**1. Introduction and Problem Statement**

Terrorism continues to pose a significant threat to global peace and stability. From organized bombings in conflict zones to isolated attacks in relatively peaceful countries, these acts of violence impact lives, disrupt economies, and challenge the international pursuit of justice and security. In the face of this global issue, one critical question emerges: *Can data be used to make sense of the patterns and evolution of terrorism?*

This study aims to explore this question by leveraging the capabilities of Big Data analytics. Aligned with **Sustainable Development Goal (SDG) 16** — *Peace, Justice, and Strong Institutions* — this project investigates how terrorist attacks have evolved over time, identifies their geographical concentrations, and evaluates their severity. The overarching goal is not merely academic; rather, this project seeks to contribute actionable insights for governments, non-governmental organizations, and policy researchers, thereby enabling more informed responses and preventive strategies.

**2. Dataset Description**

The dataset utilized in this study is the **Global Terrorism Database (GTD)**, maintained by the University of Maryland. The GTD is among the most comprehensive and publicly accessible repositories of terrorism-related data, spanning from 1970 to 2017 and encompassing over 180,000 documented incidents worldwide.

Each record in the dataset contains extensive details, including:

* The **date and location** of the attack,
* The **type of attack** (e.g., bombing, armed assault),
* The **target type** (e.g., civilians, military, police),
* The **weapon used**, and
* The number of individuals **killed** and **wounded**.

Due to the substantial volume and complexity of the data, this study employed **Apache Spark** for distributed big data processing. Analysis was conducted on **Google Colab** using **PySpark**, which allowed for scalable, cloud-based exploration and modeling of the data.

**3. Methodology and Techniques**

This section outlines the sequential process adopted to clean, analyze, and derive insights from the GTD dataset.

**Step 1: Data Loading**

The dataset was uploaded to the Google Colab environment and subsequently read into a **Spark DataFrame** using the spark.read.csv() function. This method allowed for efficient handling of large-scale records, with headers inferred to preserve column labels.

**Step 2: Data Preprocessing**

Data preprocessing involved the following steps:

* **Duplicate entries** were removed using the dropDuplicates() function.
* **Missing values** in the nkill (number killed) and nwound (number wounded) columns were filled with 0 to ensure meaningful clustering and analysis.
* **Corrupted or incomplete rows** (e.g., records with null dates) were filtered out to maintain dataset integrity.  
  To further enhance performance, the dataset was **cached** in memory, and partitioning by year was applied to accelerate time-based queries.

**Step 3: Exploratory Data Analysis (EDA)**

Using **PySpark SQL**, the dataset was queried to identify trends and patterns:

* The number of **attacks per year**,
* **Countries** with the highest number of incidents,
* **Most common attack types**.  
  Aggregations were performed using groupBy() and agg() functions, and key results were converted into **Pandas DataFrames** for visualization using **Matplotlib** and **Seaborn** libraries.

**Step 4: Machine Learning with MLlib**

For the advanced analytical component, **KMeans clustering** from **Spark MLlib** was utilized to classify terrorist attacks by **severity** based on the number of people killed and wounded.

* Data was cleaned further by casting relevant columns to integers.
* A VectorAssembler was used to create a unified feature vector from the nkill and nwound columns.
* **KMeans clustering** was applied with k=4, resulting in four distinct severity levels.  
  The clustered results were visualized, providing a data-driven perspective on the scale and impact of various terrorist incidents.

**4. Key Findings and Visualizations**

The following insights emerged from the analysis:

**4.1. Trends Over Time**

Between **2004 and 2015**, terrorist incidents increased dramatically, with a peak around **2014**. This trend coincides with the emergence and escalation of ISIS and regional insurgencies in countries such as Iraq, Syria, and Libya. The graphical representation of this data indicates a sharp rise in global conflict-related activities in the past two decades.

**4.2. Most Affected Countries**

The highest number of attacks occurred in **Iraq**, followed by **Pakistan**, **Afghanistan**, and **India**. Other significantly affected countries include **Nigeria** and the **Philippines**. This highlights that terrorism is not confined to one region but is, in fact, a **global concern**.

**4.3. Attack Types**

The most prevalent form of attack was **bombings/explosions**, followed by **armed assaults** and **assassinations**. These methods also accounted for the highest number of casualties. The data supports the assertion that **bombings**, while frequent, tend to be **especially deadly**.

**4.4. KMeans Clustering on Severity**

The application of KMeans clustering yielded the following groupings:

* **Cluster 0**: Low-impact attacks (0–1 killed or wounded),
* **Cluster 1**: Medium severity,
* **Cluster 2**: Large-scale attacks (10+ victims),
* **Cluster 3**: Catastrophic events (e.g., 9/11, major bombings).

Most attacks fell under **Cluster 0**, indicating low severity. However, **Cluster 3**, though relatively rare, encompassed **the most devastating incidents**. Visual plots of the clusters revealed a **highly skewed distribution**, where the outliers contributed disproportionately to total casualties.

**5. Conclusion and Future Work**

**Conclusion**

This study successfully demonstrated how **Apache Spark** and **Big Data technologies** can be leveraged to analyze terrorism-related data on a global scale. Key conclusions include:

* A marked **increase in global terrorist incidents**, especially in the last two decades.
* A concentration of attacks in specific countries, with varying attack types and severity.
* The effectiveness of **KMeans clustering** in identifying and interpreting the **severity levels** of attacks.

These findings contribute to a more data-driven understanding of global terrorism and can inform future **risk assessments** and **strategic planning**.

**Future Work**

* Incorporate **real-time data streams** using **Spark Streaming** to monitor emerging incidents through live sources such as news feeds or social media.
* Enhance the **machine learning pipeline** by developing a **classification model** that predicts the likelihood of an attack being deadly.
* Develop a **dashboard** using **Streamlit** or **Power BI** for interactive visualization and public accessibility of insights.