Sentiment Analysis

Introduction:

Sentiment analysis is the analysis of people's opinions, sentiments, evaluations, appraisals, attitudes and emotions in relation to entities like individuals, products, events, services, organizations and topics by classifying the expressions as negative / positive opinions.

Scope of work:

In this project you will be trying to do sentiment analysis of the opinions on twitter using Hadoop Eco-System. The reason we chose twitter is because the social media is gaining popularity for the customer reviews and it is also creating a good business-customer relationship in the market. The reviews would certainly reflect the service of the company in the market.

Objectives:

The objective of the project is to find the sentiments of the users on Twitter. Companies can analyze the sentiments of the users regarding their products / services and use it for betterment of the same.

Plan of work:

- Collect Data: The initial step is to collect all twitter tweets.
- Pre-process Data: You need to write MR job to pre-process the data
- Classification of Data: Write Hive UDF for the classification of the data into positive / negative opinions
- Print Data: Final stage would be to print the desired results with the number of good and bad tweets collected.

Architecture Diagram:

User tweets about his opinion or feedback on twitter

TWITTER

Twitter stores tweet in json format

Hadoop Pre-Processes incoming raw data (JSON) and extract relevant information from it.
Pre-processing is done with Map-Reduce job.

APACHE FLUME

Flume uses Twitter4j API to connect to Twitter Streaming API (which gives constant stream of tweets).
Flume writes Incoming tweets into HDFS

Server   Server   Server
Server   Server   Server

HADOOP CLUSTER

HIVE

Use hive udf for classifications and queries for reporting

REPORT

Report generated from HIVE Output