**Project Title:**

**Traffic Accident Data Analysis and Visualization US - Case**

**1. Project Overview:**

The project aims to analyze and visualize automobile accident data collected across 49 states in the USA from February 2016 to March 2023. The primary objective is to understand the patterns and factors contributing to accidents, such as road types, weather conditions, and time of day. By utilizing PySpark for data processing, analysis and visualization, PostgreSQL for data storage and preprocessing, and Power BI for further preprocessing and visualization, the project seeks to provide insights that can be used for improving road safety and decision-making.

**2. Project Objectives:**

* **Data Cleaning and Preparation:** Handle missing values, outliers, and feature engineering to prepare the dataset for analysis.
* **Data Processing:** Use PySpark to process , analyze, visualize and transform the data, combining various road types into a consolidated category.
* **Data Storage:** Store processed data in PostgreSQL for efficient querying and analysis.
* **Data Visualization:** Create interactive dashboards in Power BI to visualize accident trends, severity, weather conditions, and other relevant metrics.
* **Deliverables:** A detailed Power BI dashboard, processed dataset in PostgreSQL, and a comprehensive documentation of the process and findings. As the dataset is very huge near to 2.8 GB, we further reduced to publish it on the Power BI to make it less than 1 GB. The github has also limitation on the repositories and we will send the raw the link and processed data via email separately.

**3. Project Scope:**

* **Inclusions:**
  + Data preprocessing and cleaning.
  + Feature engineering and data transformations.
  + Analysis of accident data by severity, weather conditions, time of day and so on.
  + Development of interactive Power BI dashboards.
* **Exclusions:**
  + Real-time data processing or integration.
  + Analysis of data beyond the specified date range or geographic scope.

**4. Data Sources:**

* **Dataset:** US Accident data from February 2016 to March 2023.
* **Type of Data:** Structured data with features like accident severity, weather conditions, timestamps, and location information.
* **Access Method:** Data is Pyspark and fed to PostgreSQL for querying and integration into Power BI.

**5. Data Ingestion: Tools and Technologies:**

* **Tools:**
  + **PySpark:** For data processing and transformation, analysis and visualization.
  + **PostgreSQL:** For data storage and querying as further transformation.
  + **Power BI:** For visualization and dashboard creation here also made to some data pre-processing to reduce it to the desired size.

**6. Data Storage:**

* **Storage Used:** PostgreSQL database for storing processed accident data.

**7. Data Preprocessing and Processing Method:**

* **Processing Procedure:**
  + **PySpark:** Used for handling missing values, combining road types into a Road\_Type category, and performing other transformations.
* **Data Processing:**
  + **Transformations:** Consolidating various boolean road type features into a single Road\_Type column.
  + **Aggregations:** Calculating the number of accidents by road type, severity, and other metrics.
  + **Calculations:** Computing the duration between accident start and end times.

**8. Data Analysis and Visualization:**

* **Analysis Tools:**
  + **Jupyter Notebooks:** Used for data preprocessing and exploratory data analysis.
  + **Power BI:** For creating interactive dashboards and visualizations.
* **Visualization:**
  + **Dashboards:** Include visualizations such as card visuals for summary statistics, line charts for accidents by hour, and maps showing accident locations.
  + **Scatter Plots:** To analyze the correlation between accident duration and severity.

### 9. Challenges and Solutions:

* **Handling Outliers:**
  + **Challenge:** Identifying and managing outliers in the accident dataset was crucial to ensure accurate analysis. Outliers could skew the results and lead to incorrect conclusions.
  + **Solution:** Applied statistical methods and domain-specific rules to detect and address outliers. Implemented data transformation and imputation strategies to mitigate their impact.
* **Data Size Management When Exporting from PySpark to PostgreSQL:**
  + **Challenge:** The dataset processed in PySpark was large, causing challenges during export to PostgreSQL due to size and memory constraints.
  + **Solution:** Optimized the PySpark DataFrame before exporting by filtering and aggregating the data to reduce its size. Adjusted PostgreSQL configurations to handle bulk data inserts efficiently.
* **Handling Data Types When Exporting to PostgreSQL:**
  + **Challenge:** Data type mismatches between PySpark DataFrame columns and PostgreSQL table fields led to errors during the export process.
  + **Solution:** Explicitly casted PySpark DataFrame columns to the appropriate data types before exporting. Mapped the correct PostgreSQL data types during table creation to ensure compatibility.
* **Power BI Desktop Publishing Issue (Organizational Email Requirement):**
  + **Challenge:** Power BI Desktop required an organizational email account for publishing reports, which was a barrier to sharing the dashboard.
  + **Solution:** Investigated alternative publishing options, such as saving the Power BI report locally and sharing it directly, or using a personal email account with Power BI for basic features.

These challenges required creative problem-solving and optimization to ensure the project was completed successfully and within the defined scope.

**10. Team Members and Roles:**

* **Ayele Tesfaye – Project Leader and coordinator**
* **Adisu Tesfaye**
* **Betelhem Yohannes**
* **Tesfu Negase**
* Aschalew Amare