Specifications

Analysing CSV Data

In this part, you will analyze crime data from South Australia. The dataset reflects reported incidents of crime (suburb-based crime statistics for crimes against the person and crimes against property.) that occurred in South Australia since 2010.

- Crime_Statistics_SA_2010_present.csv

Step 01: Import pyspark and initialize Spark

You will use Spark Context from pyspark , which is the main entry point for Spark Core functionality. The Spark Session object provides methods used to create DataFrames from various input sources.
A DataFrame is equivalent to a relational table in Spark SQL and can be created using various functions in Spark Session .
Once created, it can be manipulated using the various domain-specific-language (DSL) functions defined in Data Frame , Column .

Write the code to create a Spark Context *object, which tells Spark how to access a cluster.* To create a Spark Context you first need to build a SparkConf object that contains information about your application. Give an appropriate name for your application and run Spark locally with as many working processors as logical cores on your machine. Write the code to create a Spark Session object that can be used to create the data frame from the input data source (CSV file).

Step 02: Create Dataframe

Write the code to create a data frame and provide the data source as the CSV file. How many records are there in the data frame?

Step 03: Write to Database

We will use MongoDB as our data source. Therefore, as a data loading step, you are required to *read the CSV file using spark session and insert all the records into MongoDB* . Use the overwrite mode when you are inserting the data.

Step 04: Read from Database

Create a Spark DataFrame to hold data from the MongoDB collection specified in the spark.mongodb.input.uri option which your SparkSession option is using. Display the schema of the data frame. **You will use this new data frame to perform all the steps mentioned below.**

Step 05: Calculate the statistics of numeric and string columns

Calculate the statistics of "Offence Count" and "Reported Date". *Find the count, mean, standard deviation, minimum and maximum for these attributes* . Explain with reasoning whether the minimum and maximum reported date is correct.

Step 06: Change the data type of a column

The Date column is in string format. You need to change it to date format using the user-defined functions (udf).

Step 07: Preliminary data analysis

Write the code to answer the following analytical queries.

- *How many level 2 offences are there? Display the list of level 2 offences.*
- *What is the number of offences against the person?*
- *How many serious criminal tresspasses with more than 1 offence count?*
- *What percentage of crimes are offences against the property?*

Step 08: Exploratory data analysis

Next, write code to analyze the following analytical queries and visualise it using the standard python library - matplotlib . Please make sure you are aware of the different factors such as *visual effects* , *coordinate system* , *data type and scale* and *informative interpretation* before data visualisation as well as you follow *clarity, accuracy* and *efficiency* .

- Find the number of crimes per year. Plot the graph and explain your understanding of the graph.

- Find the number of crimes per month. Plot the graph and explain your understanding of the graph.

- Where do most crimes take place? Find the top 20 suburbs (which would also display postcode for e.g. Caulfield-3162)?. Plot the graph and explain your understanding of the graph.

- Find the number of serious criminal trespasses by day and month. Plot a graph and explain your understanding of the graph.