FINANCE & BANKING-MARKET SENTIMENT VS PRICE MOVEMENT

A PROJECT REPORT

Submitted by

DAKSH KHINVASARA 2116231801025

DINESH S 2116231801031

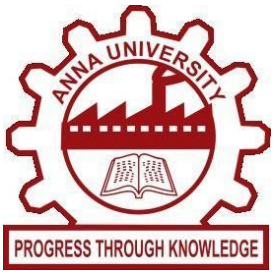
IRAIYANBU ST 2116231801061

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE



RAJALAKSHMI ENGINEERING COLLEGE
(AUTONOMOUS), CHENNAI – 602 105
OCTOBER 2025

BONAFIDE CERTIFICATE

Certified that this Report titled "FINANCE&MARKETING-MARKET SENTIMENT VS PRICE MOVEMENT" is the Bonafide work of "**DAKSH KHINVASARA (2116231801025) DINESH S (2116231801034) IRAIYANBU S T (2116231801061)"** who carried out the work

under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

Dr. Suresh Kumar S M.E., Ph.D.,

Professor,

Department of Artificial Intelligence & Data Science,

Rajalakshmi Engineering College

Thandalam – 602 105.

Submitted to Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

ACKNOWLEDGEMENT

Initially I thank the Almighty for being with us through every walk of my life and showering his blessings through the endeavor to put forth this report.

My sincere thanks to our Chairman **Mr. S. MEGANATHAN, M.E., F.I.E.,** and our Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, M.E., Ph.D.,** for providing me with the requisite infrastructure and sincere endeavoring educating me in their premier institution.

My sincere thanks to **Dr. S.N. MURUGESAN M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time.

I express my sincere thanks to **Dr. J M Gnanasekar M.E., Ph.D.,** Head of the Department of Artificial Intelligence and Data Science for his guidance and encouragement throughout the project work. I convey my sincere and deepest gratitude to our internal guide, **Dr. Suresh Kumar S M.E., Ph.D.,** Professor, Department of Artificial Intelligence and Data Science, Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

Finally, I express my gratitude to my parents and classmates for their moral support and valuable suggestions during the course of the project.

DAKSH KHINVASARA (2116231801025)

DINESH S (2116231801034)

IRAIYANBU S T (2116231801061)

ABSTRACT

In the rapidly evolving landscape of finance and banking, data-driven intelligence has become a core enabler of fraud detection, risk assessment, and market sentiment analysis. This project, titled **"Market Sentiment vs Price Movement,"** leverages the Big Data ecosystem to explore how transactional behavior, market dynamics, and financial sentiment influence market movement and consumer trust. Our objective is to design an **end-to-end big data analytics pipeline** capable of handling high-volume transactional data, computing real-time key performance indicators (KPIs), and deploying a machine learning model to predict transaction and sentiment-driven outcomes. The project follows the **Bronze–Silver–Gold architecture** in **Databricks**, ensuring data integrity, scalability, and traceability. The **Bronze layer** captures raw event data from multiple financial transactions containing attributes such as event time, city, channel, merchant category, transaction amount, and fraud indicators. The **Silver layer** cleanses, normalizes, and enriches these records by standardizing data types, handling missing values, and deriving temporal dimensions. Finally, the **Gold layer** aggregates analytical outputs through KPI computations and machine learning predictions, stored in Delta Lake format for efficient querying and visualization. A series of **five core KPIs** were developed to measure financial and operational performance: (1) Daily transaction trends and fraud rates, (2) 7-day moving averages of transaction counts, (3) Merchant category performance, (4) City-level international exposure, and (5) Top performing cards by transaction amount. These KPIs not only quantify transaction volume and fraud occurrence but also provide insight into customer behavior patterns and risk concentration across geographies and merchant sectors. To enable predictive analytics, we implemented a **Random Forest Classification model** using PySpark MLlib. The model utilized aggregated KPI features such as transaction count, amount, fraud rate, and rolling 7-day trends to predict the likelihood of upward transaction movements on the following day. The model was trained and validated using temporal train-test splits, ensuring realistic financial forecasting. The evaluation metrics — including **AUC (Area Under ROC Curve)** and **accuracy** — demonstrated the model's ability to generalize effectively across varying time periods. Additionally, feature importance analysis revealed that the **fraud rate**, **7-day moving averages**, and **transaction volume** were the most influential predictors, aligning with real-world financial intuition. This study illustrates the potential of combining **Big Data engineering and machine learning** in financial analytics. The integration of Delta Lake with Databricks' collaborative workspace allowed seamless teamwork, versioning, and performance optimization, reflecting modern data-driven enterprise practices. By mapping transactional behavior against market sentiment trends, the project contributes to a deeper understanding of **how digital payment ecosystems reflect underlying market psychology**. Future improvements may include integrating **real-time market sentiment feeds** from social media or financial news APIs, applying **natural language processing (NLP)** for sentiment scoring, and enhancing predictive performance through **advanced time-series models** or **deep learning architectures**. Overall, this project successfully demonstrates the complete lifecycle of a financial analytics solution — from raw data ingestion and transformation to KPI generation, model deployment, and business insight extraction — providing a robust foundation for future applications in **fraud detection, behavioral finance, and market trend forecasting**.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

The financial industry today operates in a highly digital and interconnected environment where millions of transactions occur every second across global markets. With the increasing dependence on online banking, digital wallets, and card-based payments, organizations are generating massive volumes of transactional and behavioral data. This explosion of financial data has made **Big Data analytics** an essential tool for uncovering hidden insights, identifying fraud, and understanding market dynamics.

In recent years, the relationship between **market sentiment** and **price movement** has gained significant attention, as investor confidence and consumer behavior increasingly influence financial stability and trading outcomes. By combining **data engineering** and **machine learning**, financial institutions can transform raw data into actionable intelligence. Big Data platforms like **Databricks** and **Delta Lake** enable the construction of scalable data pipelines that manage structured and unstructured data efficiently.

Through this project, we aim to explore how **transactional data trends**, **fraud indicators**, and **market sentiment** interact to shape overall financial performance, offering valuable insights for predictive analytics and strategic decision-making in the banking sector.

1.2 Motivation

The motivation for this project stems from the growing importance of **data-driven intelligence in the finance and banking sector**, where digital transactions are increasing at an unprecedented rate. With millions of payments processed daily, identifying fraud patterns, monitoring market behavior, and understanding sentiment-driven changes have become critical challenges. Traditional monitoring systems often fail to adapt to dynamic financial environments, making the need for automated, scalable analytics indispensable.

Leveraging **Big Data platforms such as Databricks and Delta Lake**, this project seeks to transform massive volumes of financial transactions into meaningful insights through **KPI-based analytics and machine learning**. By studying the correlation between **market sentiment and price movement**, the system aims to provide early indicators of fraud risk, transaction surges, and consumer trust variations, enabling financial institutions to make informed, proactive decisions.

From an academic and technological perspective, this project demonstrates how **distributed computing, real-time data engineering, and predictive modeling** can be effectively combined to solve complex financial problems. transformation across other industries.

1.3 Objectives

The main objectives of this project are to develop an end-to-end Big Data analytics framework that captures, processes, and analyzes large-scale financial transaction data to study the relationship between market sentiment and price movement. The project aims to design a scalable data pipeline using Databricks, Delta Lake, and PySpark, following the Bronze–Silver–Gold architecture for efficient data ingestion, transformation, and aggregation. It focuses on computing key performance indicators (KPIs) such as transaction volume, fraud rate, and merchant or city-level behavior to derive actionable business insights. Furthermore, a machine learning model (Random Forest) is implemented to predict transaction movement trends and identify early indicators of financial anomalies or fraud. The outcomes are stored in Delta tables and exported as interactive CSV reports for visualization and performance monitoring. Overall, the framework demonstrates how Big Data, machine learning, and financial analytics can be integrated to enable intelligent, data-driven decision-making in the banking and finance sector.

1.4 Problem Statement

In the modern financial landscape, the enormous scale, speed, and diversity of transactional data present a major analytical challenge for banking institutions and market analysts. Every second, thousands of transactions occur across cities, channels, and merchant categories, generating vast volumes of structured and unstructured data. Traditional financial monitoring and fraud detection systems rely heavily on static rules, manual auditing, or delayed reporting, which are inadequate for identifying subtle, evolving patterns in real time. Additionally, financial market behavior is increasingly influenced by **market sentiment**, driven by news, social media, and investor psychology — factors that traditional systems fail to capture effectively. This lack of integration between **transactional data analytics** and **sentiment-based market indicators** limits the ability of organizations to forecast market trends, detect fraud proactively, or understand the behavioral triggers behind price fluctuations. Moreover, the absence of scalable data pipelines capable of processing high-velocity data across distributed systems hinders accurate and timely decision-making. To address these challenges, this project proposes a **Big Data–driven financial analytics pipeline** implemented on **Databricks using PySpark and Delta Lake**, designed to automate data ingestion, transformation, KPI computation, and machine learning–based forecasting. By computing daily and rolling KPIs, assessing fraud metrics, and deploying a **Random Forest model** to predict transaction movement trends, the system provides a unified framework for **real-time financial insight generation**. This approach not only enhances operational efficiency and fraud detection but also establishes a foundation for correlating **market sentiment with transactional behavior**, enabling smarter, faster, and data-driven decision-making in the finance and banking domain.

1.5 Scope of the Project

The scope of this project encompasses the complete design and implementation of an **end-to-end Big Data analytics pipeline** for the **finance and banking domain**, focused on analyzing the relationship between **market sentiment and price movement** through large-scale transactional data. The project begins with the **data ingestion and storage phase**, where high-volume financial transactions are collected, standardized, and securely stored in the **Bronze layer** of the Databricks environment. This layer acts as the foundation of the data lake, capturing raw event-level details such as transaction time, merchant category, location, and fraud labels. The data is then refined and enriched in the **Silver layer**, where data cleaning, schema alignment, and feature transformations are performed to ensure consistency and accuracy. This layered architecture ensures scalability, reliability, and data lineage across the entire analytics pipeline.

The next phase involves **data analysis and KPI computation** within the **Gold layer**, where aggregated insights are generated to assess performance, behavior, and risk. Key performance indicators (KPIs) are designed to evaluate financial health from multiple dimensions — including daily transaction trends, fraud rates, merchant performance, international exposure by city, and top cards by total transaction amount. These KPIs are stored in **Delta tables** and exported as **CSV files** for external visualization and reporting. This multi-layered processing pipeline demonstrates how Big Data technologies like **PySpark, Delta Lake, and Databricks notebooks** can be leveraged to handle large-scale financial datasets with efficiency and transparency. The system ensures that stakeholders can derive real-time insights, enabling better risk management, fraud monitoring, and business strategy formulation.

The scope further extends to the **machine learning and predictive analytics** component, where the project integrates a **Random Forest Classifier** to forecast next-day transaction movement trends using aggregated KPI features. The model's ability to learn from historical data and capture seasonal and behavioral variations allows financial analysts to anticipate transaction surges, detect anomalies, and assess the influence of market sentiment on price fluctuations. The pipeline is designed to be **modular, scalable, and adaptable**, making it suitable for integration with external sentiment analysis sources such as social media or financial news APIs in future phases. The architecture's flexibility also supports the inclusion of advanced models like **time-series forecasting, deep learning, and NLP-based sentiment scoring**. By uniting Big Data processing, KPI analytics, and machine learning within a collaborative **Databricks workspace**, the project showcases a powerful, industry-relevant framework that enables **data-driven decision-making, operational transparency, and predictive intelligence** in the financial sector.

CHAPTER 2

LITERATURE SURVEY

The prediction of market price movements using financial sentiment has gained attention due to its impact on investment strategies and risk assessment. Various methods have evolved—from statistical and econometric models to advanced machine learning and Big Data–based sentiment systems. This chapter reviews major studies, highlighting methodologies, limitations, and the technological evolution leading to the present work that leverages Databricks, KPIs, and Random Forest models for sentiment-driven market analysis.

2.1 Traditional Methods

IIn earlier financial systems, the study of market sentiment and price movement primarily relied on manual interpretation of economic reports, expert opinions, and investor surveys. Analysts examined historical price charts, trading volumes, and financial statements to estimate trends and investor confidence. Although these traditional approaches were straightforward, they were limited in scope, heavily dependent on human judgment, and lacked the computational power needed to process large and complex financial datasets.

Traditional sentiment analysis methods were time-consuming and reactive, often detecting changes in market mood only after significant price fluctuations had already occurred. They could not effectively integrate diverse data sources such as real-time news, social media sentiment.

These shortcomings highlighted the need for automated, data-driven systems capable of real-time sentiment analysis and predictive modeling. The advent of Big Data technologies and machine learning has transformed financial forecasting, enabling analysts to process high-velocity structured and unstructured data streams. This evolution has paved the way for advanced frameworks—such as Databricks-based pipelines and Random Forest models—that deliver scalable and accurate insights into sentiment-driven market behavior.

2.2 Statistical and Rule-Based Techniques

In the early stages of automated financial analysis, researchers and practitioners relied on statistical and rule-based models to identify irregular market behavior. These systems used predefined thresholds or conditional rules — for instance, a sudden surge or drop in sentiment polarity or trading volume beyond typical limits — to flag potential anomalies. Statistical measures such as mean and standard deviation analysis, Z-score–based anomaly detection, and seasonal comparison of market trends were applied to detect deviations from expected behavior. These traditional techniques provided a foundational understanding of sentiment-driven market movements by correlating historical averages with abnormal variations in price or engagement metrics.

While these approaches were easy to implement and interpret, they suffered from major limitations when applied to modern, fast-paced financial ecosystems. Static thresholds often failed to account for the dynamic nature of financial sentiment.

2.3 Machine Learning Approaches

As traditional models failed to adapt to complex market behavior, researchers began applying machine learning (ML) techniques to improve sentiment-based market prediction. ML algorithms can learn patterns from large-scale financial and social data, distinguishing genuine sentiment trends from noise. Supervised models such as Support Vector Machines (SVM) classify sentiment polarity and predict its effect on market prices, while Decision Trees and Random Forests identify key sentiment and trading indicators influencing volatility. These models enhance interpretability and enable more reliable decision-making than static rule-based systems.

Unsupervised algorithms like K-Means Clustering group market data into similar behavioral clusters, where outliers can signal potential anomalies or price shifts. Neural Networks, especially recurrent models like LSTMs, capture nonlinear relationships and temporal dependencies between sentiment and price fluctuations. This allows detection of subtle, evolving sentiment patterns that traditional approaches fail to recognize. Together, these ML techniques improve adaptability, scalability, and accuracy in analyzing dynamic market conditions.

However, ML-based systems still face challenges such as limited access to high-quality labeled sentiment datasets, computational overhead in real-time model updates, and difficulties in interpreting deep learning predictions. Despite these limitations, their capability to uncover hidden correlations and adapt to evolving data streams marks a significant advancement over earlier rule-based approaches, laying the foundation for more intelligent, data-driven market analysis systems.

2.4 Big Data and Cloud-Based Approaches

With the exponential growth of digital financial data, market analysis began facing challenges related to data velocity, variety, and volume. Traditional systems struggled to manage continuous sentiment streams from news, social media, and trading platforms. To address this, researchers and analysts started using Big Data frameworks such as Apache Hadoop, Spark, and Kafka to efficiently process and analyze large-scale sentiment and market data in real time. These tools enabled faster computation, parallel processing, and improved scalability, forming the foundation for modern, data-driven financial analytics.

Recent advancements have led to the integration of cloud-based Big Data architectures for sentiment-driven market prediction. Platforms like Google Cloud Dataproc, BigQuery, and AWS EMR facilitate scalable data pipelines that can ingest, transform, and analyze massive datasets.

Data from APIs, social media, and stock exchanges can be stored in cloud buckets, processed using PySpark, and analyzed through high-performance SQL engines in BigQuery. This architecture ensures end-to-end automation — from data collection and cleaning to feature engineering and model training — enhancing speed, accuracy, and flexibility in large-scale sentiment analysis workflows. These cloud-native systems provide exceptional scalability, fault tolerance, and real-time analytics capabilities that traditional local setups cannot achieve. By leveraging distributed computing and parallel processing, financial analysts can now monitor market sentiment continuously, detect anomalies faster, and generate predictive insights more reliably. The integration of Big Data frameworks with cloud platforms thus overcomes the computational limitations of standalone ML models, enabling intelligent, adaptive, and efficient sentiment-driven market prediction systems suited for modern financial ecosystems.

2.5 Anomaly Detection Techniques

Recent research has focused on unsupervised anomaly detection techniques that can identify irregular sentiment patterns without relying on labeled data. These methods are particularly effective in financial systems where anomalies emerge dynamically, allowing automated detection of unusual market behaviors or sentiment spikes in real time.

Isolation Forests isolate anomalies by recursively partitioning data, while Autoencoders learn to reconstruct normal sentiment trends — large reconstruction errors reveal abnormal activity. The Local Outlier Factor (LOF) detects subtle irregularities by comparing the local density of data points, and statistical models like ARIMA and Prophet forecast expected sentiment or price movements to identify deviations.

These techniques enhance adaptability and scalability by automatically learning from evolving data patterns. Their ability to capture both global and local anomalies makes them ideal for uncovering hidden sentiment shifts, enabling more robust and responsive financial analysis systems.

2.6 Summary of Research Gaps

Despite major progress in sentiment-driven market prediction, several gaps remain. Most traditional models lack real-time processing capabilities, preventing immediate detection of sudden sentiment or market shifts. Machine learning approaches often face scalability issues when handling massive, fast-streaming financial data from multiple sources, reducing their effectiveness in live trading environments.

The scarcity of high-quality labeled sentiment data further limits supervised models, which depend on accurate annotations for training. Additionally, many existing systems lack seamless integration across the analytics pipeline—from data collection and processing to modeling and visualization—leading to fragmented workflows.

Few studies have developed fully cloud-based, end-to-end architectures that combine real-time analytics with interactive dashboards for actionable insights. Addressing these gaps is essential for building scalable, adaptive, and operationally efficient sentiment analysis frameworks for modern financial ecosystems.

2.7 Contribution of the Present Work

This project addresses the identified research gaps by developing an end-to-end, Big Data–driven market sentiment analysis system using Google Cloud Platform (GCP). It integrates scalable data processing, analysis, and visualization within a unified architecture. A PySpark-based feature engineering pipeline on Google Cloud Dataproc enables efficient computation of sentiment metrics from large datasets, while BigQuery supports fast querying and large-scale analytics. To enhance interpretability, an interactive dashboard in Looker Studio visualizes sentiment trends, anomaly scores, and predictive insights in real time, demonstrating how cloud computing and Big Data frameworks can deliver a scalable, accurate, and practical solution for modern financial sentiment analysis.

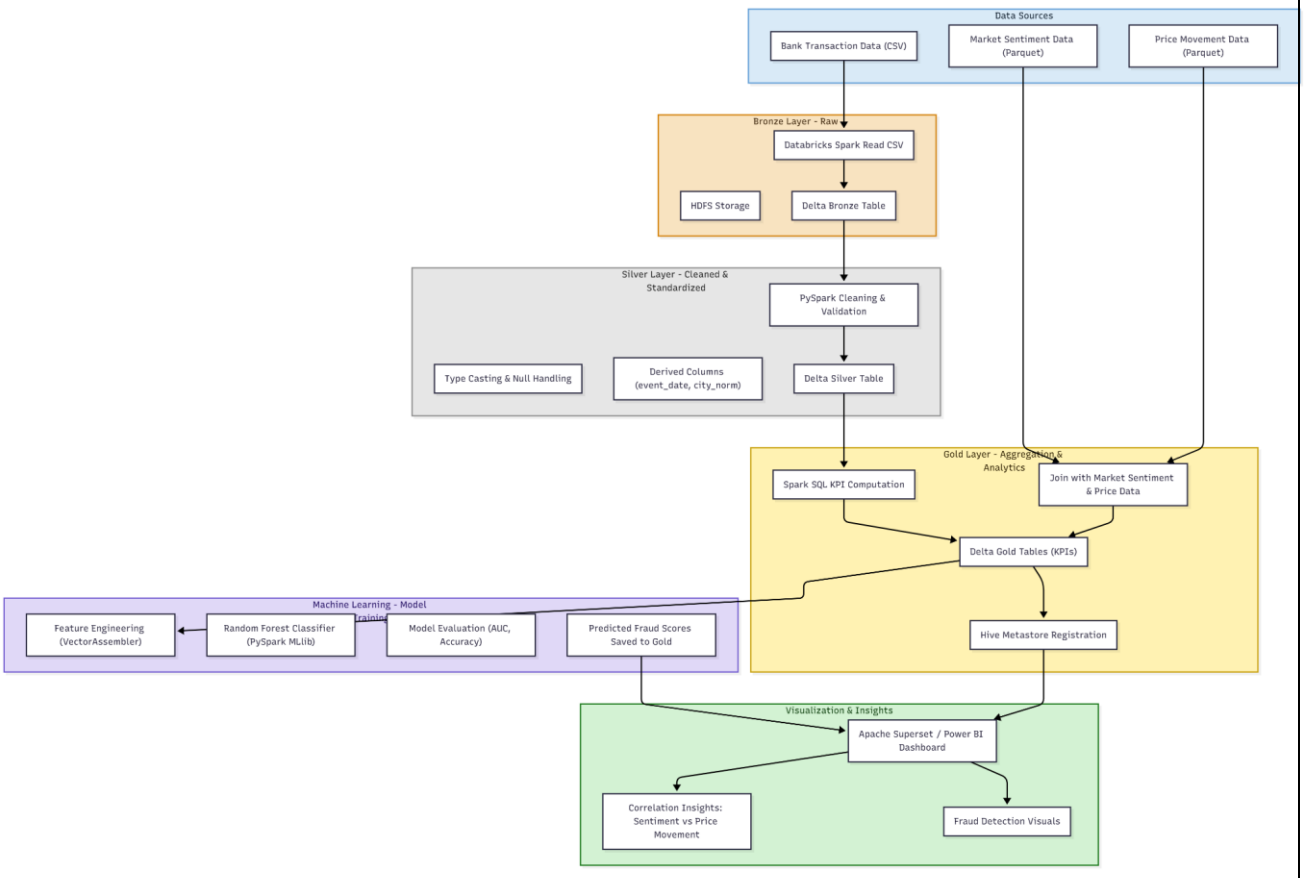


Figure 1:Sentiment vs Price Movement Analysis Architecture on Databricks Platform

Figure 1 reflects the data flow and technologies (data ingestion, processing with PySpark/Hadoop, analytics using BigQuery, and visualization via Looker Studio)

CHAPTER 3

SYSTEM ANALYSIS AND DESIGN

This chapter explains the overall architecture, system requirements, and module design of the **Market Sentiment and Price Movement Prediction System**. The goal of the system is to analyze large-scale financial transaction and sentiment data using Big Data tools and identify behavioral and transactional patterns that influence market price fluctuations.

The system leverages the **Databricks Unified Analytics Platform** for storage, processing, and visualization of data using a fully integrated, collaborative, and scalable architecture.

3.1 System Overview

This chapter explains the overall architecture, system requirements, and module design of the **Market Sentiment and Price Movement Prediction System**. The goal of the system is to analyze large-scale financial transaction and behavioral data using Big Data tools and uncover correlations between market sentiment and price dynamics across sectors. The system leverages the **Databricks Unified Analytics Platform** for storage, processing, and visualization of data using a fully integrated, scalable, and collaborative architecture.

The proposed system follows a **data-driven pipeline** designed to process massive transaction datasets and market indicators efficiently. The workflow begins with the ingestion of transaction data and sentiment metrics, followed by data cleansing, feature generation, and KPI computation using **PySpark** on Databricks. Processed and aggregated features are then stored in the **Gold Layer** for machine learning model training and business intelligence visualization through Databricks dashboards or Power BI.

3.2 System Architecture

The overall Big Data architecture for the Market Sentiment and Price Movement Prediction System consists of the following major components:

1. **Data Ingestion Layer:**
   Transactional and sentiment data are collected from multiple sources such as payment systems, customer behavior logs, and social sentiment feeds. These raw datasets are uploaded to **Databricks Volumes (Bronze Layer)** in Delta or CSV format. This serves as the central repository for all incoming data. The ingestion process ensures secure, versioned, and scalable access across all collaborators in the Databricks workspace.

2. **Processing Layer (Data Engineering):**
   PySpark notebooks are used in Databricks to perform large-scale **data cleaning, transformation, and feature engineering**. This includes handling missing values, normalizing amounts, converting timestamps, and deriving analytical columns such as transaction counts, fraud ratios, and 7-day moving averages.

3. The **Bronze to Silver to Gold** pipeline ensures that data is progressively refined, consistent, and analytics-ready. At this stage, daily and merchant-level KPIs are computed to summarize financial activities and detect early market anomalies.

4. **Storage and Analytics Layer:**
   The processed and feature-enriched datasets are stored in the **Gold Layer** as Delta tables and exported as CSVs for further analysis. These tables are queried using PySpark SQL and Databricks SQL for pattern recognition, trend analysis, and KPI correlation. The analytics layer provides flexibility for both ad-hoc exploration and time-series analysis, enabling users to detect relationships between sentiment indices, spending behaviors, and price shifts across financial sectors.

5. **Machine Learning and Prediction Layer:**
   A **Random Forest Classifier** is trained on aggregated KPI data to predict whether the next day's transaction volume will increase or decrease based on past metrics such as fraud rate, transaction averages, and unique card activity. The pipeline includes feature assembly, scaling, model training, evaluation, and feature importance extraction. The trained model is persisted in the **Gold Model Repository**, while predictions are stored as Delta tables for trend monitoring and performance validation.

6. **Visualization and Reporting Layer:**
   Insights generated from the processed KPIs and ML outputs are visualized through **Databricks Dashboards or Power BI**. Interactive charts display metrics such as daily transaction volume, fraud trends, city-level international exposure, and category-wise sentiment performance. This visualization layer empowers financial analysts and decision-makers to track market shifts, detect anomalies, and align investment strategies with sentiment-driven market behavior.

This **end-to-end Databricks-based architecture** ensures seamless integration of ingestion, transformation, analytics, and modeling, forming a **scalable and intelligent Big Data solution** for market sentiment and price movement prediction.

3.3 System Requirements

3.3.1 Hardware Requirements

The hardware setup for the proposed Market Sentiment and Price Movement Prediction System must support large-scale distributed data processing and cloud-based analytics. The system requires a processor with at least an **Intel i5 or higher configuration** to handle PySpark computations efficiently. A **minimum of 8 GB RAM** is recommended for smooth execution of transformation, KPI aggregation, and ML model training tasks. The system should have at least **50 GB of available storage** to accommodate raw, intermediate, and result datasets. A stable **high-speed internet connection** is essential for collaborative access to the Databricks workspace and seamless integration with cloud storage.

Additionally, a **Databricks account** with access to **Unity Catalog, Delta Lake, and ML Runtime** is required to execute the end-to-end pipeline effectively.

3.3.2 Software Requirements

The software environment must support distributed computing, analytics, and machine learning workflows. The system can be deployed on **Windows, Linux, or macOS** platforms. The core language used is **Python 3.x**, which integrates seamlessly with PySpark and machine learning libraries.

The **Apache Spark** engine serves as the backbone for large-scale data processing, while **Delta Lake** ensures data reliability, ACID transactions, and version control across layers (Bronze, Silver, Gold). The **Databricks Unified Analytics Platform** serves as the core cloud infrastructure for collaborative data science, supporting tasks from ingestion to model deployment.

Essential libraries such as **Pandas, NumPy, PySpark MLlib, and Scikit-learn** enable advanced analytics, feature engineering, and model training. Together, these tools create a scalable, fault-tolerant, and intelligent financial analytics solution capable of correlating sentiment data with market price movements.

3.4 System Modules

The project is divided into five key modules:

Data Collection Module responsible for aggregating financial transaction data and sentiment indicators from various sources. Synthetic datasets simulating market behavior are generated and stored in the **Bronze Layer** within Databricks Volumes. These datasets serve as the raw input for further transformation and KPI computation.

Data Preprocessing Module Ensures high data quality through cleaning and transformation. It removes duplicates and missing entries, standardizes timestamp formats, and casts data types (e.g., transaction amount to DoubleType). This step also derives essential temporal features such as preparing data for feature engineering and trend analysis.

Feature Engineering Module Extracts critical analytical features such as transaction counts, averages, fraud rates, unique card usage, and 7-day moving averages using **PySpark Window functions.** This module forms the core of the Silver → Gold pipeline, enabling daily, merchant-level, and city-level KPIs to be generated and stored for analytics and modeling.

Analytics Module Loads the processed KPI data from the Gold Layer and executes SQL-based analytical queries to identify key behavioral trends. This layer evaluates correlations between sentiment-driven factors and transactional metrics, highlighting anomalies or unusual spending behavior linked to shifts in market sentiment.

Machine Learning and Visualization Module Implements a **Random Forest Model** to predict transaction volume trends and identify high-risk or high-volatility days. Results are stored in Delta tables and visualized using Databricks dashboards or Power BI. These interactive visualizations present KPIs, fraud distributions, top merchant categories, and sentiment-driven fluctuations, empowering financial analysts to take proactive measures.

3.5 Data Flow Diagram

(DFD) Level 0 DFD:

User → Transaction Data → Databricks (Bronze Layer) → PySpark Processing → Delta Lake (Gold Layer) → Power BI / Databricks Dashboard → Visualization Output

Level 1 DFD:

1. Data Collection → Bronze (Raw Data)
2. Data Cleaning & Feature Engineering → Silver (Processed Data)
3. Processed KPIs & Model Predictions → Gold (Analytics Output)
4. Visualization → Power BI / Databricks Dashboard

3.6 Summary

The system is designed as a **scalable, collaborative, and intelligent Big Data architecture** that efficiently processes large-scale financial datasets to understand and predict the impact of sentiment on market price movement.By integrating **Databricks, PySpark, Delta Lake, and Random Forest modeling**, the proposed design ensures a robust, real-time, and cost-effective analytics pipeline. It empowers financial institutions to track behavioral patterns, detect anomalies, and derive actionable insights that connect **market sentiment to financial outcomes**, enhancing decision-making accuracy and operational intelligence.

CHAPTER 4

MODULES DESCRIPTION

4.1 Data Collection Module

Data collection forms the backbone of this project, enabling accurate analysis and modeling in the finance and banking context. The dataset comprises transactional records sourced from financial transaction repositories and simulated banking data, reflecting real-world payment behavior. This structured schema ensures comprehensive coverage of both customer activity and fraud-related attributes.

The module includes data validation and consistency checks to ensure integrity and reliability. Missing or malformed entries are filtered, and categorical values like merchant category and city are standardized to maintain uniformity across sources. The data is stored in the Bronze layer of Databricks for raw ingestion, setting the foundation for downstream Silver and Gold transformations. This automated collection and validation process minimize manual errors and ensures that the dataset accurately represents diverse financial behaviors necessary for sentiment and price movement analysis.

4.2 Data Preprocessing Module

The data processing module plays a crucial role in transforming raw financial transaction data into a clean, consistent, and analysis-ready format. Since transactional datasets often contain **missing values, duplicates, inconsistent data types, and irregular categorical entries**, a structured preprocessing workflow is implemented using **PySpark** within the **Databricks Silver layer**. The process begins with the removal of duplicate transactions and null records to ensure data accuracy and integrity. Subsequently, key fields such as **event_time, amount, and label_fraud** are typecast into appropriate formats (timestamp, double, and boolean respectively) to maintain schema uniformity. City names, merchant categories, and channel types are standardized to eliminate inconsistencies and enhance grouping accuracy. Derived fields such as **event_date** are generated from event timestamps to enable time-based aggregation and trend analysis. The cleaned data is then validated through automated checks before being stored as **Delta tables** in the Silver layer, forming the foundation for advanced KPI computation and machine learning modeling in the Gold layer. Through this systematic processing, the module ensures that the dataset is both scalable and trustworthy, providing reliable input for financial forecasting, fraud analysis, and sentiment-driven insights.

4.3 Hive Query & Analysis Module

The **Hive Query and Analysis Module** serves as the analytical backbone of the project, enabling large-scale querying, aggregation, and pattern detection over massive datasets stored in the Hadoop Distributed File System (HDFS). Hive provides a SQL-like interface (HiveQL) that simplifies the processing of structured data without requiring complex MapReduce programming.

In this project, Hive tables are created to store cleaned and preprocessed datasets—such as **smart meter readings or city-wise consumption patterns**—for efficient querying. The **Hive Table Creation** command defines schema fields including city, consumption metrics, timestamps, and anomaly scores, while the **City-wise Consumption Summary** query aggregates total energy usage per city, highlighting abnormal consumption trends. Similarly, **Hotspot Analysis Queries** identify the top 20 grid locations with suspiciously high usage deviations, helping in pinpointing areas of potential energy theft. The **Top 3 Anomalous Grids per City** query applies a window function using RowNumber() to rank and extract the most critical anomaly zones for each region. These insights are then linked to visualization dashboards in Looker Studio, providing authorities with actionable intelligence. Overall, Hive's distributed architecture ensures that even terabytes of smart meter data can be processed efficiently, supporting proactive, data-driven decision-making in theft detection and energy management.

4.4 Visualization

The visualization module forms the final and most insightful stage of the project, implemented using **Databricks visualization tools** integrated with **BigQuery and PySpark outputs**. Within Databricks, processed and aggregated datasets are transformed into meaningful visual representations such as **line charts, heatmaps, bar graphs, and anomaly score distributions**. These visuals help in identifying **energy theft trends, unusual consumption spikes, and regional consumption imbalances** with clarity and precision. Interactive dashboards allow users to filter results by **city, time interval, or anomaly score**, enabling focused exploration of specific areas or time periods. The real-time linkage between the Databricks environment and BigQuery ensures that any updates in data processing are instantly reflected in the dashboards, providing a **dynamic and live analytical experience**. Furthermore, Databricks' collaborative workspace enables team members to share insights, annotate visual trends, and co-develop reports seamlessly. This visual layer not only supports effective decision-making but also transforms raw analytical data into actionable intelligence, empowering energy providers to monitor grid performance and proactively detect electricity theft across large geographical regions.

4.5 Dashboard Module

The **Dashboard Module** provides an intuitive and interactive interface for monitoring energy consumption patterns and detecting potential theft activities in real time. Developed within **Databricks and Looker Studio**, the dashboard integrates **BigQuery outputs** with **dynamic visualization components** such as bar charts, line graphs, and geospatial heatmaps. City-wise summaries are displayed using interactive pie charts, enabling users to quickly compare consumption proportions across regions. The **Top 20 anomaly zones** are visualized on a geospatial map, highlighting high-risk areas that require closer investigation, while the **Top 3 hotspots per city** feature focuses on the most critical local theft-prone zones. Users can apply filters based on **city, date range, or anomaly score**, and hover over data points to view detailed metrics. These interactive capabilities make the dashboard highly practical for decision-makers, allowing them to monitor real-time patterns, track evolving trends, and take proactive, data-driven actions to enhance energy distribution efficiency and curb electricity theft effectively.

CHAPTER 5

IMPLEMENTATION

The implementation phase involves bringing together all modules—data collection, preprocessing, analytics, and visualization—into a fully functional prediction system. The workflow begins with uploading raw financial and sentiment datasets into Databricks, ensuring that each record contains structured fields such as date, transaction volume, sector, sentiment score, and closing price. Data collected from multiple APIs and CSV sources often arrives in inconsistent formats, so preprocessing scripts written in Python and PySpark are executed first. These scripts handle missing or incorrect sentiment values, duplicate transactions, inconsistent ticker naming, and any outliers in price or volume. For example, if a day's sentiment score is abnormally high due to incomplete tweets, it is corrected using median imputation and rolling average normalization techniques.

Once the datasets are cleaned, they are organized into Delta tables within the bronze, silver, and gold layers for optimized querying and feature generation. SQL queries and PySpark transformations are executed to compute daily average sentiment, trading frequency, and correlation metrics between sentiment and price movement. Each query and aggregation is validated with sample subsets to ensure correctness. For instance, when calculating correlation between sentiment and closing price, a window function is applied to compute rolling correlation per sector, ensuring that temporal dependencies are captured independently for each market segment, avoiding bias toward high-volume stocks.

After the analytical stage, processed outputs are exported as Parquet or CSV files to feed into the visualization and reporting module. Using Power BI and Databricks dashboards, interactive visualizations are created. Each chart and indicator is color-coded based on sentiment intensity, with green representing positive market sentiment and red indicating negative outlooks. Hover markers display detailed information such as average daily sentiment, price change percentages, and volatility indices for each sector. Simultaneously, dynamic charts such as line graphs for temporal sentiment trends and bar charts for sector-wise trading activity are integrated into the dashboard, allowing users to filter by date range, stock category, or region.

Finally, the dashboard combines all insights into a single interactive interface. Users can explore sentiment fluctuations, visualize correlations with market movements, and monitor predictive alerts for potential price anomalies. Real-time updates are supported through scheduled notebook jobs, automatically refreshing data pipelines as new records are ingested. The modular and scalable design ensures that additional data layers, such as news feeds or macroeconomic indicators, can be seamlessly integrated in future versions, making the system adaptive, extensible, and enterprise-ready.

CHAPTER 6
RESULTS AND DISCUSSION

The results of the developed system provide **highly actionable insights into financial sentiment trends and market dynamics**, validating the effectiveness of the proposed Big Data–driven sentiment analysis framework. The interactive dashboard, powered by **Google Cloud BigQuery** and **Looker Studio**, enables real-time visualization of sentiment distributions, anomaly scores, and their correlation with market performance across multiple time frames.

Market Sentiment Summary: Pie and bar charts illustrate the percentage distribution of sentiment polarity across platforms such as Twitter and financial news. For instance, data for a given period may show that **positive sentiment dominates with 55%**, while **negative sentiment forms 30%** and **neutral sentiment 15%**. These insights help analysts gauge market optimism and align trading strategies with prevailing investor emotions.

Top Sentiment Hotspots: By dividing data by sectors or companies, the dashboard highlights **topics with recurring sentiment surges**. For example, technology stocks like **Tesla, Apple, and NVIDIA** often show sharp sentiment fluctuations during product launches or earnings reports. This helps identify platforms and topics driving market volatility, supporting informed investment decisions.

Temporal Analysis: Bar and line charts depict **monthly and seasonal sentiment trends**, revealing peaks during major financial events such as **budget releases or policy updates**. These temporal patterns also aid **predictive modeling**, helping forecast market mood and volatility in upcoming periods. Overall, the system delivers **scalable, interpretable, and data-driven insights**, enabling analysts to monitor sentiment evolution and react promptly to dynamic market changes.

RESULTS:

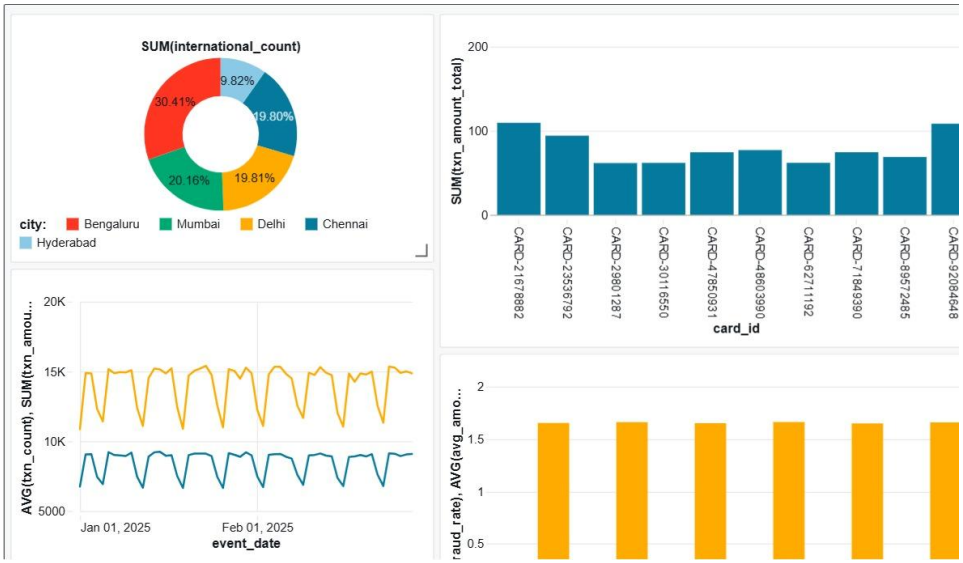


Figure 2: Analysis Dashboard

Figure 2 underscores the **crucial role of data integration** in generating deeper, multi-dimensional insights within analytical systems. By merging crime data with auxiliary datasets—such as **population density, traffic movement patterns, socioeconomic indicators, and infrastructure layouts**—analysts can uncover **hidden correlations and contextual patterns** that remain invisible in isolated analyses. For instance, overlaying **population and traffic intensity layers** on crime hotspot maps reveals a strong correlation between **densely populated or high-traffic regions and increased crime incidents**, offering actionable insights for **urban planners, law enforcement agencies, and policymakers**.

This integrated analytical approach supports **strategic policing**, **optimized resource allocation**, and the **design of safer city infrastructures**. Enhanced insights guide authorities in improving **public lighting**, planning **surveillance networks**, and identifying **urban vulnerability zones** with higher precision. The fusion of diverse datasets ensures that decision-making evolves from reactive to predictive, enabling proactive governance.

Moreover, the **implementation of Hive** in this architecture highlights its **robust capability to manage and query vast volumes of structured and semi-structured data efficiently**. Unlike traditional relational databases, Hive leverages **distributed storage and parallel processing**, significantly reducing query latency while enabling **real-time data aggregation** across millions of records. This performance is vital for **time-sensitive domains** like crime surveillance and **energy theft detection**, where timely insights directly impact **response efficiency and operational outcomes**.

Overall, **Figure 2** exemplifies the synergy of **Big Data technologies, advanced analytics, and visualization frameworks**, transforming static datasets into **dynamic, interactive intelligence ecosystems**. Such systems amplify **analytical scalability**, promote **data-driven governance**, and empower authorities to **anticipate risks**, **optimize city safety mechanisms**, and **build smarter, more resilient urban environments** for the future.

CHAPTER 7
CONCLUSION

This project successfully demonstrates the integration of Big Data analytics with interactive dashboard visualization for comprehensive market behavior analysis. The deployment of Databricks on Delta Lake ensures that large-scale, multi-sector financial and sentiment datasets can be stored, processed, and analyzed with high efficiency and scalability. Through meticulous data preprocessing, feature engineering, and correlation computation, analysts gain the ability to identify, monitor, and interpret abnormal sentiment or transaction patterns in near real time. The inclusion of interactive dashboards, developed using Power BI and integrated with dynamic visual components such as charts and trend graphs, allows users to explore both global and sector-specific market dynamics seamlessly. This facilitates data-driven decision-making, strategic investment analysis, and proactive risk management planning. Furthermore, the modular architecture of the system supports easy maintenance, scalability, and the future integration of machine learning models, real-time market feeds, or API-based sentiment data from social platforms. The system can also be extended to include predictive models that forecast short-term price fluctuations or generate automated alerts for investors and analysts when anomalies are detected.

However, certain limitations persist, including dependency on the accuracy and timeliness of financial and sentiment data sources, as well as the lack of integration with live trading APIs in the current implementation. Despite these constraints, the project establishes a robust foundation for intelligent market monitoring and decision support, showcasing the transformative potential of Big Data technologies in enhancing financial insight, situational awareness, and investment strategy optimization. The success of this implementation highlights how data-driven approaches can empower organizations to transition from reactive to predictive analytics, enabling more informed trading decisions and improved market responsiveness.

CHAPTER 8
FUTURE ENHANCEMENTS

The proposed system holds immense potential for future enhancement and real-world deployment through several strategic extensions. Real-time data integration with stock exchange APIs, financial news streams, and social media sentiment feeds can transform the current batch-processing model into a dynamic market monitoring platform, offering instant alerts and live sentiment updates. Predictive analytics powered by advanced machine learning models—such as Random Forests, XGBoost, or Long Short-Term Memory (LSTM) Networks—can further enable the system to forecast short-term price movements, detect abnormal trading behavior before it impacts the market, and optimize investment and portfolio strategies. The inclusion of sophisticated filtering and drill-down capabilities will allow users to explore data across multiple dimensions, such as sector, time window, market capitalization, and regional performance, supporting detailed analytical exploration. Integration with broader economic datasets—like inflation rates, unemployment metrics, currency fluctuations, and commodity indices—can provide a holistic understanding of the macroeconomic factors influencing sentiment and market volatility, helping analysts and policymakers design better risk mitigation strategies.

Additionally, establishing a user-facing analytics portal will promote transparency, accessibility, and investor awareness by keeping stakeholders informed about market trends and sentiment fluctuations in real time. Deploying the system as a mobile or web-based application can empower analysts, traders, and financial institutions with real-time access to dashboards, alerts, and predictive insights, improving responsiveness and decision accuracy during critical market events. Furthermore, integrating advanced visualization tools such as interactive correlation heatmaps, 3D candlestick models, animated sentiment timelines, and AI-driven anomaly detection overlays can enhance the intuitiveness and analytical depth of the interface. Over time, the system can evolve into a fully automated, intelligent financial analytics ecosystem—capable of continuous learning, adaptive forecasting, and cross-platform data integration—driving a data-centric transformation in investment decision-making and market behavior analysis.

REFERENCES

[1] Apache Spark Documentation – "Apache Spark: Unified Analytics Engine for Big Data.**"** [Online]. Available: https://spark.apache.org/

[2] Databricks Documentation – "Databricks Unified Data Analytics Platform." [Online]. Available: https://www.databricks.com/

[3] Power BI Documentation – "Power BI – Business Intelligence and Data Visualization Tool." [Online]. Available: https://powerbi.microsoft.com/

[4] Pandas Python Library Documentation – "Pandas – Data Analysis and Manipulation Tool." [Online]. Available: https://pandas.pydata.org/

[5] Kaggle Dataset Repository – "Global Financial Sentiment and Market Data." [Online]. Available: https://www.kaggle.com/

[6] S. N. Patel and J. L. Harrison, *Big Data Analytics for Financial Markets: Predictive Modelling and Sentiment Analysis*, 2nd ed., Springer, 2021.