# **CIS4560 Term Project Tutorial**

**Authors: Blanco, Andrea, Kawashima, Keith, Najera, Jose, Macias, Jennifer, Nguon, Chiv & Haghnazarian, Osheen**

#### **Instructor: Jongwook Woo**

#### **Date: 11/29/2017**

**Lab Tutorial**

**Los Angeles & Chicago Homicide Crime Analysis Hive**

**Objective**

We will analyze both data sets using Hive to see if crime rate has increased or decreased each year between both cities and compare which city has the most criminal homicide by location. We will visualize the data on a graphs comparing by year the amount of crime and also try to visualize the data using the 3D maps to show where the crime was committed.

**Introduction**

We will use an BigInsight query that analyzes the both crime data files to get insight into how crime differ between both different cities and compare the increase by year. With this analysis, we can see which city has the most crime.

In this tutorial, you'll learn how to use BigInsight to:

• Download and upload multiple csv files

• Create Hive tables to query those crime data

• Create Hive queries to analyze the crime data

• Use Tableau to retrieve the analyzed data and geolocation Map for visualization

• Use Tableau Charts for visualization

**Prerequisites**

Everything you need to go through the scripts and queries is already provisioned with the cluster. To export the analyzed data to Microsoft Excel, you must meet the following requirements:

• You must have Microsoft Excel 2010, 2013 or 2016 installed.

• You must have your Excel 3D Map enabled.

• You must have Microsoft Hive ODBC Driver to import data from Hive into Excel. Select either the 32-bit or 64-bit version based on your version of Microsoft Excel. But, BigInsights does not support it yet as of Sept 2016.

**Crime Data Loaded into BigInsights**

You need to remotely access your BigInsights that you executed in your Bluemix account using *ssh*. Below is the location of the data that are used for this crime rate analysis. You can download the data files from amazon S3 using wget shell commands in the remote node of BigInsights:

|  |
| --- |
| (Los Angeles)  wget -O Crime\_Data\_from\_2010\_to\_Present.csv [https://s3-us-west 1.amazonaws.com/dreab/Crime\_Data\_from\_2010\_to\_Present.csv](https://s3-us-west-1.amazonaws.com/dreab/Crime_Data_from_2010_to_Present.csv)  (Chicago  wget -O Crimes\_-\_2001\_to\_present.cs  <https://s3-us-west-1.amazonaws.com/dreab/Crimes_-_2001_to_present-2.csv> |

You have to upload the files to hdfs folder Crime\_Data. Run the following HDFS commands to create and list Crime\_Data directory in HDFS

|  |
| --- |
| **Creating Los\_Angeles & Chicago Crime Data File directory of HDFS:**  -bash-4.1$ hdfs dfs -mkdir Crime  -bash-4.1$ hdfs dfs -mkdir Crime/Chicago  -bash-4.1$ hdfs dfs -mkdir Crime/Los\_Angeles    **Uploading Los\_Angeles & Chicago Crime Data to hdfs successfully & make sure they uploaded:**  -bash-4.1$ hdfs dfs -put Crimes\_-\_2001\_to\_present.csv Crime Crime/Chicago/  -bash-4.1$ hdfs dfs -put Crime\_Data\_from\_2010\_to\_Present.csv Crime/Los\_Angeles/  -bash-4.1$ hdfs dfs -ls Crime/Chicago/  -bash-4.1$ hdfs dfs -ls Crime/Los\_Angeles/ |

**Creating Hive table to Query Los Angeles Crime Data**

Open hive shell environment as follows:

|  |
| --- |
| $ hive |

**The following Hive statement creates an external table for crime\_data. External tables preserve the data in the original file format, while allowing Hive to perform queries against the data within the file.**

|  |
| --- |
| DROP TABLE IF EXISTS Los\_Angeles;  --create the LACrimeData table on comma-separated crime data    CREATE EXTERNAL TABLE IF NOT EXISTS Los\_Angeles(  DRNumber INT,  DateReported STRING,  DateOccurred STRING,  TimeOccurred STRING,  AreaID STRING,  AreaName STRING,  ReportingDistrict STRING,  CrimeCode STRING,  CrimeCodeDescription STRING,  MOCodes STRING,  VictimAge STRING,  VictimSex STRING,  VictimDescent STRING,  PremiseCode STRING,  PremiseDescription STRING,  WeaponUsedCode STRING,  WeaponDescription STRING,  StatusCode INT,  StatusDescription STRING,  CrimeCode1 STRING,  CrimeCode2 STRING,  CrimeCode3 STRING,  CrimeCode4 STRING,  Address STRING,  CrossStreet STRING,  Location DOUBLE)  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  STORED AS TEXTFILE LOCATION '/user/drea17/'  TBLPROPERTIES ('skip.header.line.count'='1'); |

**Check if the table losangeles is shown**

**hive> show tables;**

**hive> select \* from Los\_Angeles LIMIT 10;**

**Result,**

****

**Creating Hive Queries to Analyze Los Angeles Crime Data**

**Most common crimes from the Los Angeles Dataset from highest to lowest**

|  |
| --- |
| Select CrimeCodeDescription, count (CrimeCodeDescription) AS total from losangeles  Group By CrimeCodeDescription  Sort By Total DESC; |

**Top 5 most common crimes in Los Angeles Dataset**

|  |
| --- |
| Select CrimeCodeDescription, count (CrimeCodeDescription) AS total from losangeles  Group By CrimeCodeDescription  Sort By Total DESC  LIMIT 5; |

**Top 5 Premises where crimes were most committed in Los Angeles**

|  |
| --- |
| SELECT Premise\_Description  COUNT (Premise\_Description) AS total  FROM losangeles  GROUP BY Premise\_Description  SORT BY total DESC  Limit 5; |

**Area with most crimes committed**

|  |
| --- |
| Select areaname, count(\*) AS cntareaname FROM losangeles  Group by areaname  Order by cntareaname  DESC; |

**Query that shows how frequent the most common crime is in the top 5 locations**

|  |
| --- |
| Select CrimeCodeDescription, count(CrimeCodeDescription) AS total, AreaName, Location  FROM losangeles  WHERE CrimeCodeDescription = ‘BATTERY - SIMPLE ASSAULT’  Group By CrimeCodeDescription, AreaName, Location  SORT BY total DESC  LIMIT 5; |

**Query that shows how frequent “Criminal Homicide” is in the dataset**

|  |
| --- |
| Select CrimeCodeDescription, COUNT(CrimeCodeDescription)  AS TOTAL FROM LosAngeles  WHERE CrimeCodeDescription = ‘CRIMINAL HOMICIDE’  GROUP BY CrimeCodeDescription; |

**Query that shows how frequent “Criminal Homicide” is in the top 5 locations**

|  |
| --- |
| Select CrimeCodeDescription, COUNT(CrimeCodeDescription)  AS TOTAL, AreaName  FROM LosAngeles  WHERE CrimeCodeDescription = ‘CRIMINAL HOMICIDE’  GROUP BY CrimeCodeDescription, AreaName  SORT BY TOTAL DESC  LIMIT 5; |

**FINAL OUTPUT**

|  |
| --- |
| INSERT OVERWRITE DIRECTORY '/user/cnguon1/output3/'  SELECT CrimeCodeDescription,address, location  FROM losangeles  WHERE CrimeCodeDescription = 'CRIMINAL HOMICIDE'; |

**Creating Hive table to Query Chicago Crime Data**

|  |
| --- |
| DROP TABLE IF EXISTS chicago;  --create the CHICrimeData table on comma-separated crime data  CREATE EXTERNAL TABLE IF NOT EXISTS chicago(  ID STRING,  CaseNumber STRING,  DateOccured STRING,  Block STRING,  IUCR STRING,  PrimaryType STRING,  Description STRING,  LocationDescription STRING,  Arrest STRING,  Domestic STRING,  Beat STRING,  District STRING,  Ward STRING,  Communityarea STRING,  FBIcode STRING,  XCoordinate STRING,  YCoordinate STRING,  Year STRING,  UpdatedOn STRING,  Latitude STRING,  Longitude STRING,  Location STRING  )  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  STORED AS TEXTFILE LOCATION '/user/YOURUSER/Crime/Chicago/'  TBLPROPERTIES ('skip.header.line.count'='1'); |

**Creating Hive Queries to Analyze Chicago Crime Data**

**Most common crimes from the Chicago Dataset from highest to lowest**

|  |
| --- |
| set hive.cli.print.header=true;  Select primarytype, count (primarytype) AS total  from chicago  group by primarytype  sort by total DESC; |

**Top 5 most common crimes in Chicago**

|  |
| --- |
| Select primarytype, count (primarytype) AS total  from chicago  group by primarytype  sort by total DESC  limit 5; |

**Query to Show Top 5 least committed crimes in Chicago**

|  |
| --- |
| Select primarytype, count (primarytype) AS total  from chicago  group by primarytype  sort by total ASC  limit 5; |

**Top 5 Premises where crimes are most committed in Chicago**

|  |
| --- |
| SELECT locationdescription,  COUNT(locationdescription) AS total  FROM chicago  GROUP BY locationdescription  SORT BY total DESC  Limit 5; |

**District with most reported incidents**

|  |
| --- |
| SELECT district,  count(\*) AS cntdistrict,  XCoordinate,  YCoordinate  FROM chicago  GROUP BY district,XCoordinate,YCoordinate  ORDER BY cntdistrict DESC  LIMIT 5; |

**Top 3 Location Descriptions of where crimes were committed in Chicago**

|  |
| --- |
| Select LocationDescription, count (LocationDescription) AS total  FROM chicago  GROUP by LocationDescription  SORT by total DESC  Limit 3; |

**Area with most crimes committed in Chicago**

|  |
| --- |
| Select block, count(\*) AS cntareaname FROM chicago  Group by block  Order by cntareaname DESC  Limit 20; |

**Query that shows how frequent the most common crime is in the top 5 locations**

|  |
| --- |
| Select PrimaryType, count(PrimaryType) AS total, Block  FROM Chicago  WHERE PrimaryType = ‘THEFT’  Group By PrimaryType, Block  SORT BY total DESC  LIMIT 5; |

**Query that shows how frequent “Homicide” is in the dataset**

|  |
| --- |
| Select PrimaryType, COUNT(PrimaryType) AS TOTAL  FROM chicago  WHERE PrimaryType = ‘HOMICIDE’  GROUP BY PrimaryType; |

**Query that shows how frequent “Homicide” is in the top 5 locations**

|  |
| --- |
| Select PrimaryType, COUNT(PrimaryType) AS TOTAL, Block  FROM Chicago  WHERE PrimaryType = ‘HOMICIDE’  GROUP BY PrimaryType, Block  SORT BY TOTAL DESC  LIMIT 5; |

**Query listing all homicide type crimes with latitude, longitude coordinates**

|  |
| --- |
| SELECT primarytype, latitude, longitude  FROM chicago  WHERE primarytype = 'HOMICIDE'; |

**FINAL OUTPUT**

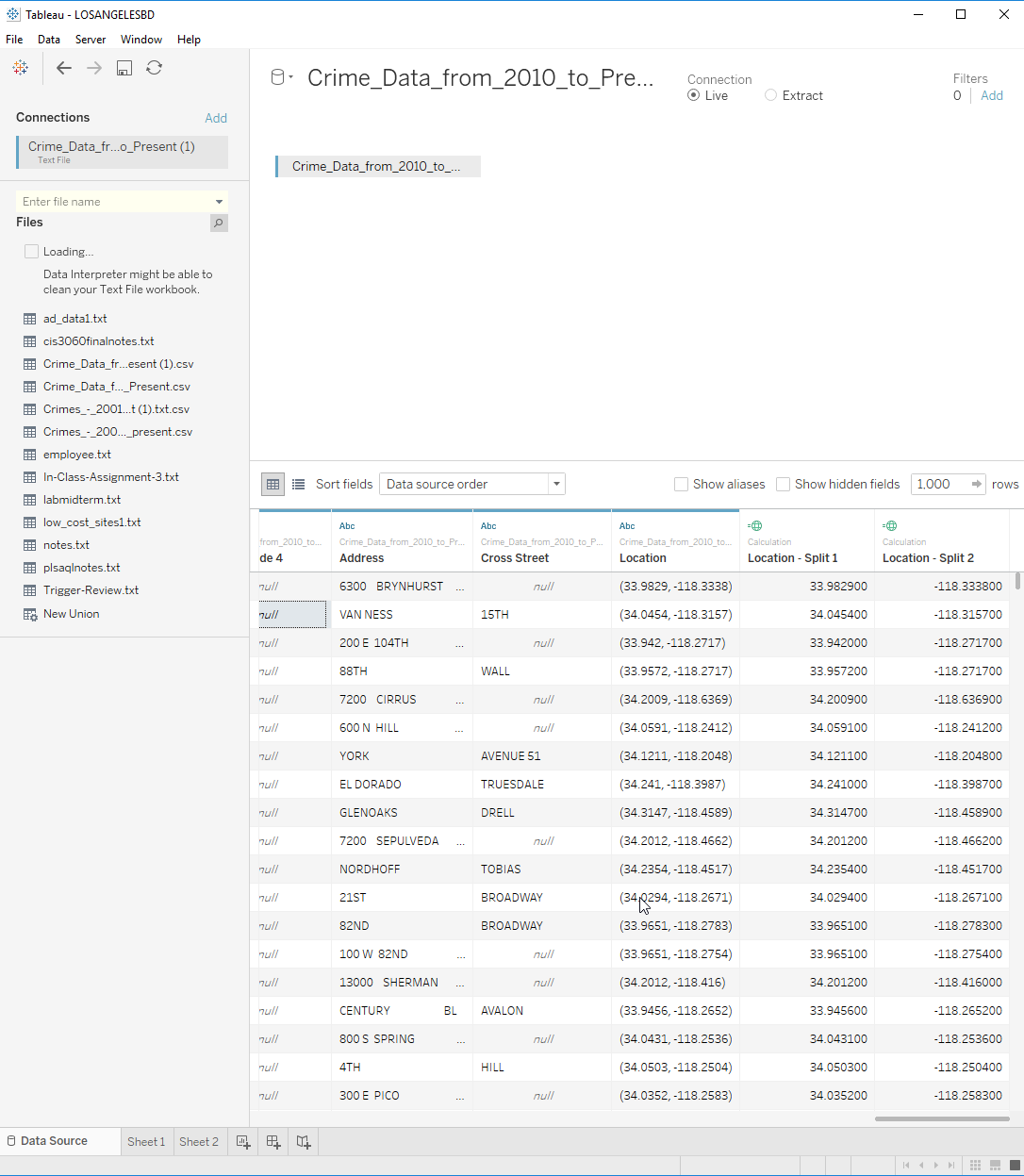
|  |
| --- |
| INSERT OVERWRITE DIRECTORY '/user/cnguon1/output/'  SELECT primarytype, latitude, longitude  FROM chicago  WHERE primarytype = 'HOMICIDE'; |

**Loading Data into Tableau**

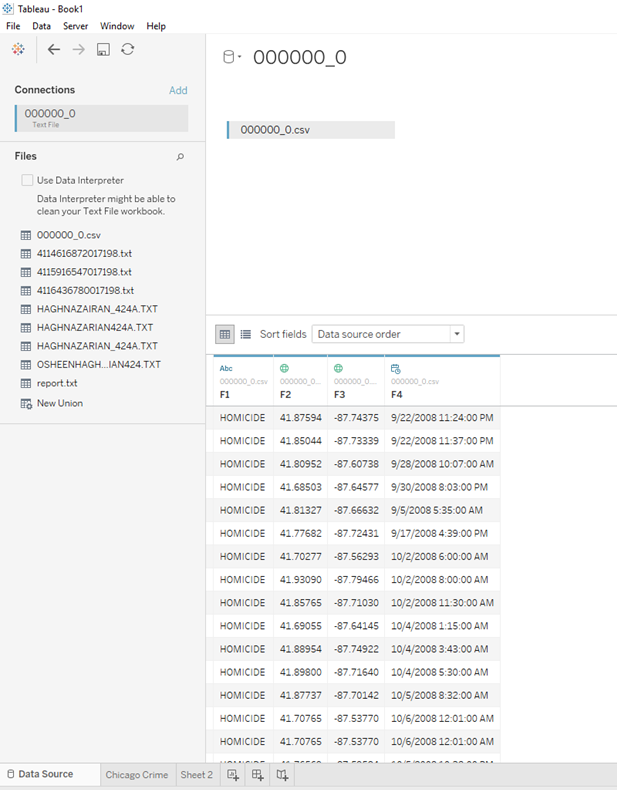
1.Open your Tableau to connect your server. You need to select Text File to open the file 000000-0

2.You will see the following data at Data Source

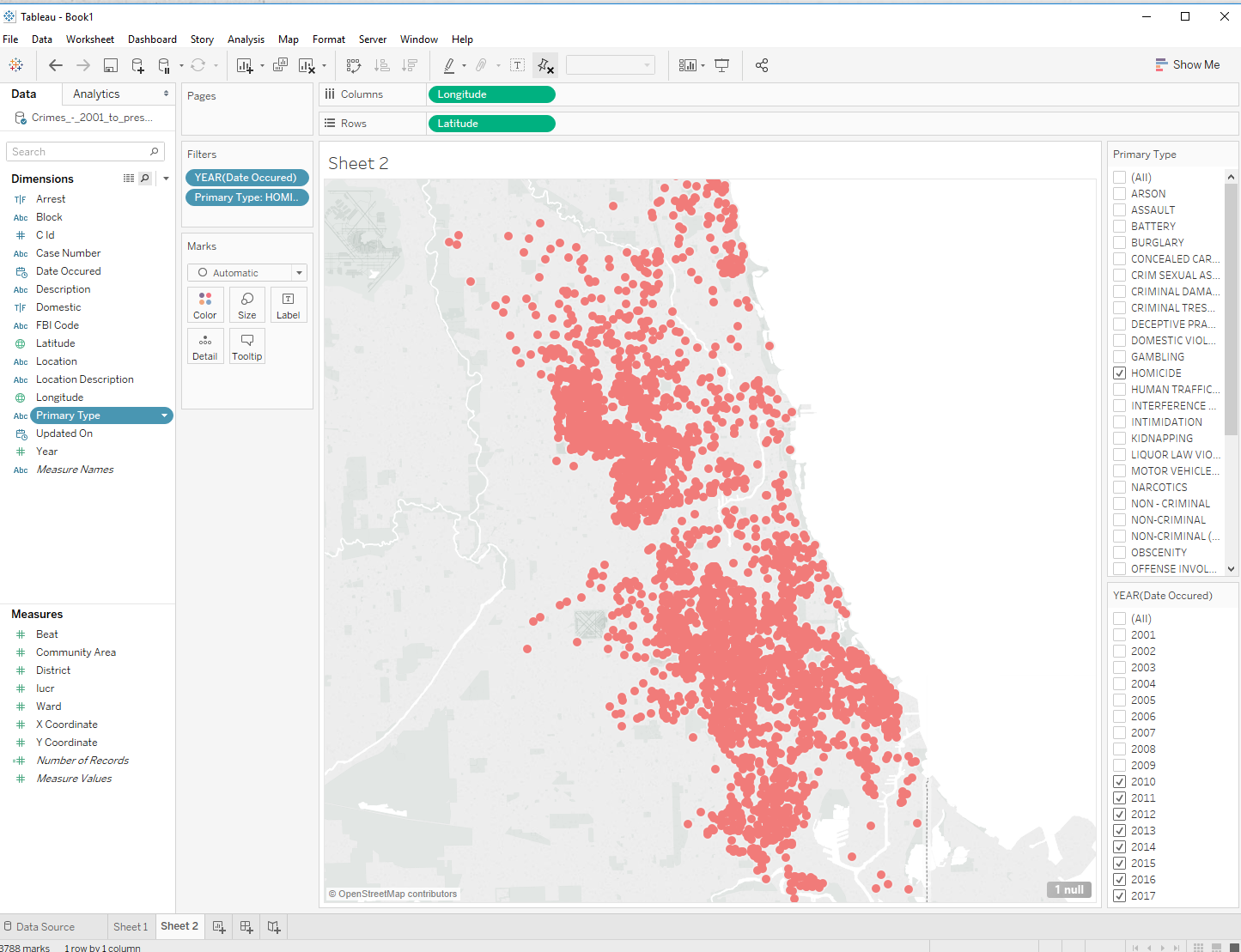
LA LOADING DATA



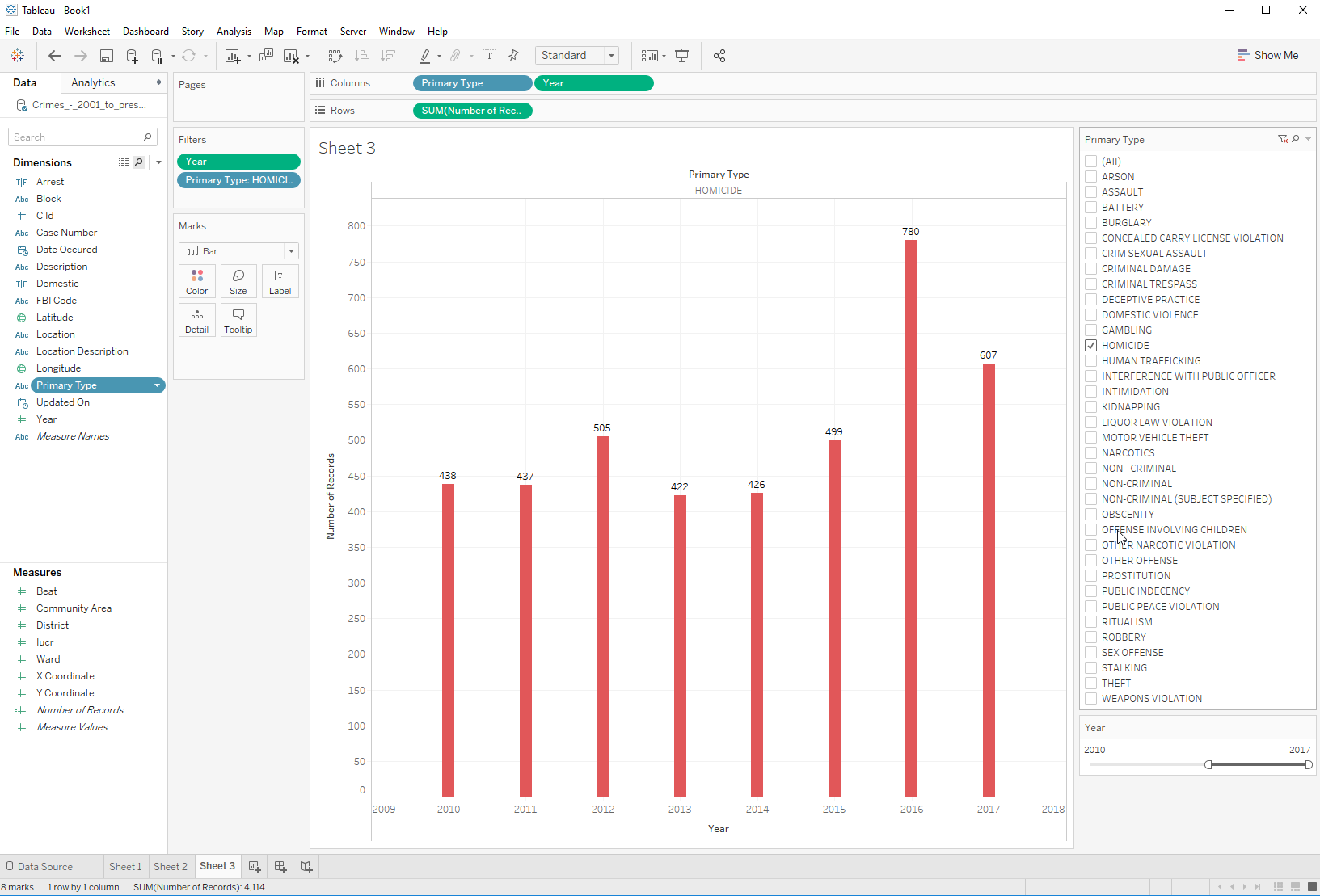
CHICAGO DATA



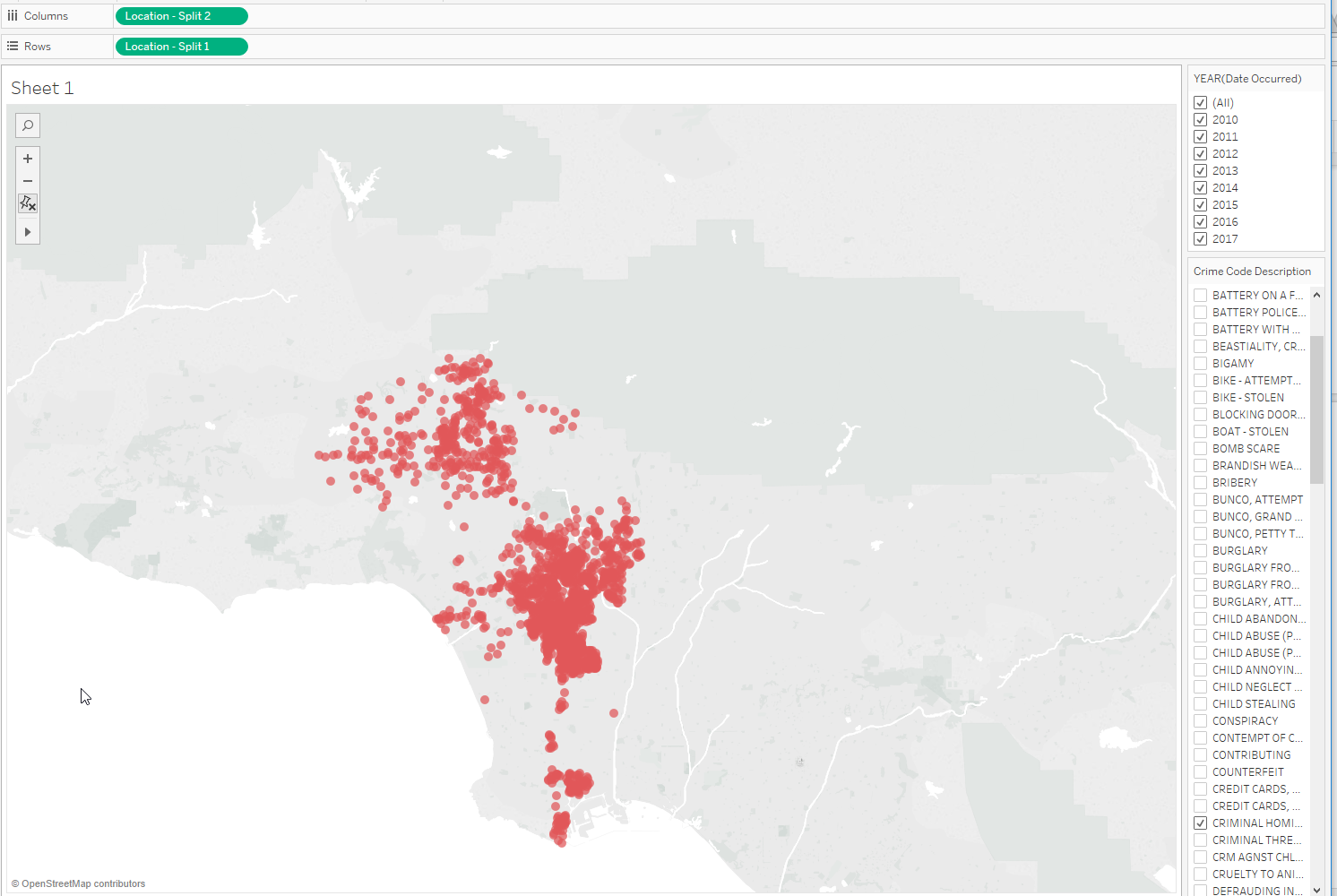
2010-2017 homicide geolocation Chicago



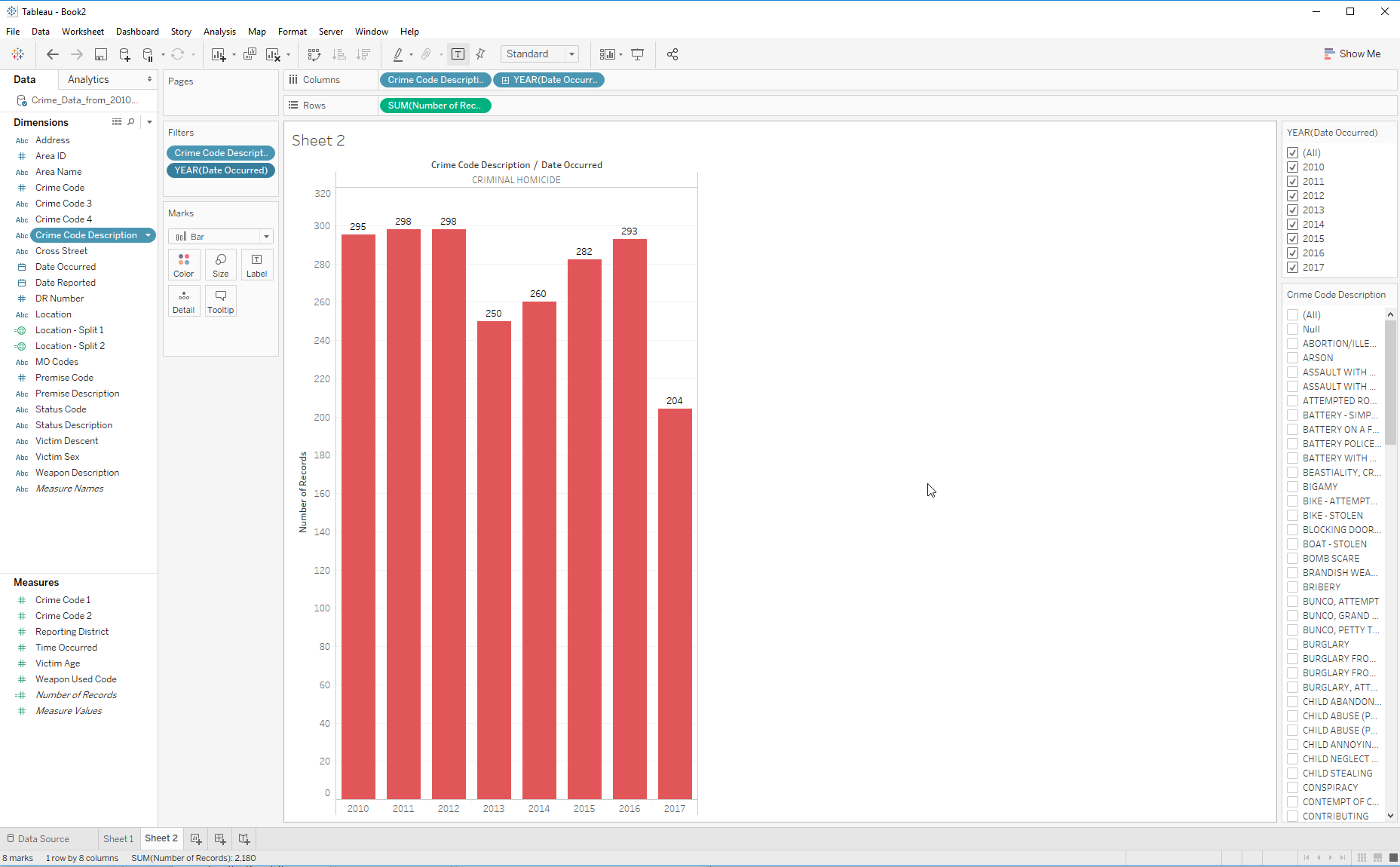
2010-2017 homicide Chicago



Los Angeles geo location



2010-2017 homicide Los Angeles



**Summary**

In this tutorial you learned how IBM BigInsight can be used to analyze data using Apache Hive. You went through a flow to understand how the raw data is first uploaded to HDFS, and then loaded to Hive tables for performing queries. And, you learned how to import the results of Hive queries into Tableau. Finally, you can create GeoLocation and time chart to represent crime location in the map and the amount of crime per year in the Chart to represent the homicide crime in both cities.

**References**

<https://s3-us-west-1.amazonaws.com/dreab/Crime_Data_from_2010_to_Present.csv>

<https://s3-us-west-1.amazonaws.com/dreab/Crimes_-_2001_to_present.csv>