MSc ANALYTICS
BIG DATA PLATFORMS (MSCA 31013)

ASSIGNMENT 2

Big Data Exploration on HDFS and Hive

Problem: Data exploration of Chicago Crimes data using Hive, HDFS and Python

Dataset

Data: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
Metadata: https://dev.socrata.com/foundry/data.cityofchicago.org/6zsd-86xi

Data Loading

For all the questions below provide the commands/queries for HDFS and Hive. You may use multiple queries where applicable. Also submit snapshots of the results/logs in a word or pdf format below each query.

1) Create an external Hive table from this data set called chicago_crimes in a database named as <your userid>. (Try to match the column names from the metadata link above. Ensure that column names have no spaces or special characters)
2) Load all the chicago crimes data (~ 1.5 GB) from 2001 to present from the Chicago city data portal into chicago_crimes.

Data Manipulation - HIVE

Answer the following questions by issuing Hive queries against your table:

3) What are earliest and most recent dates of the crimes recorded in the dataset and what are the types of those crimes. (Dates might vary based on when you download the dataset)
4) List the top 5 and bottom 5 primary crime types based on total count of occurences
5) Which location descripton has the highest number of homicides associated with it ?
6) Which are the most dangerous and least dangerous police districts in the Chicago area?
7) What is the average number assaults per month that occurred in 2018. Has that number increased since the prior period ?
8) From chicago_crimes table create a smaller (summarized) external table in Hive (that supports questions 9 and 10) and download this summarized table to your computer as a CSV file.

Data Visualization - Python

9) Plot a horizontal bar chart with Community (Y axis) and Count of crimes involving children (X axis)
10) Plot a heatmap between Crime Types vs Community and Count (color/number) in each cell.
Community Names: https://www.chicagotribune.com/chi-community-areas-htmlstory.html