DON BOSCO INSTITUTE OF TECHNOLOGY

Bengaluru, Karnataka – 74

TEAM COUNT: FOUR

PROJECT TITLE: REAL TIME SOCIAL MEDIA ANALYTICS PIPELINE : BUILDING A ROBUST DATA PROCESSING FRAMEWORK

TEAM LEAD NAME: RAHEESHA S

TEAM LEAD CAN ID: CAN_33695975

1) NAME: RAHEESHA S

   CAN ID : CAN_33695975

   ROLE : PROJECT MANAGER & RESEARCHER

2) NAME: SANJANA D

   CAN ID : CAN_33698964

   ROLE: MACHINE LEARNING ENGINEER

3) NAME: MAHESH N

   CAN ID : CAN_33760190

   ROLE: BACKEND DEVELOPER

4) NAME: SUNIL KUMAR T R

   CAN ID: CAN_33706158

   ROLE: FORTEND DEVELOPER

Real-Time Social Media Analytics Pipeline: Building a Robust Data Processing Framework

Phase 1: Problem Definition and Data Understanding 1.1 Project Overview

Real-time Social Media Analytics Pipeline The exponential growth of social media platforms has produced a massive amount of user-generated content. Organizations and researchers alike leverage this data to extract insights, monitor trends, and enhance decision making. A real-time social media analytics pipeline is designed to process, analyze, and visualize data at scale, delivering actionable intelligence almost instantly. Key components of the pipeline: Data ingestion:

Collect data streams from social media platforms such as Twitter, Facebook, or Instagram using APIs or scraping tools. Support structured and unstructured data formats such as text, images, and videos. Data preprocessing: Clean and transform raw data by removing noise, de-duplicating records, and handling missing values. Apply natural language processing (NLP) techniques for text normalization and sentiment analysis. Real-time stream processing: Use technologies such as Apache Kafka, Apache Flink, or Spark Streaming to process high-velocity data in real time. Implement filtering, enrichment, and aggregation of data streams. Data storage: Employ scalable storage solutions such as NoSQL databases (e.g., MongoDB, Cassandra) or distributed file systems (e.g., Hadoop HDFS). Store data in a format optimized for analysis and retrieval. Analytics and Visualization: Implement machine learning algorithms to identify trends, anomalies, and patterns in data. Provide dashboards using tools such as Tableau, Grafana, or custom web applications for real-time visualization. Scalability and Fault Tolerance: Ensure that the system can handle increasing data loads by incorporating scalable architecture. Integrate failover mechanisms to maintain uptime and reliability.

1.2 Objective of the Project

The objective is to create a system that can continuously ingest, process, and analyze social media data in real-time and to collect social media data from various plaforms and to analyze the incoming data immediately.

Potential Applications:

The application is to monitor and analyze the sentiments of publics and to identify the topics and to conduct the market research, to conduct market research, management of brand reputation and to detect crisis, to market real-time market decisions on based of live social media platforms.

1.3 Dataset Overview and Data Requirements

The dataset may consist of continuous streaming of social media posts from different platforms like Instagram, Twitter, Facebook, etc, and to capture details to post texts, and information of the user and to potential sentiment scores.

Dataset components:

User information

Post details

Timestamp

Engagement metrics

Sentiment score

Dataset requirements:

High volume

Data variety

Data quality

Potential datasource

Web scraping

Overview of the dataset

It typically consists of continuous streaming of social media posts from different platforms like Instagram, Twitter, Facebook, etc,, which includes information of users, timestamps, and metadata which allows for real-time analyses of sentiments, trends, and behaviour of the user as the data is generated.

Dataset name

User_id

Timestamp

Platform

Text Sentiment

Hashtag

Location

1.4 Conclusion of Phase 1

Building a robust real-time social media analytics pipeline enables organizations and researchers to derive valuable insights from dynamic social media data. The framework should prioritize scalability, low latency and high fault tolerance to address the unique challenges of real-time analytics. By leveraging advanced technologies and methodologies, organizations can monitor public sentiment, predict emerging trends, and improve decision-making processes to stay ahead of the competitive landscape