Problem Statement

Objective

In this case study, you will be working on California Traffic Collision Data Analysis using Apache Spark, a powerful distributed computing framework designed for big data processing. This assignment aims to provide hands-on experience in analyzing large-scale traffic collision datasets using PySpark. You will apply data analytics techniques to clean, transform, and explore crash data, drawing meaningful insights to support traffic safety and urban planning. Beyond understanding how big data tools optimize performance on a single machine and across clusters, you will develop a structured approach to analyzing crash trends, identifying high-risk locations, and evaluating contributing factors to traffic incidents.

Business Value:

Traffic collisions pose significant risks to public safety, requiring continuous monitoring and analysis to enhance road safety measures. Government agencies, city planners, and policymakers must leverage data-driven insights to improve infrastructure, optimize traffic management, and implement preventive measures. In this assignment, you will analyze California traffic collision data to uncover patterns related to accident severity, location-based risks, and key contributing factors. With Apache Spark's ability to handle large datasets efficiently and AWS S3's scalable storage, transportation authorities can process vast amounts of crash data in real time, enabling faster and more informed decision-making.

As an analyst examining traffic safety trends, your task is to analyze historical crash data to derive actionable insights that can drive policy improvements and safety interventions. Your analysis will help identify high-risk areas, categorize accidents by severity and contributing factors, and store the processed data for scalable and long-term storage. By leveraging big data analytics and cloud-based storage, urban planners and traffic authorities can enhance road safety strategies, reduce accident rates, and improve public transportation planning.

Dataset Overview

Context

The dataset used in this analysis consists of California traffic collision data obtained from the Statewide Integrated Traffic Records System (SWITRS). It includes detailed records of traffic incidents across California, covering various attributes such as location, severity, involved parties, and contributing factors. The dataset has been preprocessed and transformed using PySpark to facilitate large-scale analysis. By leveraging Apache Spark, we ensure efficient data handling, enabling deeper insights into traffic patterns, accident trends, and potential safety improvements.

Content

- The dataset is a .sqlite file structured into four primary tables, each containing detailed information about traffic collisions across California.
- **Key attributes in Collisions Table (50+ columns)**
  - **Collision Date**: The exact date of the reported traffic incident.
  - **Primary Collision Factor**: The main cause of the accident as recorded in police reports (e.g., speeding, DUI, failure to yield).
  - **Severity**: The impact of the crash, categorized based on injuries, fatalities, or property damage.

  - Other fields include environmental conditions, roadway characteristics, and lighting conditions at the time of the accident.
- **Key attributes in Parties Table (30+ columns)**
  - **Involved Parties**: Details of individuals and entities involved, including drivers, passengers, and pedestrians.
  - **Party Type**: Specifies whether the individual was a driver, pedestrian, cyclist, or other road user.
  - **Violation Code**: Indicates any traffic law violations committed by the party.
  - Victims Table (20+ columns) - Contains specific details about injuries sustained, including severity levels, protective equipment used, and victim demographics.
- **Key attributes in Locations Table (15+ columns)**
  - **Latitude & Longitude**: Geographic coordinates of the accident site for spatial analysis.
  - **Road Type & Intersection Details**: Information on whether the accident occurred at an intersection, highway, or local road. This dataset enables analysis of accident trends, high-risk locations, and contributing factors, providing valuable insights for traffic management and policy-making.

Acknowledgements

This dataset is sourced from publicly available records from the SWITRS database and is intended for educational and analytical purposes.

Download

The dataset files can be accessed here in the next section.

Scoring and Penalty

- **Total Marks**: **200** (130 for code notebook and 70 for report)
- **Extension and Penalty**: As given in your learner handbooks

Instructions

1. Each learner should attempt this assignment individually.
2. Programming Language: Python
3. You will be provided with the dataset and a starter notebook. You have to perform analyses in the starter notebook only.
4. It is very important that you do not change any headings, subheadings, questions or tasks in your notebook as it can cause problems with grading.
5. For analyses and processing tasks, you should use only the following libraries: NumPy, Pandas, Matplotlib, Seaborn, and Plotly.
6. The data will have inconsistencies and outliers please handle them as per your understanding and mention them in your report.
7. You are encouraged to search the web and consult AI tools for conceptual understanding. However, using plagiarized or AI-generated code is strictly prohibited and strongly discouraged.
8. Submitting plagiarized and AI-generated code or reports will result in significant penalties to your scores.

Submission Guidelines

1. You are required to upload your solution files in the submission field.
2. You are required to submit **two** files:
   (a) an **Interactive Python Notebook** (.ipynb) that contains your code
   (b) a **Report Document** (.pdf) that presents your visualisations, analysis, results, insights, and outcomes.
3. Note that these files should only be generated from the starter files provided to you.
4. Both your Jupyter notebook and report should contain your name and the assignment title in the format: `"ETL_Crash_Analysis_<your_name>"`
5. Mention all assumptions made in the report.
6. Your answers to all the tasks mentioned in the starter notebook should be present in the report. Any graphs/plots you generate for analysis should also be attached to the report.

Results Expected from Learners

Present the overall approach of the analysis in a report document. Mention the problem statement and the analysis approach briefly.

In the starter notebook, you will find headings, subheadings, and checkpoints stating the tasks you need to perform. The marks associated with each checkpoint will also be mentioned in the notebook. Keep in mind not to edit the cells with marking schemes and questions. You can find a brief description of the tasks below.

1. **Data Preparation [5 marks]**
   The dataset consists of structured tables containing traffic collision data. Before conducting any analysis, it is essential to ensure that the data is properly formatted and structured for efficient processing.
   (a) Check for data consistency and ensure all columns are correctly formatted.
   (b) Apply sampling techniques if needed to extract a representative subset for analysis.
   (c) Structure and prepare the data for further processing and analysis.
2. **Data Cleaning [20 marks]**
   (a) Fixing Columns [5 marks]
   (b) Handling Missing Values [10 marks]
   (c) Handling Outliers [5 marks]
   Hints:
   - Note that it is not necessary to replace the missing value in EDA, if you have to replace it, what should be the approach? Mention the approach.
   - Identify if there are outliers in the dataset. Also, mention why you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.
3. **Exploratory Data Analysis and Traffic Collision ETL [95 marks]**
   (a) **EDA: Finding Patterns [40 marks]**
      i. Classify variables into categorical and numerical. [5 Marks]
      ii. Analyze the distribution of collision severity. [5 Marks]
      iii. Examine weather conditions during collisions. [5 Marks]
      iv. Analyze the distribution of victim ages. [5 Marks]
      v. Study the relationship between collision severity and the number of victims. [5 Marks]
      vi. Analyze the correlation between weather conditions and collision severity. [5 Marks]

vii. Visualize the impact of lighting conditions on collision severity. [5 Marks]
viii. Extract and analyze weekday-wise collision trends. [5 Marks]
ix. Assess the number of collisions occurring on different days of the week. [5 Marks]
x. Study spatial distribution of collisions by county. [5 Marks]
xi. Generate a scatter plot to analyze collision locations geographically. [5 Marks]
xii. Extract and analyze collision trends over time. Analyze yearly, monthly and hourly trends in collisions. [10 Marks]

(b) **Querying [55 marks]**
i. Load the processed dataset as CSV files in S3 bucket. [3 Marks]
ii. Identify the top 5 counties with the highest number of collisions. [7 Marks]
iii. Identify the month with the highest number of collisions. [7 Marks]
iv. Determine the most common weather condition during collisions. [7 Marks]
v. Calculate the percentage of collisions that resulted in fatalities. [7 Marks]
vi. Find the most dangerous time of day for collisions. [7 Marks]
vii. Identify the top 5 road surface conditions with the highest collision frequency. [7 Marks]
viii. Analyze lighting conditions that contribute to the highest number of collisions. [7 Marks]

4. **Conclusion [10 marks]**

Final insights and recommendations:

(a) Recommendations to improve road safety by identifying high-risk locations and peak accident times to guide infrastructure improvements and targeted enforcement.
(b) Suggestions for optimizing traffic management by analyzing trends in collision severity, weather conditions, and lighting to improve road design and traffic signal timing.
(c) Propose data-driven policy changes to enhance pedestrian and cyclist safety based on collision trends involving vulnerable road users.
(d) Identify potential high-risk zones for proactive intervention by examining geographic collision density and historical accident data.
(e) Assess the impact of environmental factors such as weather, road surface conditions, and lighting on accident frequency and severity.
(f) Develop predictive models to anticipate collision hotspots and support proactive safety measures.

Points to note:

- Conclude the analysis by summarizing key findings and business implications.
- Explain the results of univariate, segmented univariate, and bivariate analyses in real-world traffic safety and policy terms.
- Include visualizations and summarize the most important results in the report. You are free to choose the graphs that best explain numerical/categorical variables.
- Insights should explain why each variable is important and how they can influence traffic safety policies and urban planning.

5. **Visualization Integration [Optional]**

Enhance the project by incorporating a visualization component that connects the processed data stored in an S3 bucket to a business intelligence tool such as Tableau or Power BI. This involves:

(a) Setting up the connection between the S3 bucket and the chosen visualization tool.
(b) Importing the processed dataset for analysis and visualization.
(c) Creating interactive dashboards to explore key trends and insights.
(d) Ensuring data updates are reflected dynamically in the visualization tool.

Evaluation Rubrics

The following rubrics will be used while judging your solutions to the above tasks.

Table 1: Rubrics

Criteria	Meets expectations	Does not meet expectations
Data Understanding	1. All data quality issues, such as missing values, outliers, and invalid transactions, are correctly identified and documented. 2. Variables are correctly classified into categorical and numerical types, and their meanings are clearly interpreted and explained in the comments or text.	1. Data quality issues, including missing values, outliers, or invalid records, are overlooked or incorrectly identified. 2. Variables are not correctly classified, or their meanings are not provided or are incorrectly explained.
Data Cleaning and Manipulation	1. Data quality issues, such as missing values and redundancies, are resolved appropriately using suitable methods. 2. Data is converted into a suitable and convenient format for analysis using correct techniques. 3. Manipulation of strings and dates is done correctly wherever required	1. Data quality issues are not addressed correctly. 2. The variables are not converted to an appropriate format for analysis. 3. String and date manipulation is not done correctly or is done using complex methods

Continued on next page

Table 1: Rubrics (Continued)

Criteria	Meets expectations	Does not meet expectations
Data Analysis (EDA) ETL	*These come from the analyses in your code and insights reported in your report.* 1. The analysis effectively addresses key traffic safety concerns, identifying trends, patterns, and risk factors related to collisions. 2. The ETL process is correctly implemented, ensuring a clean and structured dataset, including handling missing values and normalizing location-based data. 3. At least five key variables are identified (e.g., collision severity, road conditions, weather, time of day, driver demographics), with a focus on safety insights. 4. Derived metrics (e.g., top 5 counties with the most collisions, highest-risk time periods) are clearly justified and effectively used in analysis. 5. Univariate and bivariate analyses are conducted to uncover meaningful relationships (e.g., collision frequency by weather conditions, accident severity by vehicle type). 6. Key insights are documented with clear explanations, highlighting critical factors contributing to traffic collisions. 7. Well-designed visualizations (e.g., heatmaps, bar charts, line graphs) effectively communicate trends and findings, with clear labels and relevance to conclusions.	*The code and report will be graded for following a reasonable order along with the notebook and implementing the points mentioned here.* 1. The analysis does not effectively address key traffic safety concerns or lacks a clear structure, making it difficult to interpret. 2. The ETL process is incomplete or incorrect, leading to inconsistencies in the dataset. 3. Key variables are not identified, and the analysis lacks depth in exploring collision trends. 4. Derived metrics are missing, irrelevant, or not effectively used in the analysis. 5. Univariate and bivariate analyses are incomplete or fail to uncover meaningful relationships. 6. Insights are missing, unclear, or misinterpreted, overlooking key safety trends. 7. Visualizations are missing, poorly formatted, or unclear, making interpretation difficult.

Continued on next page

Table 1: Rubrics (Continued)

Criteria	Meets expectations	Does not meet expectations
Presentation and Recommendations	1. The report has a clear structure, is not too long, and explains the most important results concisely in simple language. 2. The recommendations to solve the problems or the outcomes and insights, whichever is applicable, are realistic, actionable and coherent with the analysis. 3. If any assumptions are made, they are stated clearly.	1. The report lacks structure, is too long or does not put emphasis on the important observations. The language used is complicated for business people to understand. 2. The recommendations to solve the problems or the outcomes are either unrealistic, non-actionable or incoherent with the analysis. 3. Contains unnecessary details or lacks important ones. 4. Assumptions made, if any, are not stated clearly.
Conciseness and readability of the code	1. The code is concise and syntactically correct. Wherever appropriate, built-in functions and standard libraries are used instead of writing long code (if-else statements, loops, etc.). 2. Custom functions are used to perform repetitive tasks. 3. The code is readable with appropriately named variables and detailed comments are written wherever necessary.	1. Long and complex code is used instead of shorter built-in functions. 2. Custom functions are not used to perform repetitive tasks resulting in the same piece of code being repeated multiple times. 3. Code readability is poor because of vaguely named variables or lack of comments wherever necessary.