Real-time Social Media Analytics Tool for Twitter

ITI41 - Data Management

***The objectives in this document present the baseline of the project. Sky is the limit. Act as a data engineer who is gathering requirement and building a robust solution to ensure satisfying customer needs. You are free to choose additional tools, Integrate secondary data sources …etc.*

Introduction

The project aims at building a data platform for real time moderation and analytics of twitter data. The implementation will utilize different big data technologies as Spark, Kafka and Hive, in addition to visualization tools for data discovery and delivering insights.

Objective

1- Use python to stream data from twitter as a source to Kafka with the right configuration
2- Use Kafka queuing system for twitter messages
3- Use python to create the destination topics with the right configuration
4- Use Spark structured streaming to fetch and process data from Kafka.
5- Analyse the tweets sentiment using a simple approach
6- reply with the right message based on the tweet's sentiment.
7- Output data from spark streaming to HDFS in the form of parquet files
8- Build a Hive data model (Staging tables + Analytical tables) that use modelling best practices for big data
9- Use PowerBI to create a dashboard that fetch data from Hadoop. **The trainee is expected to act as an analyst who is willing to deliver some actionable information to business users.**
10- **Bonus:** Compare between SparkSQL connector and Conventional ODBC connectors (for Hive) in terms of Performance.
11- Create a monitoring script in python that utilizes Ambari API to check the important services.

Deliveries

1- Each trainee is expected to deliver a link to a Github repo including all the project code/config files. You have to add a README file in Markdown format explaining general info about the project(technologies used, image of your models , samples from data across the different steps of the flow … etc.)

2- Each trainee will present her/his work in a 20-minute powerpoint presentation and a live demo for their work.

Technologies to be used

1- For streaming data from twitter use
   a. Flume OR Spark
2- For queuing and persisting messages
   a. Kafka
3- For processing data from Kafka to Hive
   a. Spark Streaming
4- For SQL on Hadoop
   a. Hive
5- For visualization
   a. PowerBI/Tableau
   b. Bonus: Apache Superset

Required Architecture