**Project Description:** The goal of this project is to develop a predictive model that can detect potential insurance fraud claims using historical data. The model will be built using PySpark and Python, leveraging the strengths of both technologies to handle large datasets and complex computations.

**Milestones:**

1. **Data Ingestion and Preprocessing**
   * Collect and integrate historical insurance claims data from various sources
   * Clean and preprocess the data to remove missing values, outliers, and irrelevant features
   * Transform(standardize) the data into a suitable format for analysis
2. **Exploratory Data Analysis (EDA)**
   * Perform statistical analysis and data visualization to understand the distribution of variables and relationships between them
   * Identify and select potential features that may indicate fraudulent activity
3. **Feature Engineering**
   * Create new features that can help detect fraudulent claims, such as:
     + Claim frequency and severity
     + Claimant behavior patterns
     + Network analysis of claim relationships
4. **Model Development**
   * Train and evaluate machine learning models using PySpark MLlib, such as:
     + Logistic Regression
     + Decision Trees
     + Random Forest
     + Gradient Boosting
   * Tune hyperparameters to optimize model performance
5. **Model Evaluation and Selection**
   * Evaluate the performance of each model using metrics such as accuracy, precision, recall, and F1-score
   * Select the best-performing model for deployment