# Integrated Big Data Pipeline for Social Media Sentiment Analysis using Spark, MongoDB, and PostgreSQL

## Introduction

This project implements a mini big data system to collect, process, and analyze real-time Twitter data. The aim is to extract public sentiment towards specific topics and present the results in a structured and visual format.

## System Overview

Tweets are collected in real-time using the Tweepy Python library and stored in MongoDB as raw JSON documents. Apache Spark processes the data, performs cleaning, and applies sentiment analysis using a natural language processing library. The processed results are stored in PostgreSQL for structured querying.

## Architecture

1. Data Collection: Tweepy API → MongoDB

2. Processing & Analysis: Spark DataFrame transformations + Sentiment model

3. Storage: PostgreSQL relational tables

## Results

The system successfully classifies tweets into Positive, Negative, and Neutral categories. SQL queries provide insights such as sentiment distribution over time and most active users.

## Conclusion

The pipeline demonstrates how big data tools can work together for real-time sentiment analysis. Future improvements include integrating a live dashboard and expanding to multiple social media platforms.