Project

Team 14

Title: Evaluation of framework performing Sentiment Analysis on Twitter Stream Data

Description: Our project aims to perform sentiment analysis depending on the chosen topic (which will be an input to the system) and find out the major talked about issues in relation to that. We will use Twitter's real-time streaming data and perform the visualization on the results.

Team Members:
Li Shi – lshi7
Navjot Singh – nsingh9
Shriyansh Yadav – scyadav

Deliverables

1. Apache Kafka - to buffer the incoming Twitter Stream data
2. An architectural framework based on Hadoop ecosystem
3. Components of the ecosystem consisting of HDFS, YARN, and Apache Spark on Amazon's EC2 instances
4. A NoSQL database (preferably MongoDB) to store the tweets as a key value pair
5. An application hosted on AWS that will show the analysis in the form of charts/graphs
    a. Dynamic to choose the kind of chart/graph based on user
    b. Ability to show the analysis based on a particular sentiment

Status

Done:

- Successfully running 4 EC2 nodes on AWS (under two accounts)
- Completed the setup of HDFS with 1 master node and 2 data nodes
- Completed the setup of MongoDB in 1 node
- The hadoop ecosystem on AWS has also been setup with YARN and Zookeeper

In process:

- Working on the setup of MongoDB connectors for Hadoop on all the nodes
- Working on the setup between Twitter API and Kafka

Issues

- The IPs of nodes on AWS are dynamic so rebooting leads to a disruption of the hadoop ecosystem which then needs a reconfiguration
- Dealing with limit on the number of Free Tier EC2 instances
    - Planning to add additional available VCLs, if need be

- Dealing with slow streaming of Twitter Data:
    - Since, there is a limit on the number of requests that can be made to the Twitter Stream API we plan to use downloaded datasets
- Finding proper techniques to overcome shortcomings in analyzing a Tweet
    - Complete expression of thought is limited because of limit on the length of message
    - Presence of grammatical/spelling errors, use of emoticons, and colloquialism in the Tweets