**Credit Card Fraud Detection using Parallel Processing with Dask and Isolation Forest**

**Subject:**

**Parallel and Distributed Computing**

**Team Members:**

* **Akash Kumar A** (22PD03)
* **Kaviya N.S** (22PD16)

**Content:**

This project focuses on detecting fraudulent credit card transactions using a scalable and efficient parallel processing system. By integrating **Dask**, a parallel computing library in Python, and **Isolation Forest**, an unsupervised anomaly detection model, we created a framework that processes large-scale transaction data and detects potential frauds effectively.

The dataset used is from **Kaggle** (mlg-ulb/creditcardfraud), which includes anonymized features of transactions. The system performs data preprocessing, feature engineering, and anomaly detection in a parallelized manner, reducing execution time compared to serial processing.

The detected frauds are compared with the actual labels in the dataset to evaluate the system using **precision and recall** metrics. Visualizations such as fraud distribution by time and amount are used to interpret the results.

**Use of Each Function:**

* **upload\_dataset()**  
  Prompts the user to upload the creditcard.csv file and validates the uploaded file.
* **preprocess\_data(df)**  
  Handles missing values and normalizes numerical features using StandardScaler for better model performance.
* **engineer\_features(df)**  
  Extracts meaningful features like time\_hour and amount\_zscore to enhance fraud detection.
* **detect\_fraud(df, contamination=0.01)**  
  Applies the **Isolation Forest** algorithm to identify outliers (potential frauds) based on the input features.
* **process\_transactions(df, contamination)**  
  Integrates preprocessing, feature engineering, and anomaly detection into one pipeline.

**Technology Used:**

* **Python** – Core programming language.
* **Dask** – For parallel and distributed data processing.
* **Pandas & Dask DataFrame** – For handling and partitioning the data.
* **Scikit-learn** – For using the Isolation Forest model and preprocessing tools.
* **Matplotlib & Seaborn** – For data visualization and fraud pattern analysis.
* **Google Colab** – For cloud-based execution.
* **Kaggle Hub** – For dataset access.