location_Description STRING,
Arrest BOOLEAN,
Domestic BOOLEAN,
Beat INT,
District INT,
Ward INT,
Community_Area INT,
FBI_Code INT,
X_Coordinate INT,
Y_Coordinate INT,
Year INT,
Updated_On STRING,
Latitude DOUBLE,
Longitude DOUBLE,
Location DOUBLE,
Zipcode INT,
City STRING,
State STRING)
ROW FORMAT SERDE'org.apache.hadoop.hive.serde2.OpenCSVSerde'WITH
SERDEPROPERTIES ("separatorChar" = ",","quoteChar" =
"\"")TBLPROPERTIES ("skip.header.line.count"="1");
```

e. Load the data into the table chicago_crime_data

```
load data local inpath '/home/ngupta8/chicagooutput.csv' into table
chicago_crime_data;
```

f. Check whether data is uploaded properly or not.

```
select * from chicago_crime_data limit 10;
```



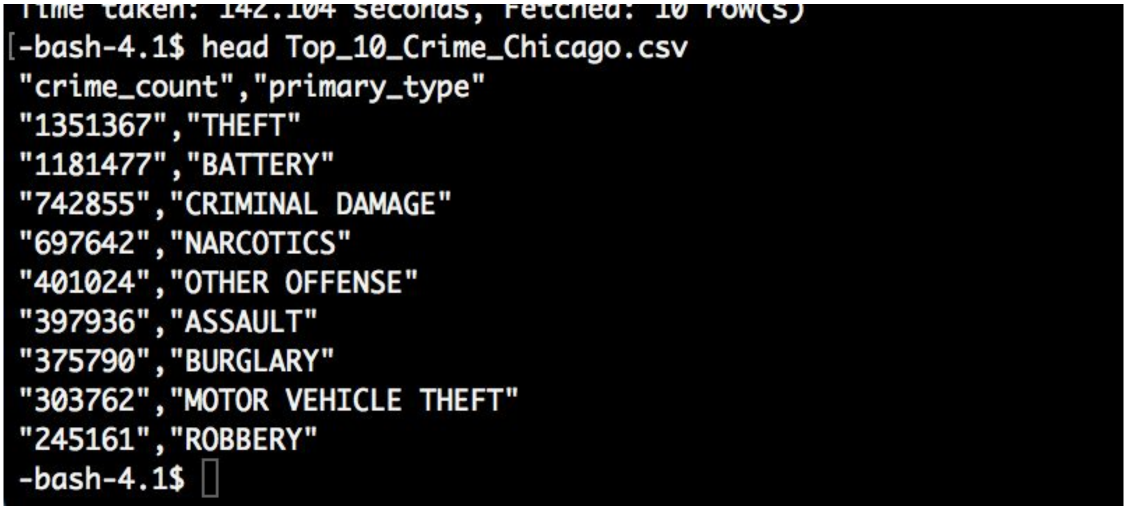
```
[     > select * from chicago_crime_data limit 10;
OK
8370681 HT604074        11/24/2011 10:30:00 PM  019XX N HALSTED ST      0890    THEFT   FROM BUILDING   BAR OR TAVERN   false   false   1813    018     43      7       06      11706
45      1913217 2011    02/04/2016 06:33:39 AM  41.917351505    -87.648472787   (41.917351505, -87.648472787)   60614   Chicago IL
8370682 HT603998        11/25/2011 12:10:00 AM  017XX N MONTICELLO AVE  0460    BATTERY SIMPLE  STREET  false   false   2535    025     26      22      08B     1151775 1911343 20110
2/04/2016 06:33:39 AM   41.912601655    -87.717851085   (41.912601655, -87.717851085)   60647   Chicago IL
8370683 HT603977        11/24/2011 11:30:00 PM  052XX W BLOOMINGDALE AVE        0320    ROBBERY STRONGARM - NO WEAPON   SIDEWALK        false   false   2532    025     37      25  0
3       1141362 1911472 2011    02/04/2016 06:33:39 AM  41.913154323    -87.756103035   (41.913154323, -87.756103035)   60639   Chicago IL
8370684 HT603933        11/24/2011 03:00:00 PM  023XX N KNOX AVE        0620    BURGLARY        UNLAWFUL ENTRY  APARTMENT       false   false   2522    025     31      19      05  1
145005  1915197 2011    02/04/2016 06:33:39 AM  41.923308042    -87.742625052   (41.923308042, -87.742625052)   60639   Chicago IL
8370685 HT603919        11/24/2011 10:55:00 PM  022XX N MENARD AVE      0560    ASSAULT SIMPLE  RESIDENCE PORCH/HALLWAY false   false   2515    025     37      19      08A     11373
56      1914321 2011    02/04/2016 06:33:39 AM  41.921045348    -87.770751746   (41.921045348, -87.770751746)   60639   Chicago IL
8370687 HT604059        11/25/2011 02:02:00 AM  055XX W MONTROSE AVE    1320    CRIMINAL DAMAGE TO VEHICLE      ALLEY   false   false   1624    016     38      15      14      11385
61      1928706 2011    02/04/2016 06:33:39 AM  41.960497478    -87.765974413   (41.960497478, -87.765974413)   60630   Chicago IL
8370688 HT603982        11/24/2011 11:50:00 PM  054XX S HALSTED ST      031A    ROBBERY ARMED: HANDGUN  SIDEWALK        false   false   0934    009     20      61      03      11718
92      1868918 2011    02/04/2016 06:33:39 AM  41.795764158    -87.645195709   (41.795764158, -87.645195709)   60609   Chicago IL
8370690 HT604109        11/25/2011 03:15:00 AM  063XX S KOSTNER AVE     1310    CRIMINAL DAMAGE TO PROPERTY     RESIDENCE       false   false   0813    008     13      65      14  1
148131  1862174 2011    02/04/2016 06:33:39 AM  41.777746452    -87.732502051   (41.777746452, -87.732502051)   60629   Chicago IL
8370691 HT604012        11/25/2011 12:30:00 AM  069XX S STATE ST        0313    ROBBERY ARMED: OTHER DANGEROUS WEAPON   CTA BUS STOP    false   false   0322    003     6       69  0
3       1177495 1859111 2011    02/04/2016 06:33:39 AM  41.768727829    -87.624945647   (41.768727829, -87.624945647)   60621   Chicago IL
8370692 HT603952        11/24/2011 04:00:00 PM  073XX S MOZART ST       0610    BURGLARY        FORCIBLE ENTRY  RESIDENCE       false   false   0835    008     18      66      05  1
158647  1855773 2011    02/04/2016 06:33:39 AM  41.75997307     -87.694124242   (41.75997307, -87.694124242)    60629   Chicago IL
Time taken: 0.098 seconds, Fetched: 10 row(s)
hive> 
```

Query -1: Top 10 crimes:

```
hive -e "set hive.cli.print.header=true ;use crime_data ;Select
count(*) as crime_count,Primary_Type from chicago_crime_data group by
Primary_Type order by crime_count DESC limit 10;"| perl -lpe
's/"/\\"/g; s/^|$/"/g; s/\t/","/g' > Top_10_Crime_Chicago.csv ;
```

Check the file Top_10_Crime_Chicago.csv:

```
head Top_10_Crime_Chicago.csv
```



```
Time taken: 142.104 seconds, Fetched: 10 row(s)
[-bash-4.1$ head Top_10_Crime_Chicago.csv
"crime_count","primary_type"
"1351367","THEFT"
"1181477","BATTERY"
"742855","CRIMINAL DAMAGE"
"697642","NARCOTICS"
"401024","OTHER OFFENSE"
"397936","ASSAULT"
"375790","BURGLARY"
"303762","MOTOR VEHICLE THEFT"
"245161","ROBBERY"
-bash-4.1$ 
```

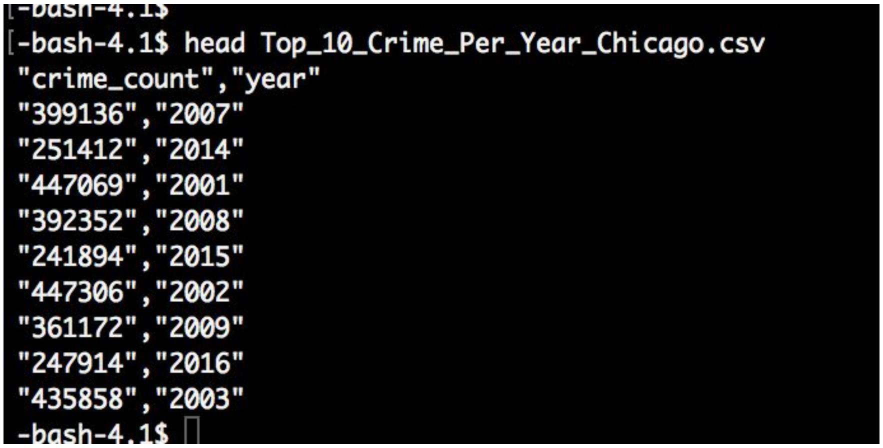
Query -2: Top 10 crimes per year:

```
hive -e 'set hive.cli.print.header=true ; SET
```

```
hive.groupby.orderby.position.alias=true ;use crime_data; select
count(*) as
crime_count,from_unixtime(unix_timestamp(Crime_date,"mm/dd/yyyy"),"Y"
) as year from chicago_crime_data where Primary_Type is not NULL and
Primary_Type != "" AND Primary_Type IN(
"THEFT",
"BATTERY",
"CRIMINAL DAMAGE",
"NARCOTICS",
"OTHER OFFENSE",
"ASSAULT",
"BURGLARY",
"MOTOR VEHICLE THEFT",
"ROBBERY",
"DECEPTIVE PRACTICE") group by 2'| perl -lpe 's/"/\\"/g; s/^|$/"/g;
s/\t/","/g' >Top_10_Crime_Per_Year_Chicago.csv ;
```

Check the file Top_10_Crime_Per_Year_Chicago.csv:

```
head Top_10_Crime_Per_Year_Chicago.csv
```



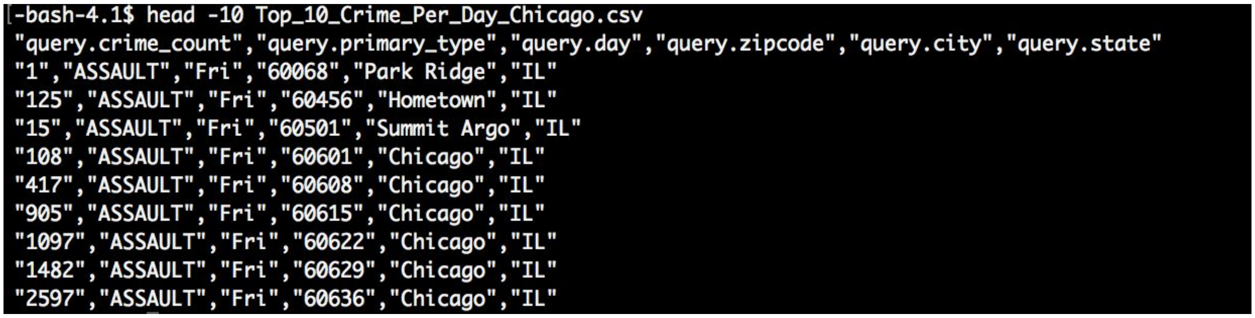
```
-bash-4.1$
[-bash-4.1$ head Top_10_Crime_Per_Year_Chicago.csv
"crime_count","year"
"399136","2007"
"251412","2014"
"447069","2001"
"392352","2008"
"241894","2015"
"447306","2002"
"361172","2009"
"247914","2016"
"435858","2003"
-bash-4.1$
```

Query -3: Top 10 crimes by day:

```
hive -e 'set hive.cli.print.header=true ; SET
hive.groupby.orderby.position.alias=true ; use crime_data ; with
query as (Select count(*) as
crime_count,Primary_Type,from_unixtime(unix_timestamp(Crime_date,"mm/
dd/yyyy"),"E") as day ,zipcode,city,state from chicago_crime_data
where Primary_Type is not NULL and Primary_Type != ""  AND
Primary_Type IN(
"THEFT",
"BATTERY",
"CRIMINAL DAMAGE",
"NARCOTICS",
"OTHER OFFENSE",
"ASSAULT",
"BURGLARY",
"MOTOR VEHICLE THEFT",
"ROBBERY",
"DECEPTIVE PRACTICE") AND zipcode != "NONE" group by Primary_Type,3,
zipcode,city,state) select * from query where day != "NULL";'| perl
-lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g'
>Top_10_Crime_Per_Day_Chicago.csv;
```

Check the file Top_10_Crime_Per_Day_Chicago.csv:

```
head -10 Top_10_Crime_Per_Day_Chicago.csv
```



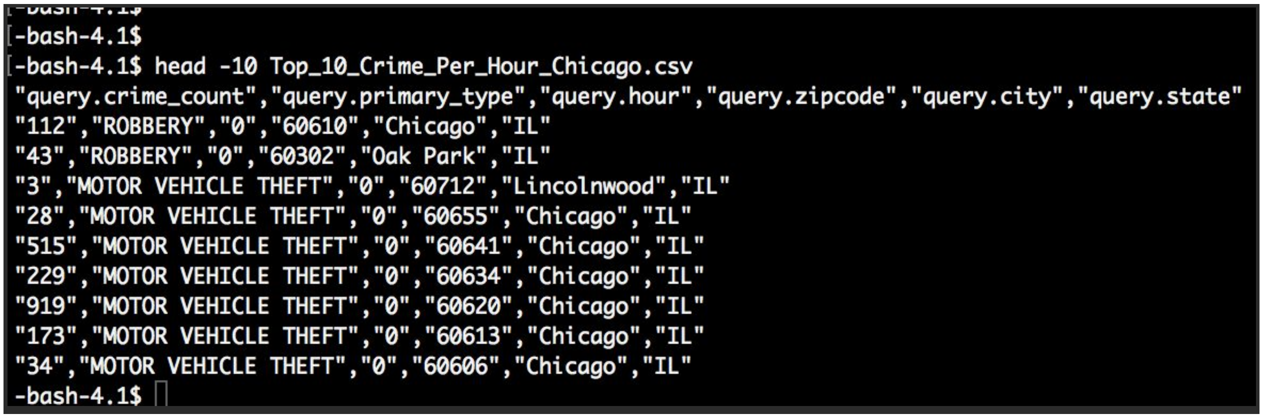
```
-bash-4.1$ head -10 Top_10_Crime_Per_Day_Chicago.csv
"query.crime_count","query.primary_type","query.day","query.zipcode","query.city","query.state"
"1","ASSAULT","Fri","60068","Park Ridge","IL"
"125","ASSAULT","Fri","60456","Hometown","IL"
"15","ASSAULT","Fri","60501","Summit Argo","IL"
"108","ASSAULT","Fri","60601","Chicago","IL"
"417","ASSAULT","Fri","60608","Chicago","IL"
"905","ASSAULT","Fri","60615","Chicago","IL"
"1097","ASSAULT","Fri","60622","Chicago","IL"
"1482","ASSAULT","Fri","60629","Chicago","IL"
"2597","ASSAULT","Fri","60636","Chicago","IL"
```

Query -4: Top 10 crimes by hour:

```
hive -e 'set hive.cli.print.header=true ; SET
hive.groupby.orderby.position.alias=true ; use crime_data ; with
query as (Select count(*) as
crime_count,Primary_Type,from_unixtime(unix_timestamp(Crime_date,"mm/
dd/yyyy hh:mm:ss a"),"H") as hour ,zipcode,city,state from
chicago_crime_data where Primary_Type is not NULL and Primary_Type !=
""  AND Primary_Type IN(
"THEFT",
"BATTERY",
"CRIMINAL DAMAGE",
"NARCOTICS",
"OTHER OFFENSE",
"ASSAULT",
"BURGLARY",
"MOTOR VEHICLE THEFT",
"ROBBERY",
"DECEPTIVE PRACTICE") AND zipcode != "NONE" group by Primary_Type,3,
zipcode,city,state) select * from query order by hour;'| perl -lpe
's/"/\\"/g; s/^|$/"/g; s/\t/","/g'
>Top_10_Crime_Per_Hour_Chicago.csv;
```

Check the file Top_10_Crime_Per_Hour_Chicago.csv:

```
head -10 Top_10_Crime_Per_Hour_Chicago.csv
```



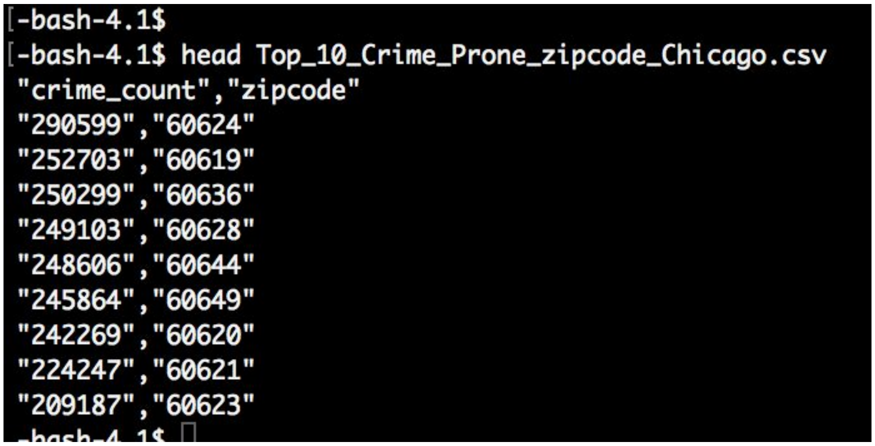
```
[-bash-4.1$
[-bash-4.1$ head -10 Top_10_Crime_Per_Hour_Chicago.csv
"query.crime_count","query.primary_type","query.hour","query.zipcode","query.city","query.state"
"112","ROBBERY","0","60610","Chicago","IL"
"43","ROBBERY","0","60302","Oak Park","IL"
"3","MOTOR VEHICLE THEFT","0","60712","Lincolnwood","IL"
"28","MOTOR VEHICLE THEFT","0","60655","Chicago","IL"
"515","MOTOR VEHICLE THEFT","0","60641","Chicago","IL"
"229","MOTOR VEHICLE THEFT","0","60634","Chicago","IL"
"919","MOTOR VEHICLE THEFT","0","60620","Chicago","IL"
"173","MOTOR VEHICLE THEFT","0","60613","Chicago","IL"
"34","MOTOR VEHICLE THEFT","0","60606","Chicago","IL"
-bash-4.1$
```

Query -5: Top 10 crime prone zipcodes:

```
hive -e 'set hive.cli.print.header=true;use crime_data ; select
count(*) as crime_count,zipcode from chicago_crime_data where zipcode
is not NULL and zipcode !="NONE" and zipcode != "" group by zipcode
order by crime_count DESC limit 10;'| perl -lpe 's/"/\\"/g;
s/^|$/"/g; s/\t/","/g' > Top_10_Crime_Prone_zipcode_Chicago.csv;
```

Check the file Top_10_Crime_Prone_zipcode_Chicago.csv:

```
head Top_10_Crime_Prone_zipcode_Chicago.csv
```



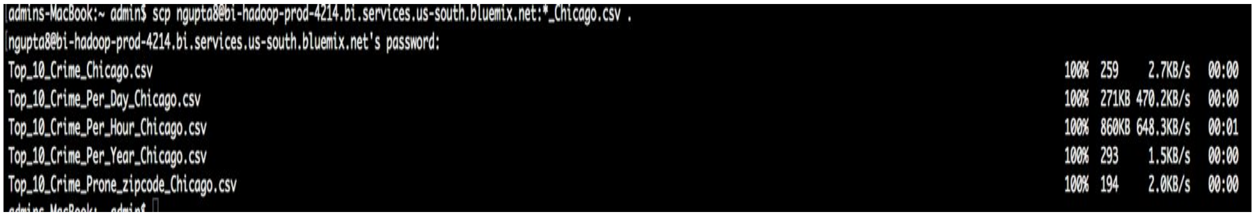
```
-bash-4.1$
-bash-4.1$ head Top_10_Crime_Prone_zipcode_Chicago.csv
"crime_count","zipcode"
"290599","60624"
"252703","60619"
"250299","60636"
"249103","60628"
"248606","60644"
"245864","60649"
"242269","60620"
"224247","60621"
"209187","60623"
-bash-4.1$
```

Download processed data on local:

Note:

1. Change the username(ngupta8) and the SSH Host link.
2. For windows system use ftp client like winscp

```
scp ngupta8@bi-hadoop-prod-4214.bi.services.us-south.bluemix.net:*_Chicago.csv .
```



```
admins-MacBook:~ admin$ scp ngupta8@bi-hadoop-prod-4214.bi.services.us-south.bluemix.net:*_Chicago.csv .
ngupta8@bi-hadoop-prod-4214.bi.services.us-south.bluemix.net's password:
Top_10_Crime_Chicago.csv                                          100% 259     2.7KB/s   00:00
Top_10_Crime_Per_Day_Chicago.csv                                  100% 271KB 470.2KB/s   00:00
Top_10_Crime_Per_Hour_Chicago.csv                                 100% 860KB 648.3KB/s   00:01
Top_10_Crime_Per_Year_Chicago.csv                                 100% 293     1.5KB/s   00:00
Top_10_Crime_Prone_zipcode_Chicago.csv                            100% 194     2.0KB/s   00:00
admins-MacBook:~ admin$
```

Step 4: Visualize the data:

1. Install Tableau and use Tableau to make graphs

Graph 1 : Top 10 Crime Types (Horizontal Bar Chart)

1. Import the Top_10_Crime_(City).csv of the appropriate city into Tableau

2. Drag “No. of Crime” to Columns, “Crime Type” to Rows

3. Drag “Crime Type” to Color

4. Go to Analytics (Next to Data), drag “Average Line” to the middle of the dashboard

5. Title the chart “(City) Top 10 Crime Types”

6. Save

7. Repeat the same steps for all cities (Los Angeles, New York, Chicago).

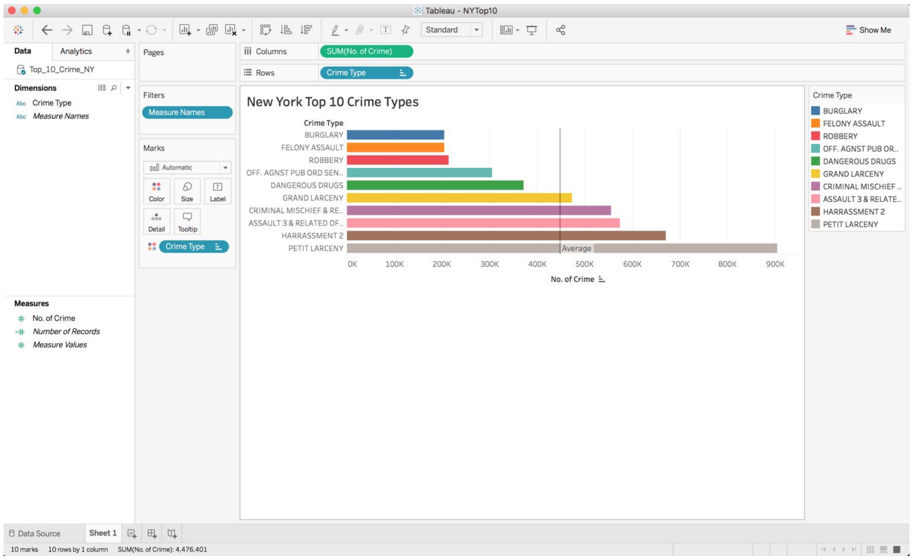


Tableau - NYTop10

Data Source: Top_10_Crime_NY

Dimensions: Crime Type, Measure Names

Measures: No. of Crime, Number of Records, Measure Values

Filters: Measure Names

Marks: Automatic; Color: Crime Type

Columns: SUM(No. of Crime)

Rows: Crime Type

10 marks   10 rows by 1 column   SUM(No. of Crime): 4,476,401

Graph 2 : Crime Overview, by Year (Line Chart)

1. Import the Top_10_Crime_Per_Year(City).csv of the appropriate city into Tableau

2. Drag “Year” to Columns, “Crime Count” to Rows

3. Drag “Crime Count” to Color and Size

4. Change chart type from Automatic to “Line”

5. Title the chart “(City) Crime Overview, by Year”

6. Save

7. Repeat the same steps for all cities (Los Angeles, New York, Chicago).

(continued on next page)

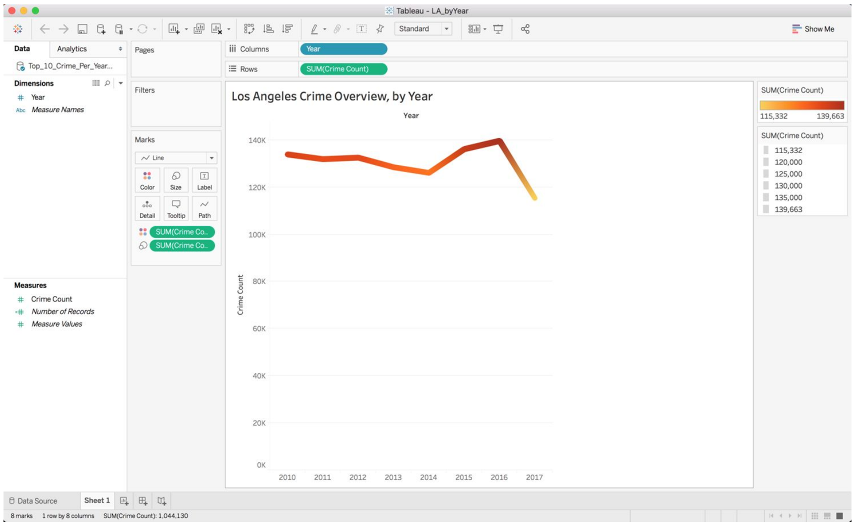


Tableau - LA_byYear

Top_10_Crime_Per_Year...

Columns: Year

Rows: SUM(Crime Count)

Marks: Line

SUM(Crime Co..

SUM(Crime Co..

Los Angeles Crime Overview, by Year

8 marks 1 row by 8 columns SUM(Crime Count): 1,044,130

Graph 3: Crime Overview, by Day (Multiple Line Chart)

1. Import the Top_10_Crime_Per_Day(City).csv of the appropriate city into Tableau
2. Drag "Day" to Columns, "Crime Count" to Rows
3. Drag "Crime Count" to Color and Label
4. Drag "Day" to Detail, and sort it.
5. Change chart type from Automatic to "Line"
6. Title the chart "(City) Crime Overview, by Day"
7. Save
8. Repeat the same steps for all cities (Los Angeles, New York, Chicago).

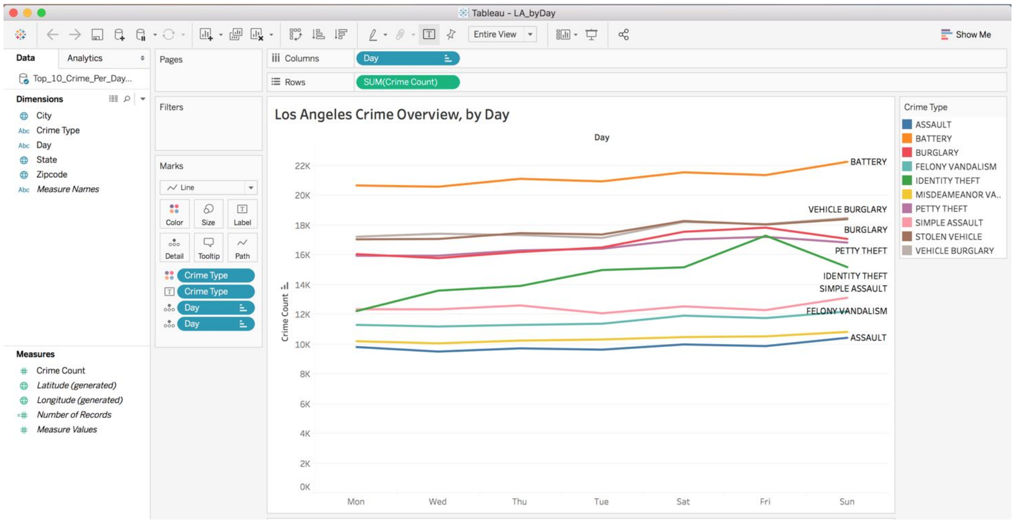


Tableau - LA_byDay

Top_10_Crime_Per_Day...

Columns: Day

Rows: SUM(Crime Count)

Marks: Line

Crime Type

Crime Type

Day

Day

Los Angeles Crime Overview, by Day

Graph 4 : Crime Overview, by Hour (Heat Graph)

1. Import the Top_10_Crime_Per_Hour(City).csv of the appropriate city into Tableau

2. Drag “Crime Type” to Columns, “Hour” to Rows

3. Drag “Crime Count” to Color

4. Title the chart “(City) Crime Overview, by Hour”

5. Save

6. Repeat the same steps for all cities (Los Angeles, New York, Chicago).

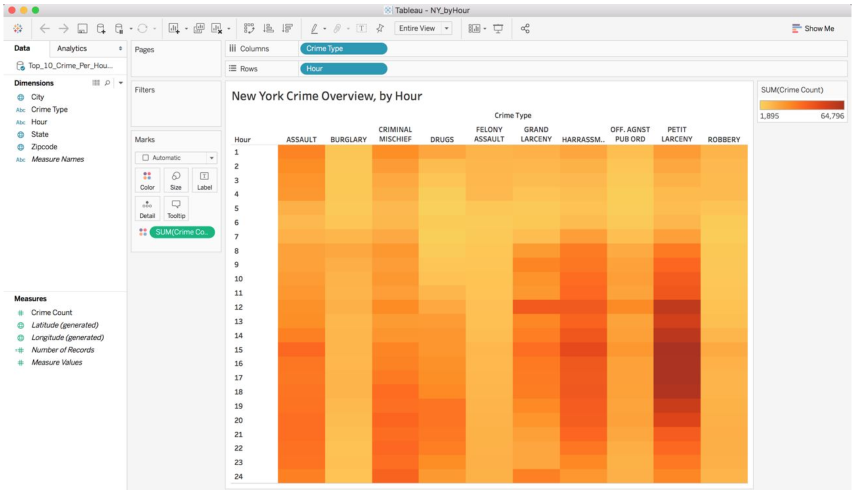


Tableau - NY_byHour

Columns: Crime Type

Rows: Hour

Marks: Automatic, SUM(Crime Co..

New York Crime Overview, by Hour

<u>Crime Overview, by zipcode (Multiple block Chart)</u>

1. Import the Top_10_Crime_Prone_Zipcode(City).csv of the appropriate city into Tableau

2. Drag “Longitude” to Columns, “Latitude” to Rows

3. Drag “Crime Count” to Color

4. Drag “Zipcode” to Detail and label

5. On the top menu bar, select “Map”, then select “Map Layer”

6. Make sure Base, Land Cover, Coastline, Streets and Highways are selected

7. Then, at the bottom, at the Data Layer portion, select Household Income Median” for layer, “ZipCode” for by, and “Blue-Green Gradient” for using.

8. Change chart type from Automatic to “Line”

9. Title the chart “(City) Crime Overview, by Zipcode”

10. Save

11. Repeat the same steps for all cities (Los Angeles, New York, Chicago).

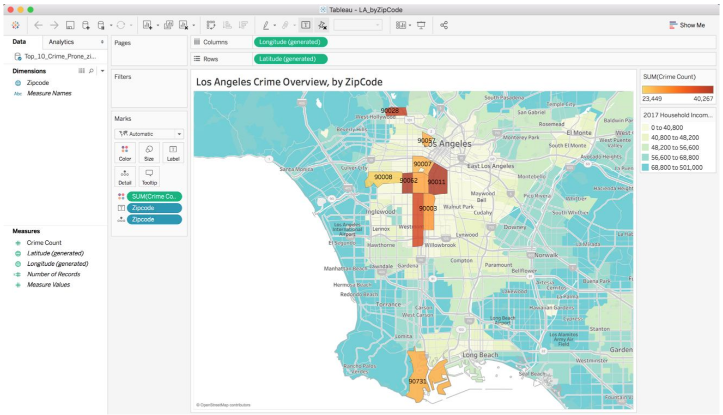


Tableau - LA_byZipCode

Show Me

Data | Analytics

Top_10_Crime_Prone_zi...

Dimensions
- Zipcode
- *Measure Names*

Measures
- Crime Count
- *Latitude (generated)*
- *Longitude (generated)*
- *Number of Records*
- *Measure Values*

Pages

Filters

Marks

Automatic

Color | Size | Label | Detail | Tooltip

- SUM(Crime Co..
- Zipcode
- Zipcode

Columns: Longitude (generated)

Rows: Latitude (generated)

Los Angeles Crime Overview, by ZipCode

SUM(Crime Count)

23,449 40,267

2017 Household Incom...

▨	0 to 40,800
▨	40,800 to 48,200
▨	48,200 to 56,600
▨	56,600 to 68,800
▨	68,800 to 501,000