TERM PAPER
EXECUTIVE SUMMARY

Name: Sarvesh Krishnan Rajendran
BUID: U86908171

Technology: Apache Kafka
Domain: Real time data streaming and Distributed event streaming

In an era where real-time insights are critical for decision-making, organizations are increasingly focused on technology that allow for real-time data intake, processing, and analysis. With the fast growth of connected devices, digital platforms, and transactional systems, there is an increasing need for a scalable and dependable solution capable of handling huge data streams and processing them in real time. Traditional batch processing systems sometimes struggle with the intricacies of real-time data streams because of their latency and inefficiency.

Apache Kafka is a distributed event streaming technology that manages real-time data streams at high throughput and low latency. Kafka's design, which separates data producers and consumers, enables scalable and fault-tolerant data processing. Unlike typical message brokers or database systems, Kafka organizes data into topics with partitioned parts, enabling for simultaneous data handling and seamless processing even as data volumes increase.

Kafka's capacity to handle high-velocity data streams makes it an excellent choice for a variety of applications, including log aggregation, real-time analytics, and event sourcing. It efficiently manages the storage and consumption of event data across different platforms, serving as a foundation for data pipelines in industries such as finance, telecommunications, and retail. This project used Apache Kafka to create real-time data streaming pipelines for applications like heart disease prediction and sentiment analysis. Kafka made it possible for PySpark to efficiently accept, process, and stream real-time data by saving pre-trained models and preprocessing procedures as reusable pipelines.

Furthermore, Kafka interfaces well with cloud storage options such as AWS S3 and analytics systems, and Google Cloud Platform enabling enterprises to expand their data pipelines. It supports major APIs and connections, making it extremely flexible to a variety of contexts and use cases. With advanced capabilities such as distributed messaging, real-time analytics, and stream processing, Kafka is a robust and versatile platform for handling large-scale data streams in real time.