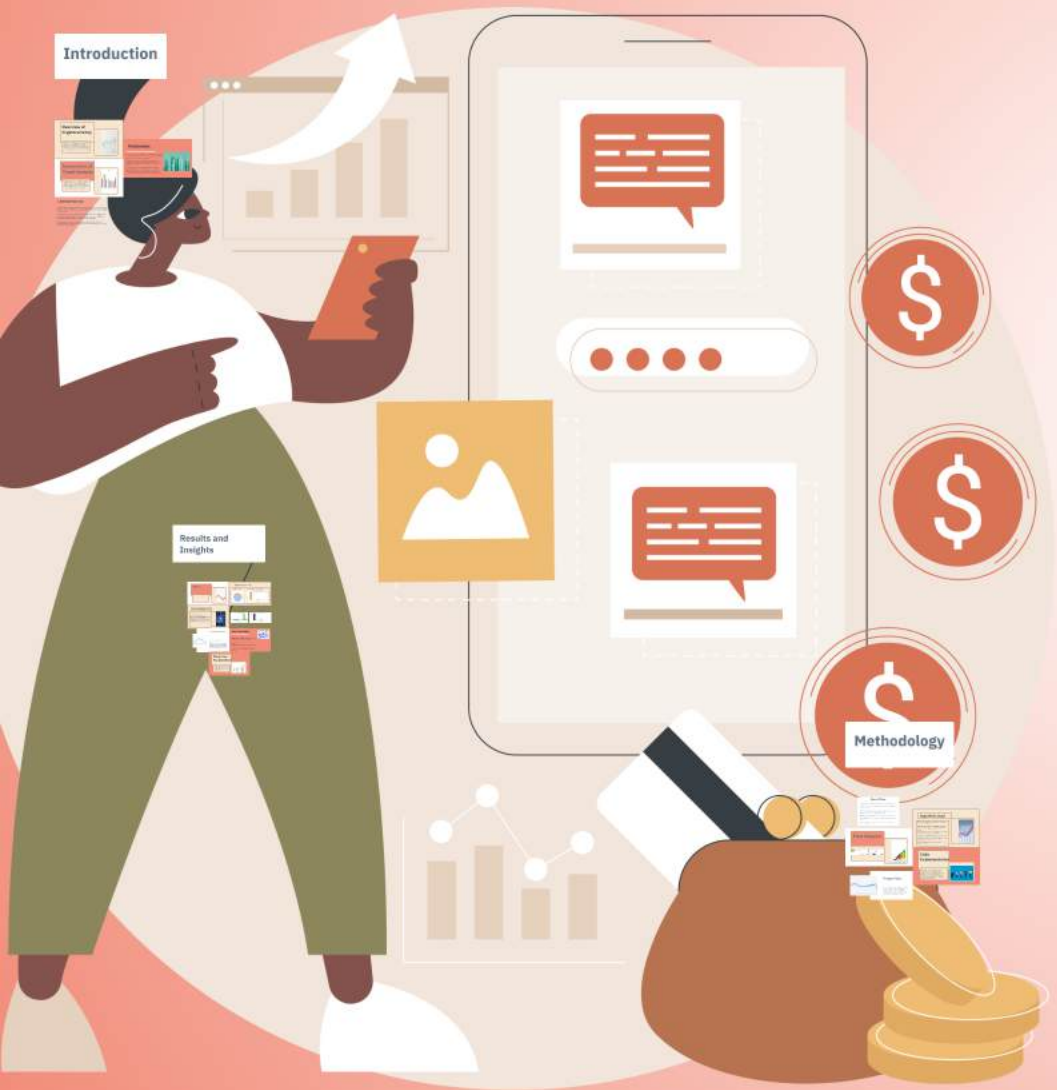
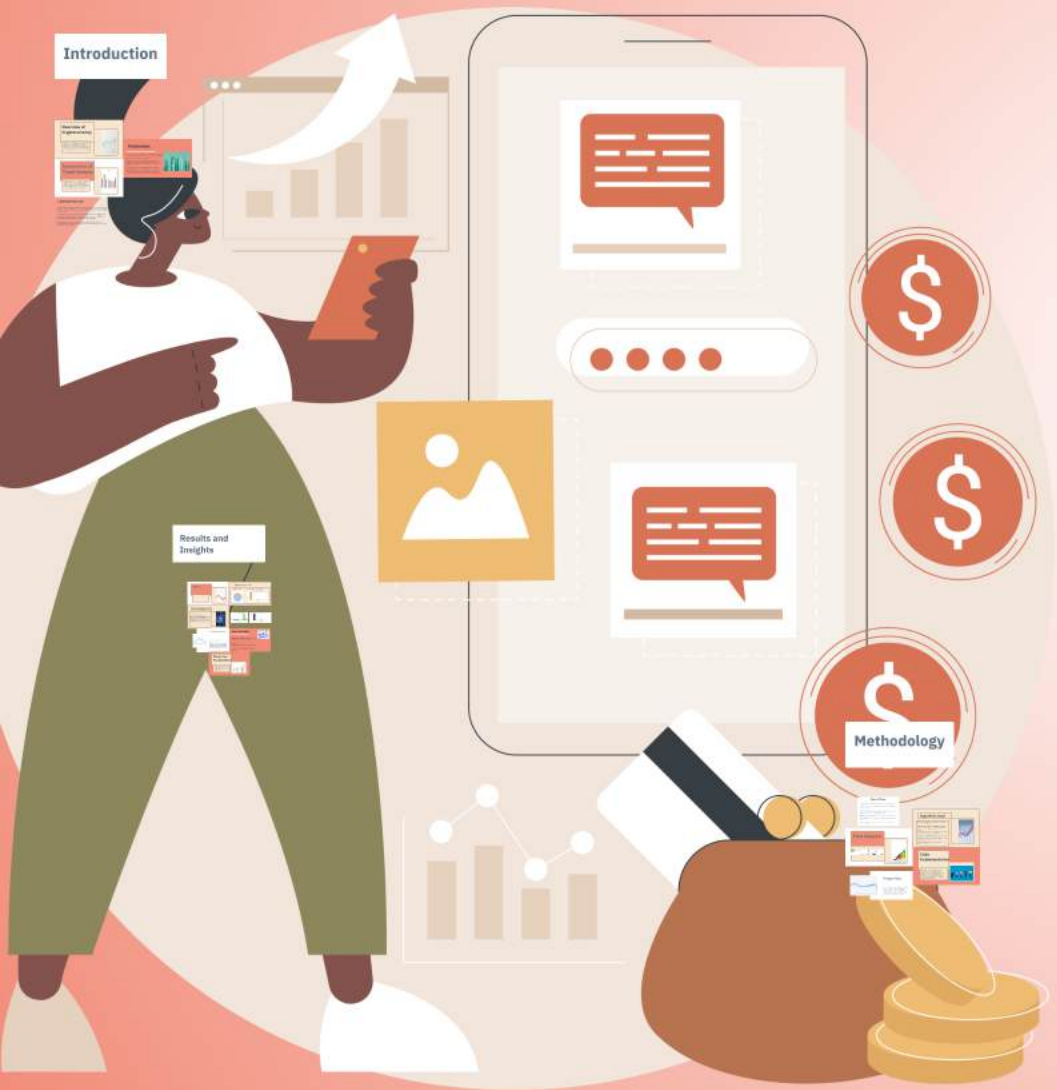


Real-Time Cryptocurrency Trend Analysis

By Group 4
017557462–Manjot Singh
017536623–DarpankumarJiyani
017514978– Saqib Chowdhury
017512781– Sai Prasad Thalluri



Real-Time Cryptocurrency Trend Analysis

By Group 4
017557462–Manjot Singh
017536623–DarpankumarJiyani
017514978– Saqib Chowdhury
017512781– Sai Prasad Thalluri

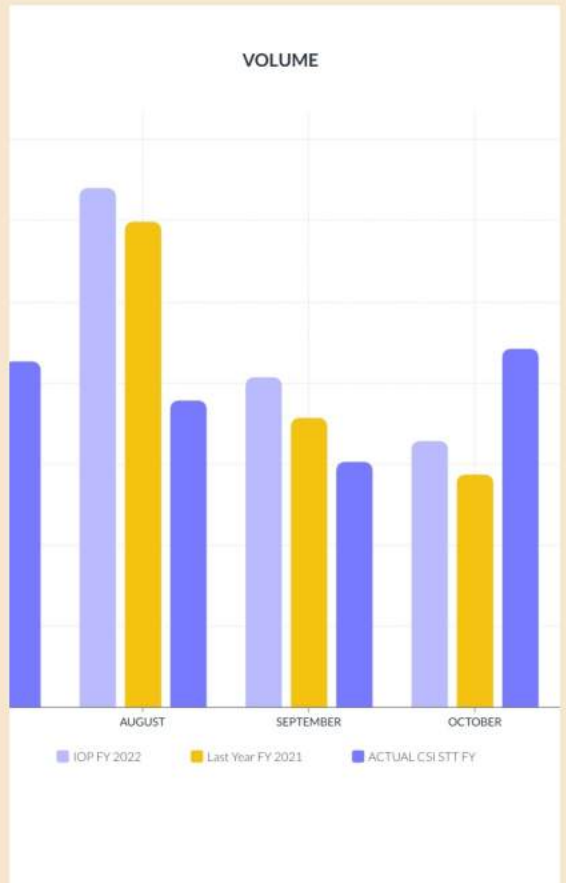
Introduction

Overview of Cryptocurrency

Cryptocurrency is a digital or virtual form of currency that relies on cryptography for security. Unlike traditional currencies, cryptocurrencies operate on decentralized networks based on blockchain technology, ensuring transparency and reducing fraud.

Importance of Trend Analysis

Trend analysis in cryptocurrency allows investors and analysts to identify patterns and predict future price movements. Understanding market trends is essential for making informed investment decisions and gaining competitive advantages.



Motivation

Issues such as regulatory uncertainty, security vulnerabilities, and the volatility remain significant barriers to their mainstream adoption.

In response to these challenges, there is demand for systems capable of analyzing vast volumes of real-time cryptocurrency data, while ensuring the privacy and security.

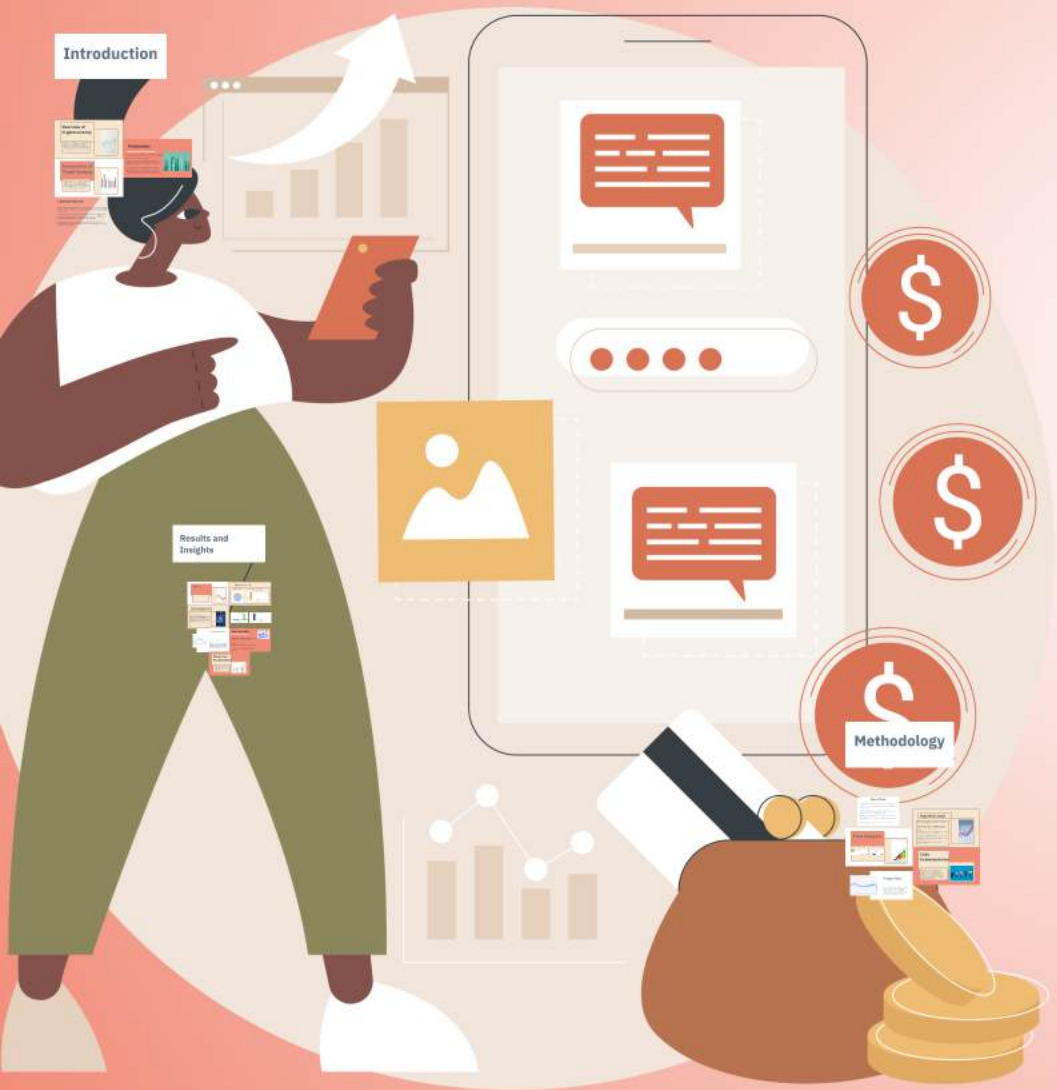
By leveraging big data technologies such as Hadoop, Spark, and Kafka, our project aims to provide real-time analysis of cryptocurrency trends, offering predictive insights and deeper understanding of market behaviors.

Literature Survey

Big Data Tools: Studies show that tools like Kafka and Hadoop/Spark are effective for scalable, real-time data processing in high-velocity environments.

Distributed Algorithms: Research highlights the use of MapReduce for large-scale aggregations, while Flajolet-Martin and DGIM are proven methods for approximations in data streams.

Cryptocurrency: Literature emphasizes the need for real-time streaming analytics to track trends and detect anomalies in volatile cryptocurrency markets.



Real-Time Cryptocurrency Trend Analysis

By Group 4
017557462–Manjot Singh
017536623–DarpankumarJiyani
017514978– Saqib Chowdhury
017512781– Sai Prasad Thalluri

Dataset

Data Collection

Source:
The dataset was collected in real-time from the CoinGecko API, which provides comprehensive cryptocurrency market data.

Coins Included:

Bitcoin, Ethereum, Dogecoin, USD Coin, Tether, Binance Coin, Lido Staked Ether, Solana, XRP, and Cardano.

Data Attributes

Basic Info: Cryptocurrency ID, name, symbol, and image URL.

Price Details: Current price, 24-hour price change, percentage change, and all-time high (ATH).

Market Data: Market capitalization, market cap rank, 24-hour volume, and circulating/total/max supply.

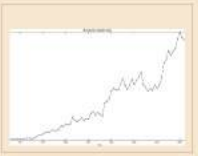
Historical Data: All-time low (ATL) with percentage change since ATL, and ATH with percentage change since ATH.

Real-Time Ingestion:

Data was streamed into Apache Kafka using a Kafka Producer, fetching data every 60 seconds.

Ensured continuous updates to reflect live market conditions.

Storage:
Streamed data was consumed by the Kafka Consumer and stored in HDFS on AWS EMR for distributed processing and analysis.



Data Collection

Source:
The dataset was collected in real-time from the CoinGecko API, which provides comprehensive cryptocurrency market data.

Coins Included:

Bitcoin, Ethereum, Dogecoin, USD Coin, Tether, Binance Coin, Lido Staked Ether, Solana, XRP, and Cardano.

Data Attributes

Basic Info: Cryptocurrency ID, name, symbol, and image URL.

Price Details: Current price, 24-hour price change, percentage change, and all-time high (ATH).

Market Data: Market capitalization, market cap rank, 24-hour volume, and circulating/total/max supply.

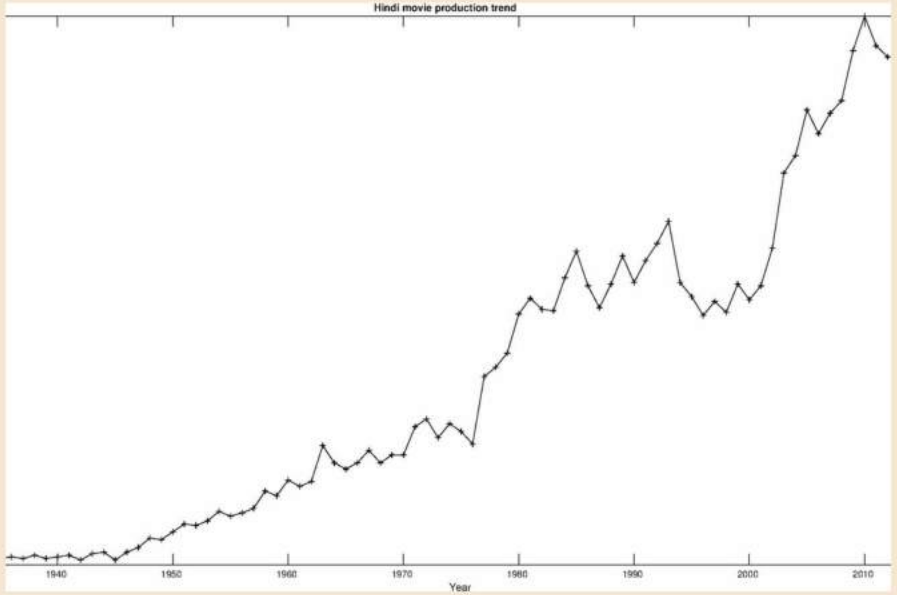
Historical Data: All-time low (ATL) with percentage change since ATL, and ATH with percentage change since ATH.

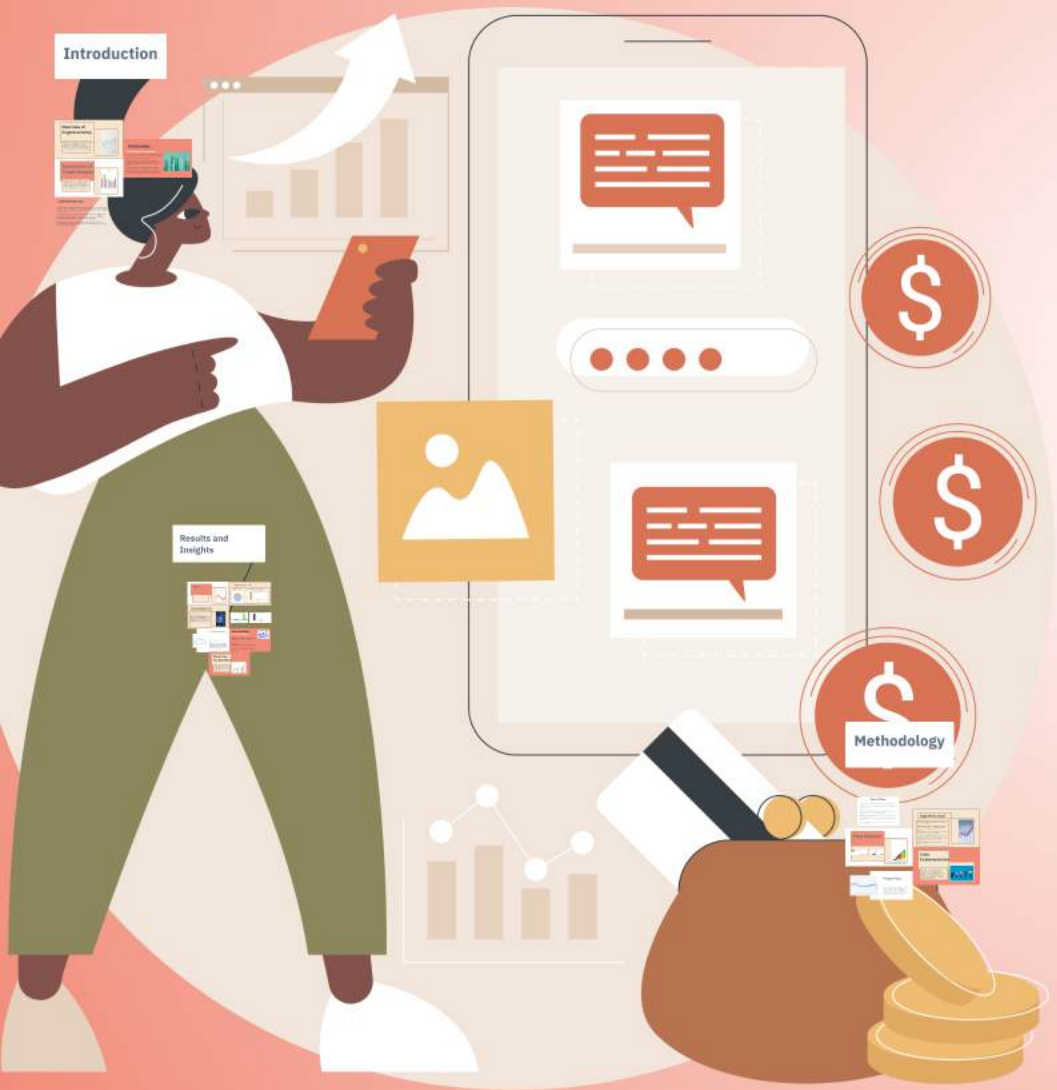
Real-Time Ingestion:

Data was streamed into Apache Kafka using a Kafka Producer, fetching data every 60 seconds.

Ensured continuous updates to reflect live market conditions.

Storage:
Streamed data was consumed by the Kafka Consumer and stored in HDFS on AWS EMR for distributed processing and analysis.

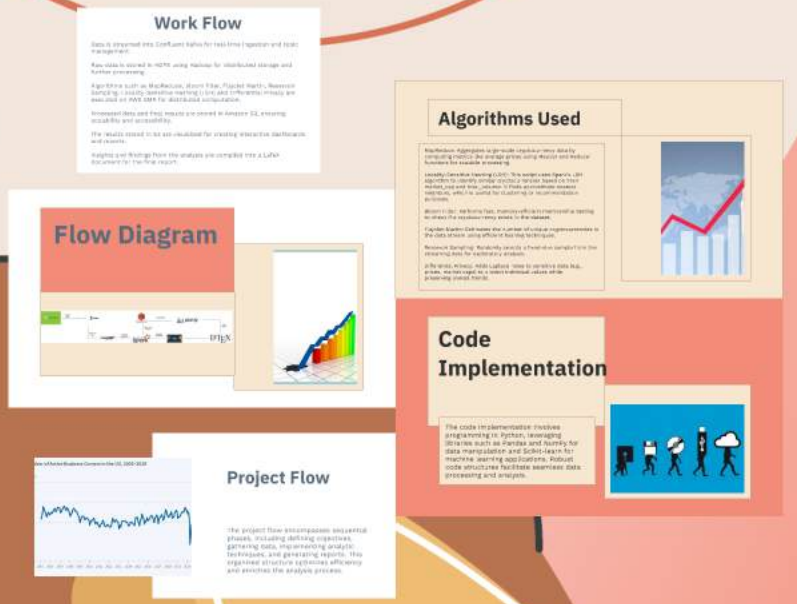




Real-Time Cryptocurrency Trend Analysis

By Group 4
017557462–Manjot Singh
017536623–DarpankumarJiyani
017514978– Saqib Chowdhury
017512781– Sai Prasad Thalluri

Methodology



Work Flow

Data is streamed into Confluent Kafka for real-time ingestion and topic management.

Raw data is stored in HDFS using Hadoop for distributed storage and further processing.

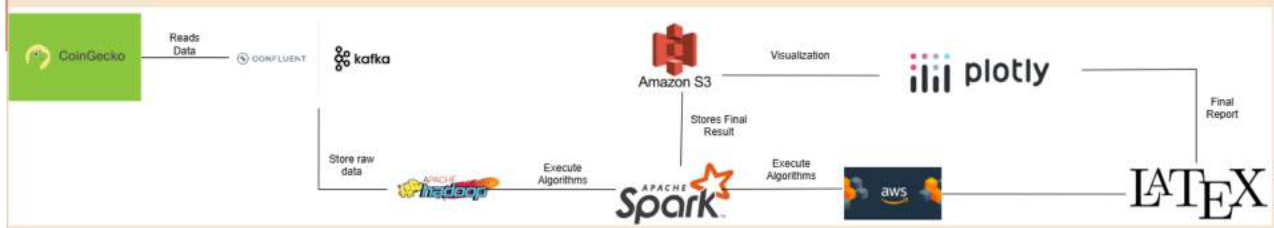
Algorithms such as MapReduce, bloom filter, Flajolet Martin, Reservoir Sampling, Locality-Sensitive Hashing (LSH) and DIfferential Privacy are executed on AWS EMR for distributed computation.

Processed data and final results are stored in Amazon S3, ensuring scalability and accessibility.

The results stored in S3 are visualized for creating interactive dashboards and reports.

Insights and findings from the analysis are compiled into a LaTeX document for the final report.

Flow Diagram



Algorithms Used

MapReduce: Aggregates large-scale cryptocurrency data by computing metrics like average prices using Mapper and Reducer functions for scalable processing.

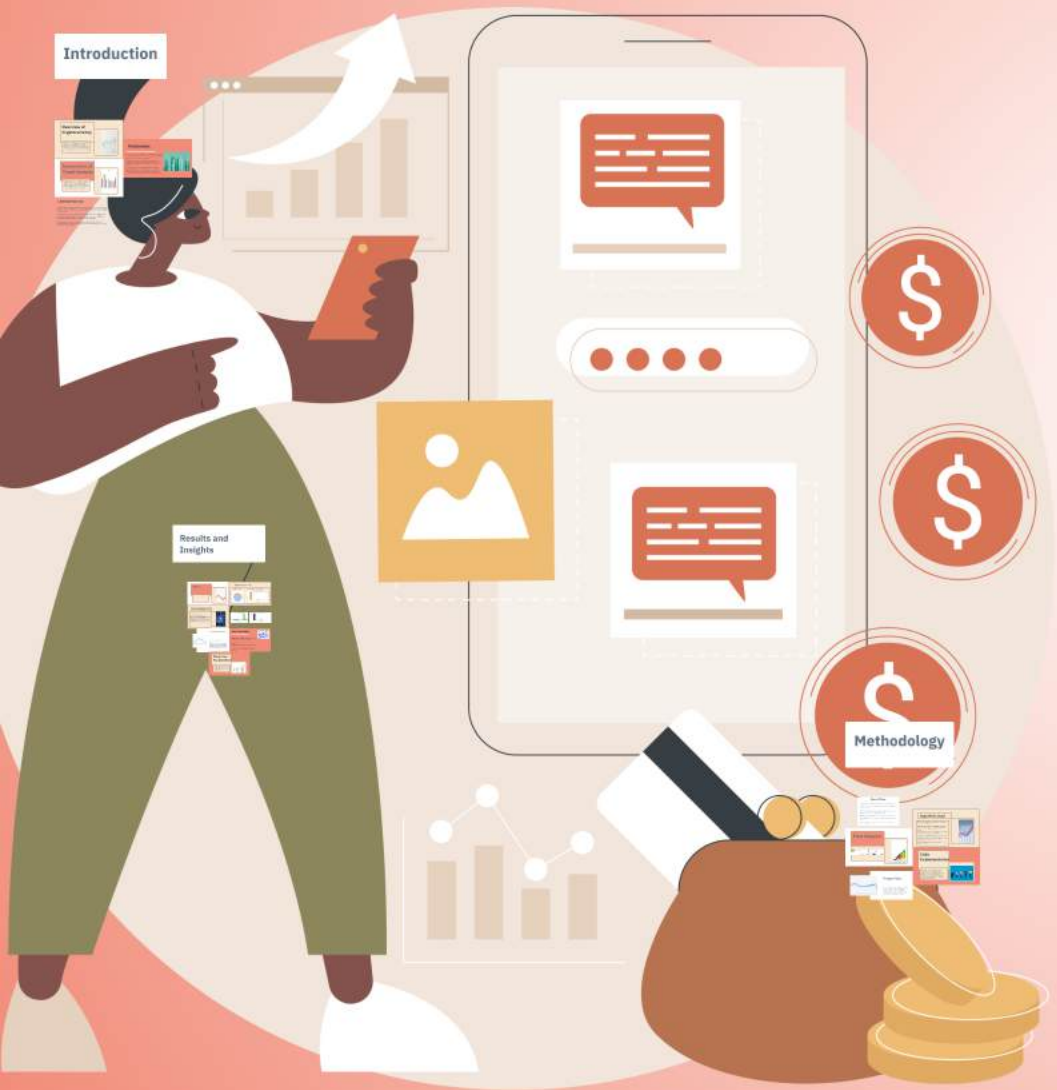
Locality-Sensitive Hashing (LSH): This script uses Spark's LSH algorithm to identify similar cryptocurrencies based on their market_cap and total_volume. It finds approximate nearest neighbors, which is useful for clustering or recommendation purposes.

Bloom Filter: Performs fast, memory-efficient membership testing to check if a cryptocurrency exists in the dataset.

Flajolet-Martin: Estimates the number of unique cryptocurrencies in the data stream using efficient hashing techniques.

Reservoir Sampling: Randomly selects a fixed-size sample from the streaming data for exploratory analysis.

Differential Privacy: Adds Laplace noise to sensitive data (e.g., prices, market caps) to protect individual values while preserving overall trends.



Real-Time Cryptocurrency Trend Analysis

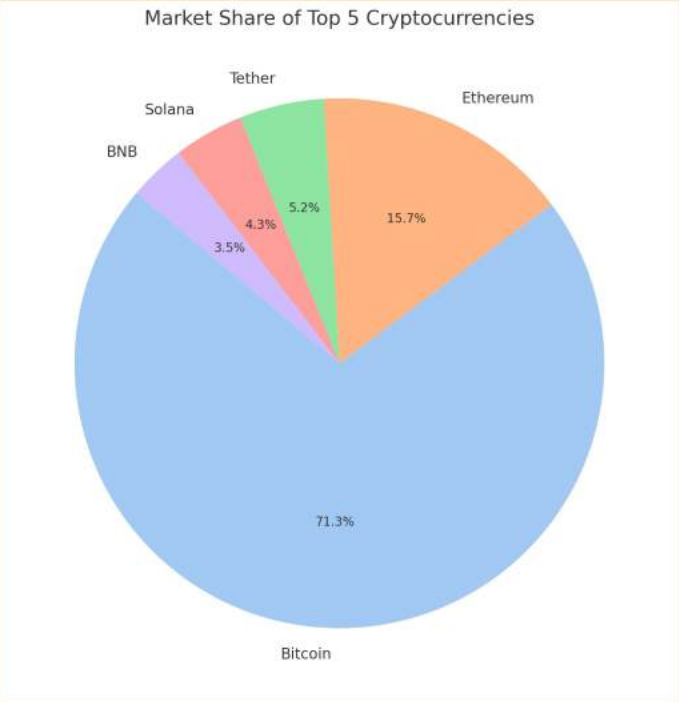
By Group 4
017557462–Manjot Singh
017536623–DarpankumarJiyani
017514978– Saqib Chowdhury
017512781– Sai Prasad Thalluri

Results and Insights

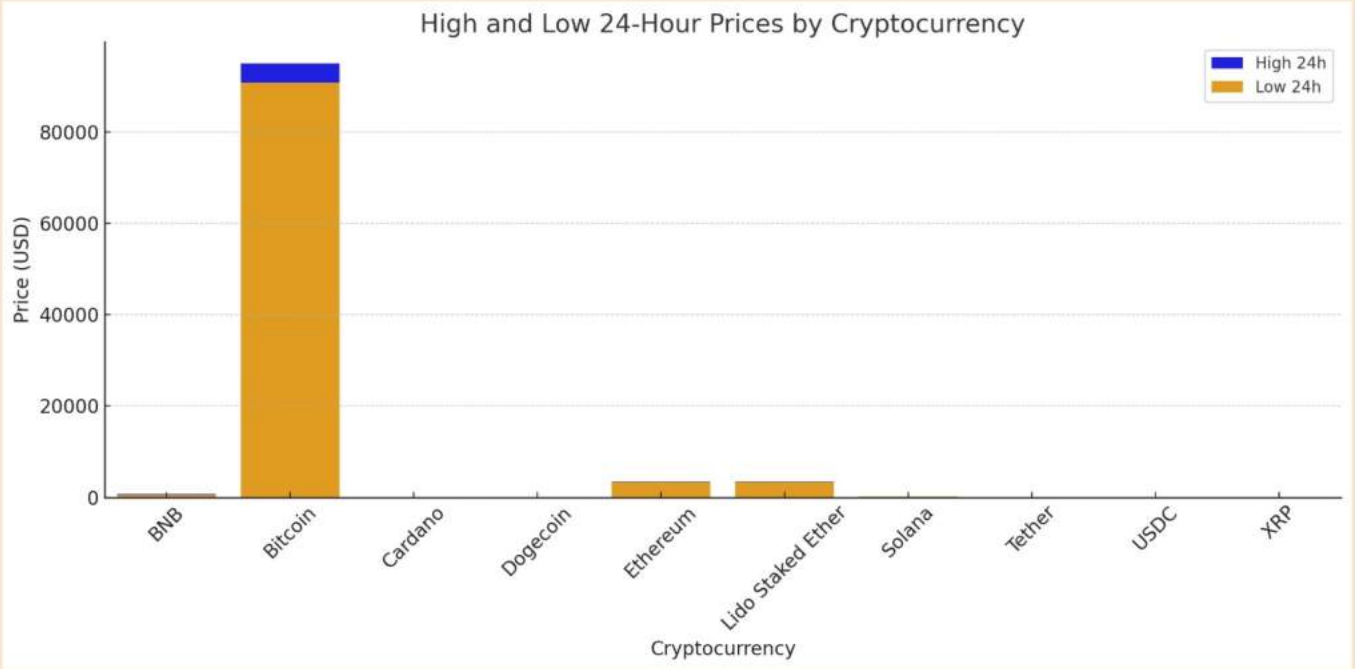
Demo

Insights and Visualizations

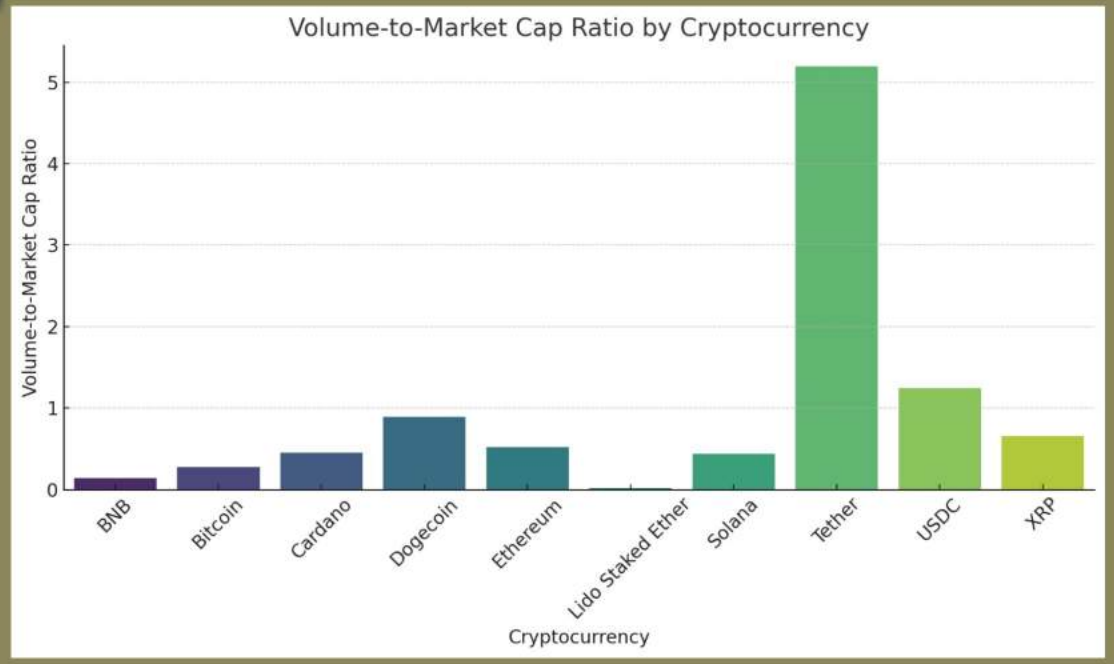
Market Share of Top 5 Cryptocurrencies



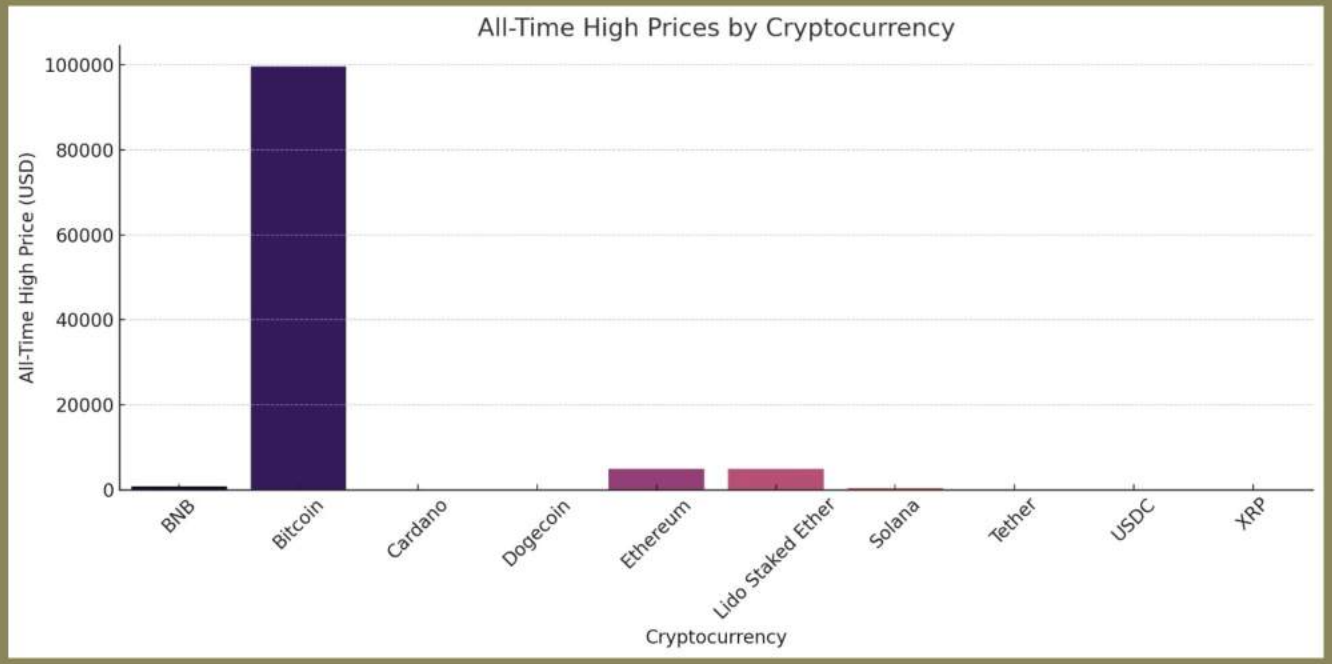
High and Low 24-Hour Prices by Cryptocurrency



Volume-to-Market Cap Ratio by Cryptocurrency



All-Time High Prices by Cryptocurrency



Technical Difficulties

Handling inconsistent CoinGecko API responses and ensuring efficient Kafka streaming posed challenges.

Setting up and optimizing Hadoop and Spark on AWS EMR for compatibility with Kafka and HDFS was complex.

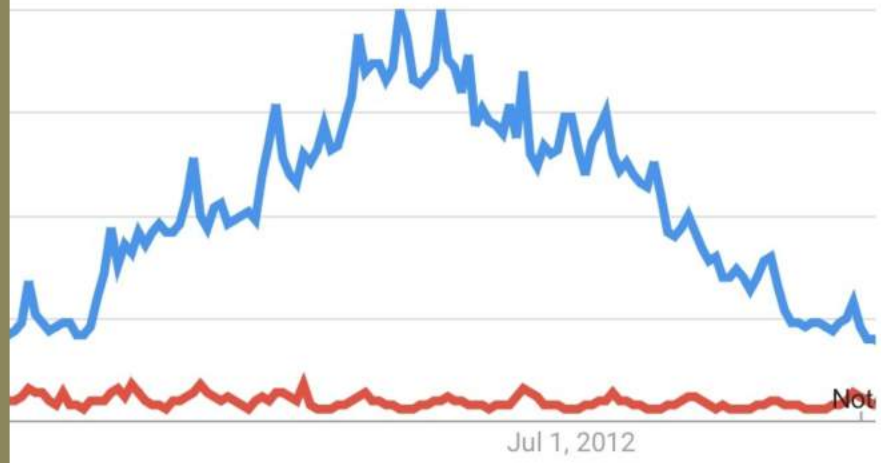
Ensuring accuracy and scalability for Flajolet-Martin, DGIM, and Reservoir Sampling in a distributed environment required optimization.

Managing real-time JSON data in HDFS, maintaining schema consistency, and optimizing storage for fast retrieval were challenging.

Balancing Differential Privacy's noise addition with data accuracy required careful tuning.

Optimizing resource allocation on AWS EMR and fine-tuning Spark and Kafka configurations to handle growing data volumes was demanding.

Implementing fault tolerance for network disruptions, Kafka failures, and data loss required robust configurations.



Conclusion and Impact

Conclusion

The project successfully demonstrated the development of a scalable, real-time big data pipeline for cryptocurrency market analysis. By integrating tools like Kafka, Hadoop, and Spark with advanced algorithms such as MapReduce, Flajolet-Martin, and Differential Privacy, the system effectively processed and secured real-time data streams. The pipeline provided actionable insights while ensuring data privacy and scalability, making it a robust solution for analyzing dynamic market trends.

Impact

This project showcases the potential of big data technologies to handle high-velocity, large-scale data streams in real-time. It provides a framework that can be applied across industries like finance, e-commerce, and IoT for trend detection, anomaly analysis, and real-time decision-making. By integrating privacy-preserving techniques and explainable AI, it emphasizes the importance of building transparent and trustworthy analytical systems.

Key Learnings

Gained hands-on experience with Kafka for real-time data streaming and managing high-velocity data flows.

Learned to configure and use Hadoop and Spark on AWS EMR for distributed storage and processing, ensuring scalability and performance.

Implemented and optimized advanced algorithms like MapReduce, Flajolet-Martin, and DGIM, understanding their application to streaming data.

Explored Differential Privacy to secure sensitive data while maintaining analytical accuracy.

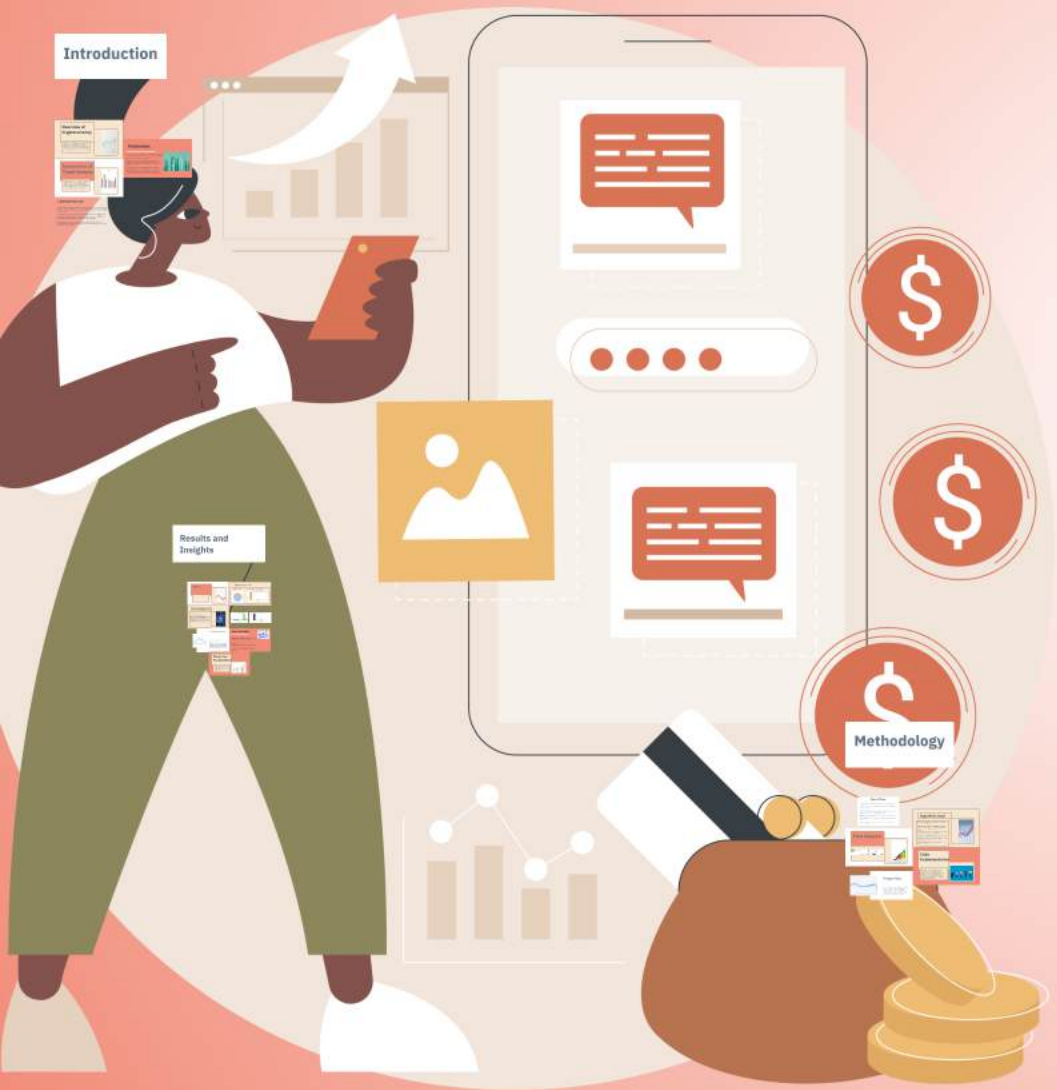
Integrated Explainable AI (SHAP) to enhance transparency and interpretability in insights.

Understood the challenges of managing real-time systems, including fault tolerance, scalability, and resource optimization in cloud environments.

Developed skills in visualizing results using Plotly and creating professional reports with LaTeX.

Thank You/ Any Questions

Thank you for your attention. This concludes our presentation. We welcome any questions regarding our project and findings.



Real-Time Cryptocurrency Trend Analysis

By Group 4
017557462–Manjot Singh
017536623–DarpankumarJiyani
017514978– Saqib Chowdhury
017512781– Sai Prasad Thalluri