MSc ANALYTICS
BIG DATA PLATFORMS (MSCA 31013)

Assignment 3 – Exploratory Data Analysis using Spark

Instructions:

- Provide a single Python notebook **(pynb) file AND a pdf or HTML verison** of the same.
- This assignment must be executed using a big data cluster on the cloud (GCP DataProc or Amazon EMR)
- Build a Spark application and solve the following problems only using PySpark APIs (or Spark SQL).
- Pandas dataframes are not distributed, so you may not use pandas **other than** for plotting summarized data.
- Unless explicitly specified, the question applies to the entire dataset.
- Make assumptions where needed and document them along with your submission.

Problem

- Data exploration of Chicago crimes data (~ 2 GB) from 2001 to present using PySpark
- Continue to work with the Chicago crimes data and load it on to a Data Lake GCS bucket (or Amazon S3)

Datasets:

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
https://data.cityofchicago.org/Health-Human-Services/Daily-Chicago-COVID-19-Cases-Deaths-and-Hospitaliz/kxzd-kd6a

Metadata:

https://dev.socrata.com/foundry/data.cityofchicago.org/6zsd-86xi

Community Names:

https://en.wikipedia.org/wiki/Community_areas_in_Chicago

Exploratory Data Analysis using PySpark

1) PySpark environment and Data Ingestion (10)
    a. Print the configuration settings of the PySpark environment
    b. Read the Chicago crimes dataset into a PySpark dataframe
    c. Print summary statistics of the data set
    d. Inspect the data partitions and repartition if needed

2) Data Transformation (15)
    a. Drop the columns beat, ward, latitude and longitude columns
    b. Convert remaining columns to appropriate data types. Make your best assumptions by sampling the data. View schema again to ensure that data types have been converted.
    c. Add a month column and community name (from metadata) to the dataset

3) Explore data by crime attributes (15)
    a. Group and count crimes where description begins with the word “aggravated”
    b. Which crime type is the most prevalent in apartments and which community has it occurred the most
    c. What is the maximum number of weapons violations per month that occurred in 2020.
    d. What percentage of the domestic crimes led to an arrest ?

4) Explore data by date and time (15)
    a. Which day of the week and which month have the most and the least crimes on average
    b. Which date had the most number of homicides in the dataset. How many days passed between this date and the next highest number of homicides
    c. Plot a monthly time series line chart of all crimes for the last 3 years
    d. Plot a year over year comparison for 3 years (2020, 2021, 2022) by top 5 crime types.

5) Explore by location (15) *Hint: Use spark window functions*
    a. Use a window function to calculate the community rank based on total crime figures (highest to lowest), where the community with the highest crime will have rank 1. Your results set should have 1 row for each community, with a column for the community name and the rank. You can also add a column with the total crime count if it helps you.
    b. Use a window function to calculate a rolling 7 day sum of crimes over time *within each community* Your results set should have 3 columns: community, date, and the rolling/lagging 7 day sum.
    c. Use window functions to calculate a 7 day moving average and cumulative **sum** of crimes over time *within each community*. Your results set should have 4 columns: community, date, the 7 day moving average, and the cumulative sum.
    d. Cross-tabulate Crime Types vs Location description and visualize it through a heatmap

6) Impact of Covid-19 (30)
    a. Bring in daily Covid cases data from the City of Chicago data portal and load into your data lake or Hive table.
    b. Create summarized daily total counts of the daily crime data **by crime type**
    c. Join daily total covid cases and death data with daily chicago crimes data starting Jan 2020.
    d. Perform a thorough analysis in PySpark on how Covid-19 has impacted various types of crimes compared to previous years.

Hint

You can change the specific window function applied, the partitioning, the ordering, and the window frame to solve each of the questions.

Resources:

Spark SQL window functions https://spark.apache.org/docs/latest/sql-ref-syntax-qry-select-window.html

https://www.georgiaruralhealth.org/blog/what-is-a-moving-average-and-why-is-it-useful/

https://stackoverflow.com/questions/45806194/pyspark-rolling-average-using-timeseries-data