Spark and Scala 201 Course Project

Project Context

Crime in Chicago has been tracked by the Chicago Police Department's Bureau of Records since the beginning of the 20th century. The city's overall crime rate, especially the violent crime rate, is substantially higher than the US average. Chicago was responsible for nearly half of 2016's increase in homicides in the US, though national crime rates stayed near historic lows.

Use publically available crime data and develop spark based application to create data visualizations that are possible with a large, comprehensive data set. Each row of data is a story unto itself.

The Proposed application intension is to develop a model for the Chicago Police department to provide platform for deriving high level summary of statistics in terms of no. of crimes, patterns, type of crimes etc., Following are few queries to consider:

1. Find number of crimes that happened under each FBI code.
2. Find number of 'NARCOTICS' cases filed in the year 2015.
3. Find the number of theft related arrests that happened in each district.
4. Find number of crimes happened per year
5. Find number of crimes per month.
6. Where do most crimes take pace?
7. Which days have the highest number of crimes?
8. Calculate number of domestic crimes by day and hour
9. How many number of crime incidents happened on street?
10. What categories of crime exhibited the greatest year-over-year increase between 2015 and 2016?
11. Which month generally has the greatest number of motor vehicle thefts?
12. Number of crimes committed by primary type since 2001

Data Set:

Source: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard/5cd6-ry5g

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. 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, addresses are shown at the block level only and specific locations are not identified.

Disclaimer: These crimes may be based upon preliminary information supplied to the Police Department by the reporting parties that have not been verified. The preliminary crime classifications may be changed at a later date based upon additional investigation and there is always the possibility of mechanical or human error. Therefore, the Chicago Police Department does not guarantee (either expressed or implied) the accuracy, completeness, timeliness, or correct sequencing of the information and the information should not be used for comparison purposes over time.

The dataset contains more than 65,000 records/rows of data and cannot be viewed in full in Microsoft Excel. Therefore, when downloading the file, select CSV from the Export menu. Open the file in an ASCII text editor, such as WordPad, to view and search.

Content

1. **ID -** Unique identifier for the record.
2. **Case Number -** The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
3. **Date -** Date when the incident occurred. this is sometimes a best estimate.
4. **Block -** The partially redacted address where the incident occurred, placing it on the same block as the actual address.

5. **IUCR -** The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description.
6. **Primary Type -** The primary description of the IUCR code.
7. **Description -** The secondary description of the IUCR code, a subcategory of the primary description.
8. **Location Description -** Description of the location where the incident occurred.
9. **Arrest -** Indicates whether an arrest was made.
10. **Domestic -** Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
11. **Beat -** Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts.
12. **District -** Indicates the police district where the incident occurred.
13. **Ward -** The ward (City Council district) where the incident occurred.
14. **Community Area -** Indicates the community area where the incident occurred. Chicago has 77 community areas.
15. FBI Code - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications
16. **X Coordinate -** The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
17. **Y Coordinate** - The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
18. **Year -** Year the incident occurred.
19. **Updated On -** Date and time the record was last updated.
20. **Latitude -** The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.

21. **Longitude -** The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
22. **Location -** The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

Note:

Develop platform with help of RDD's and Spark SQL. Dataset should first be loaded to HDFS then Learners should move it to RDD's and Data frame.

Solution expectation:

Step 1: Load datasets to HDFS
Step 2: Create RDD through external sources such as a shared file system, HDFS
Step 3: Choose any 6 questions from the above and write Spark queries using RDD
Step 4: Create Data Frames
Step 5: Create Spark SQL queries for any 6 questions from the above

Procedure to submit the solution:

1. Submit both solution document for each questions along with screen capture of output from your screen.
2. Solution document should contain respective program/query/script for the corresponding questions.
3. Submit your solution as per guidelines shared by program management team