**Anti-Money Laundering Detection System**

**Project Overview**

**🎯 Objective**

To detect and classify potentially fraudulent financial transactions using big data tools and machine learning, enabling early identification of money laundering activities.

**🧱 Architecture**

* **Data Source**: Raw transaction data stored in **AWS S3**
* **Preprocessing**: Data cleaned and transformed using **PySpark** and **SparkSQL**
* **Modeling**: Fraud detection model trained using **Spark MLlib**
* **Visualization**: Dashboards built using **Apache Superset**
* **Output**: Predictions on transaction risk, model metrics, and visual insights

**🔧 Tech Stack**

| **Tool** |  |
| --- | --- |
| S3 Bucket | Raw data storage |
| PySpark | Distributed data processing |
| SparkSQL | SQL-based data cleaning |
| Spark MLlib | Machine learning model training |
| Superset | Interactive dashboards |

**📊 Workflow**

1. **Ingest** data from S3 into Spark
2. **Clean & preprocess** using SparkSQL
3. **Train ML model** using Spark MLlib (e.g., Random Forest)
4. **Evaluate** model performance (accuracy, precision, etc.)
5. **Visualize** results using Superset dashboards

**🚀 Future Enhancements**

* **Real-Time Detection**: Integrate Kafka and Spark Streaming for live fraud alerts.
* **Graph Analysis**: Use Neo4j to uncover hidden fraud networks.
* **.Dashboard Expansion :** Add filters and real-time metrics in Superset.

**Output:**

**A graph with blue lines

AI-generated content may be incorrect.**

**Description**: This chart illustrates the daily fraud rate, calculated as the ratio of flagged transactions to total transactions. A rising trend may indicate increased fraudulent activity or improved detection accuracy.

A graph of blue lines

AI-generated content may be incorrect.

**Description**: This visualization highlights regions with the highest concentration of fraud. It can help prioritize investigations or resource allocation for compliance teams.

A pie chart with a number of percentages

AI-generated content may be incorrect.

**Description**: This chart summarizes the model's predictions. A balanced distribution indicates the model is identifying fraud without excessive false positives, while a skewed chart may suggest under- or over-detection.