MSc ANALYTICS
BIG DATA PLATFORMS (MSCA 31013)

Assignment 2 - Exploratory Data Analysis using Hive

Instructions:

- For all the questions below provide the commands/queries for HDFS/Hive.
- Submit snapshots of the results/logs in a word or pdf format below each query.
- You may use multiple queries where applicable.
- Unless explicitly specified, the question applies to the entire dataset.
- Make assumptions where needed and document them in your notebook

Problem: Data exploration of Chicago crimes data (~ 2 GB) from 2001 to present 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 – HDFS, Hive

1) Load crimes data directly (as crimes.csv) from the city of chicago data portal and store in HDFS
2) Create an external Hive table from this data set called **chicago_crimes** in a database with name as <**your userid> on RCC.** (Try to match the column names from the metadata link above. Ensure that column names have no spaces or special characters)
3) Load data from crimes csv into chicago_crimes Hive table.

Data Manipulation - Hive

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

4) 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)
5) List the top 5 and bottom 5 primary crime types based on total count of occurences
6) Which location descripton has the highest number of homicides associated with it ?
7) Which are the most dangerous and least dangerous police districts in the Chicago area?
8) What is the average number assaults per month that occurred in 2019. Has that number increased since the prior period ?
9) 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

10) Plot a horizontal bar chart with Community (Y axis) and Count of crimes involving children (X axis)
11) 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