**THE DATA TRIBE: Milestone – 02**

Team Members: Santhosh Reddy Vantikommu, Jyothsna Eleti, Sri Harshika Sattor, Arun Sai Ram Nuvvula, Sreeja Rao Ambati

## 1 Title of Project:

**Improving Safety by Analyzing Big Data on Deadly Crashes from Toronto Police**

## 2 Project Idea:

- This project aims to enhance road safety in Toronto using data on fatal collisions from the Toronto Police Services.

- Key steps include data collection, cleaning, exploratory analysis, predictive modeling, and geospatial analysis to identify high-risk areas.

- Additionally, the project involves launching public awareness campaigns and providing data-driven policy recommendations.

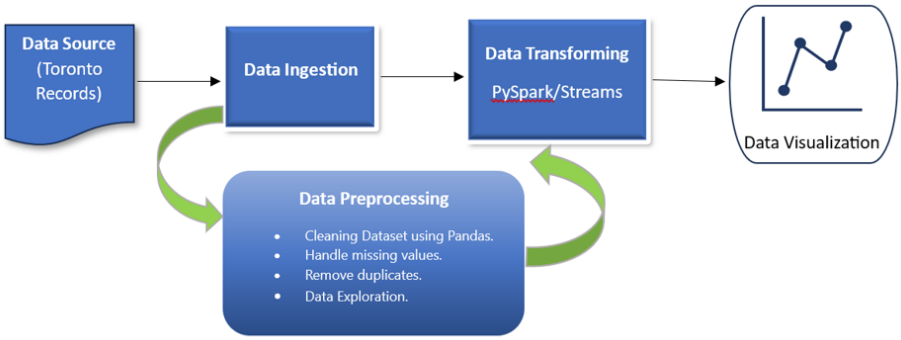
- Goals include reducing fatal collisions, increasing public awareness, and influencing policy decisions through advanced data analytics.

- Ultimately, the project aims to significantly contribute to improving road safety and saving lives in Toronto through big data analytics.

## 3 Tools and Technologies:

1. Programming Language:
   - Python: Python is widely used in big data for its simplicity and extensive libraries, playing a key role in data processing workflows and analytics. It is commonly applied in frameworks like Apache Spark and tools such as Spark for scalable and efficient data processing.
   - Java: Java is pivotal in big data with its role in major processing frameworks like Apache Hadoop. It's also essential for implementing analytics and machine learning algorithms, contributing to scalable and distributed data processing.
2. Data Analysis and Visualization:
   - Pandas: Pandas, a Python library, are employed in big data for data manipulation and analysis, facilitating tasks such as cleaning, transforming, and aggregating datasets efficiently. Its Data Frame structure is instrumental in handling large-scale data within big data processing workflows.
   - Matplotlib: Matplotlib, a Python library, is utilized in big data for creating visualizations and plots, aiding data analysts and scientists in interpreting large datasets. Its capabilities enhance the presentation of insights derived from extensive data in platforms like Jupyter notebooks within big data workflows.
3. Jupyter Notebooks: Interactive development and documentation.

4. Version Control: Git and GitHub for version control and collaborative development.
5. Project Documentation: Microsoft Word for more formal documentation.
6. Communication and Presentation: Microsoft PowerPoint for creating project presentations.




## 4 Architecture Summary:

- Data Source: When raw data is extracted for later transformation and loading into a destination database, it is called a data source in the ETL (Extract, Transform, Load) process. Kaggle is a well-known site for datasets and data science challenges, which we will be utilizing. The dataset is obtained for insights and analysis.
- Data Ingestion: The method of bringing raw data into a processing or storage system from different sources for analysis and modification is known as data intake. PySpark ensures scalability and distributed processing by facilitating smooth data input and effectively loading the Kaggle dataset into a PySpark Data Frame.2
- Data Preprocessing: To improve the quality of raw data, it entails cleaning and arranging it. Pandas are essential for preparing data because they provide strong tools for handling null values, cleaning, and transforming data before transferring it to PySpark.
- Data Transformation: The process of transforming unstructured data into a format of choice is known as data transformation. PySpark and Java Streams work together to provide reliable data transformation by fusing Java Streams' adaptability and distributed processing capabilities to create complex stream-based transformations.
- Data Visualization: The process of representing data using graphical components like graphs, charts, and maps is known as data visualization. It offers

a visual context that makes it easier to see patterns, trends, and insights in the data. The flexible visualization toolkit Matplotlib is used to create static, animated, and interactive charts that provide data exploration as a visual component.

## 5 Project Goals:

- Identify the most common vehicle speed involved in crashes.
- Determine the district with the highest number of fatal car crashes.
- Find the street with the highest number of fatal crashes.
- Analyze the frequency of fatal crashes during the day compared to nighttime.
- Determine the number of deadly collisions on rainy versus dry roads.
- Investigate the year with the highest number of fatal auto accidents.
- Identify the traffic control with the highest fatality rate.
- Determine the exact number of fatal collisions involving drivers exceeding the speed limit.
- List the number of fatal collisions involving pedestrians exercising the right-of-way.
- Determine the drivers' condition during the most fatal collisions.