Dev Gupta *48188665*

Project Proposal:

Fraud Detection in Financial Transactions

Project Overview:

Financial fraud is a pervasive issue, causing staggering losses of over $41 billion globally each year. This project addresses the critical need to combat financial fraud through advanced data analytics techniques.

Motivation: Financial fraud poses a significant threat to individuals, businesses, and the financial industry as a whole. Detecting and preventing fraudulent activities is paramount for preserving financial stability and safeguarding the interests of consumers and institutions.

Overall Objective and Features: This project aims to harness the power of big data analytics to uncover fraudulent patterns within vast financial transaction datasets. Key project features include:

- **Data Analysis**: Analyzing a dataset of 594,000 credit card transactions, which includes transaction amount, merchant details, customer information, transaction categories, and limited demographic data for both customers and merchants.
- **Machine Learning**: Building robust machine learning models using Apache Spark, a distributed data processing framework, to identify fraudulent transactions.
- **Proactive Fraud Prevention**: Empowering financial institutions to take proactive measures against fraud by creating a reliable system for fraud detection and prevention.

Limitations of Traditional Computing Solutions: Traditional computing approaches often fall short in effectively addressing financial fraud due to several limitations:

- **Scalability**: Traditional systems struggle to handle and process the massive volumes of data generated by financial transactions in real-time.
- **Complex Patterns**: Fraudulent activities exhibit intricate and evolving patterns that are challenging to detect using rule-based systems or basic statistical methods.
- **Speed**: The fast-paced nature of financial transactions demands rapid analysis and decision-making, which is often beyond the capabilities of traditional systems.

Benefits Brought by Cloud Computing: This project leverages cloud computing to overcome the limitations of traditional computing solutions:

- **Scalability**: Cloud-based solutions, such as Apache Spark, can easily scale horizontally to accommodate growing datasets and processing demands.
- **Cost-Efficiency**: Cloud computing allows financial institutions to pay only for the resources they use, reducing operational costs compared to maintaining extensive on-premises infrastructure.

  - **Flexibility**: Cloud environments provide flexibility in terms of resource provisioning, making it possible to adjust computing resources based on real-time requirements.

Technical Solutions:

- **Analytics Environment**: Jupyter Notebook
- **Language**: Python
- **Data Processing:** Spark SQL and DataFrames for distributed data queries and preprocessing
- **Modelling**: Spark ML for machine learning model training and evaluation
- **Data Storage**: Cloud Storage for efficient storage of large datasets

Objectives:

- Leverage big data technologies like Spark and MongoDB to handle large volumes of financial transaction data.
- Evaluate model performance using metrics like precision, recall, F1-score to ensure robust fraud detection capabilities.
- Build reusable libraries and components for fraud modelling that can be adapted to new data sources and types of financial fraud.

Architecture:

- The project will leverage a managed Data cluster to execute PySpark jobs. Raw transaction data in CSV format will be loaded from Cloud Storage and converted into Parquet for efficient SQL queries.
- Spark SQL will be used for data exploration, visualization, and feature engineering. Logistic Regression , Decision Tree and Random Forest models will be trained on the processed data using Spark ML. Model hyperparameters will be tuned using cross-validation.
- The final models will be evaluated on a test set for accuracy metrics like F1 score.

Compute

- 1 x n1-standard-2 VM instance (2 vCPUs, 4GB RAM) for Jupyter Notebook and Docker - $25/month

Storage

- 100 MB MongoDB document database - $0.25/GB/month = $0.025/month
- 1 GB disk for VM - $0.04/GB/month = $0.04/month

Networking

- Egress traffic - $12/month (conservative estimate)

Total Estimated Cost:

- Compute: $25
- Storage: $0.025+ $0.04 = $0.065
- Networking: $12

Overall Estimated Total Cost: $27.065 per month

Dataset Source: Fraud-Detection-in-Python/fraudData.csv at master · gouldju1/Fraud-Detection-in-Python (github.com)