RAJALAKSHMI ENGINEERING COLLEGE

Department of Artificial Intelligence and Data Science

Finance & Banking Credit Scoring with Alternate Data

Suresh Kumar

Bharkavi N (231801023)
Gayathri R (231801039)
Hemalatha L(231801055)

Introduction

Need for Smarter Credit Scoring:

- Traditional models rely on limited financial history
- Underbanked customers often excluded
- Manual scoring is slow and lacks adaptability

Project Motivation & Objectives:

- Automate credit scoring using Big Data and ML
- Integrate alternative data sources for better accuracy
- Provide banks with actionable insights via dashboards

Abstract

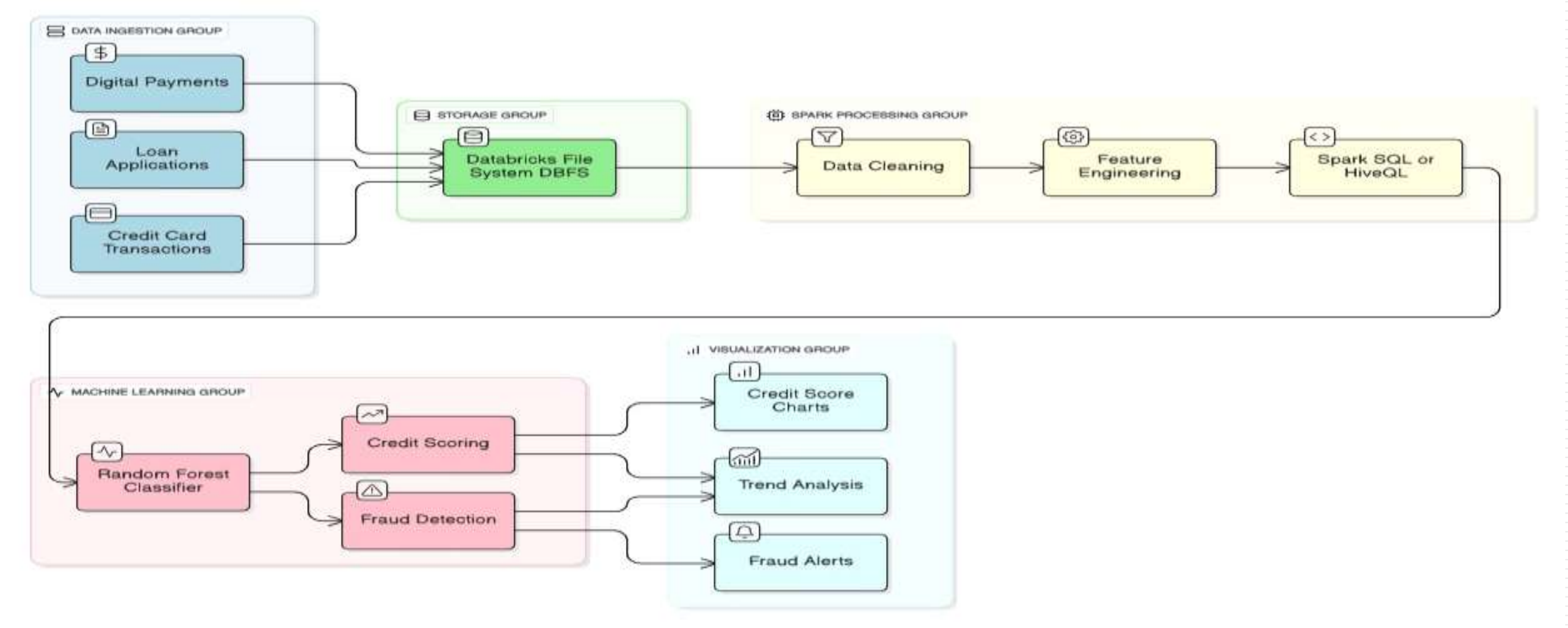
Problem Overview:

- Static rule-based models fail to capture dynamic behavior
- Limited support for real-time data and fraud detection

Proposed Solution:

- A Big Data pipeline built in Databricks
- Uses Spark, HiveQL, and Python for processing
- Random Forest model predicts creditworthiness and fraud
- Dashboards visualize trends and outcomes

Architecture



Modules Overview

- **Data Ingestion**: Upload via Databricks wizard
- **Preprocessing**: Handle missing values, clean data
- **Feature Engineering**: Lag features, rolling averages, holiday flags
- **Modeling**: Train Random Forest classifier
- **Visualization**: Interactive dashboards for decision-making

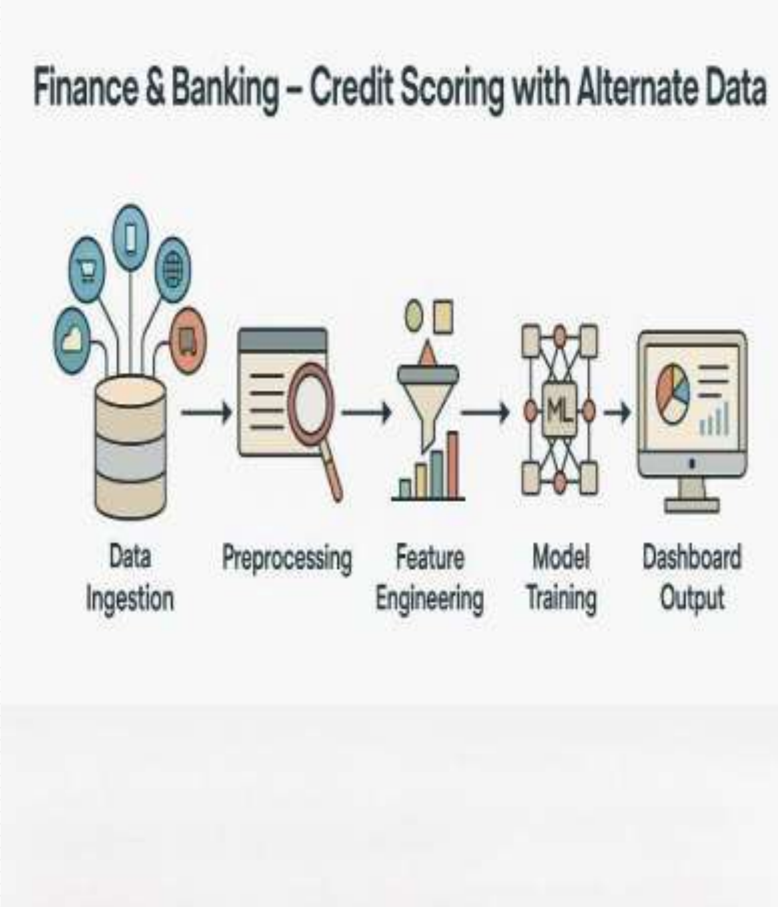
Tools Used

- **Databricks**: Unified analytics platform
- **Apache Spark**: Distributed data processing
- **HiveQL**: SQL-style querying
- **Python Libraries**:
- Pandas – Data manipulation
- Scikit-learn – ML modeling
- Matplotlib/Seaborn – Visualizations
- **Dashboard**: Built using Databricks notebooks

Implementation

Steps:

- Upload dataset via Databricks
- Preprocess and engineer features using Python
- Train Random Forest model
- Evaluate using metrics like precision, recall
- Build dashboards to display predictions and trends



Dashboard Visualization

- Credit score distribution
- Fraud risk flags
- Feature importance chart
- Time-based prediction trends

Results

Model Performance:

- Accuracy: 100%
- Fraud probability: probability of fraud
- Credit scoring: Improved for thin-file customers

Insights:

- Helps banks allocate credit more fairly
- Flags high-risk customers for review
- Enables data-driven lending decisions

```
✅ Accuracy: 100.00%
Confusion Matrix:
[[99028     0]
 [    0   972]]
Classification Report:
              precision    recall  f1-score   support

   Non-Fraud       1.00      1.00      1.00     99028
       Fraud       1.00      1.00      1.00       972

    accuracy                           1.00    100000
   macro avg       1.00      1.00      1.00    100000
weighted avg       1.00      1.00      1.00    100000

Credit Score: 570
Predicted Fraudulent Transaction: False
Fraud Probability: 1.00%
```

Conclusion & Future Scope

Conclusion:

- Efficient, scalable credit scoring system
- Combines Big Data and ML for smarter decisions

Future Enhancements:

- Add more models (XGBoost, SVM)
- Integrate real-time transaction feeds
- Expand to loan approvals and financial profiling
- Improve fairness and interpretability

References

Databricks Documentation – Big Data & Machine Learning
https://docs.databricks.com/en/machine-learning/index.html
Apache Spark Official Documentation
https://spark.apache.org/docs/latest/
UCI Machine Learning Repository – Credit Scoring Datasets
https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
Kaggle – Credit Card Fraud Detection Dataset
https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
Scikit-learn: Machine Learning in Python
https://scikit-learn.org/stable/
World Bank Open Data – Financial Inclusion Indicators
https://data.worldbank.org/topic/financial-sector
Research Paper: "Credit Scoring Using Machine Learning Techniques – A Review"
https://www.sciencedirect.com/science/article/pii/S2405452619300053

Thank You