Interim Project Report

**a) No-SQL tool: JAQL**

Based on the project requirement and our research on various tools. We finally decide to use JAQL as a No-SQL to store and retrieve data to and from HDFS. Following are the reasons behind the decision.

1. Here we primarily work on twitter data which is in JSON format. Jaql is primarily a query language for JSON, but it supports more than just JSON.
2. Jaql is much like a blend of Pig and Hive.
3. IBM uses Jaql as primary data processing language for BigInsights.

Apart from the above mentioned reasons there are several other features like support for different tyoes of data(structured, semi-structured, no structured), ability to perform parallel processing of queries by MapReduce method and ideal to use for highly nested data. These encouraged us to use Jaql as a No-SQL tool for our project. Which looks like a perfect fit for our project requirement.

**b) Software Modules to develop:**

This project consists of following modules:

1. User Interface: This frontend allows users to pose queries and see results in the form of visualiztions. User interface is a webpage which forwards the request along with the proper parameters to Query generator/processor. After the processing is done UI visualizes the results from query processor.
2. Query generator/processor: After user poses a querie, the request is forwarded to this module. This acts as a controller for generation of queries based on the input from the user and replies back the results.
3. Visualizer: This module is the creator of nice visualizations for the requested queries from user.
4. StreamTweets: Primary goal of this modules is to live stream twitter's data and store it to local disk.
5. StoreTweets: This module helps in pushing the streamed data on to HDFS using Jaql script.

**c)Thoughts on analytic queries:**

Following are the analytic queries we are implementing

1. Trend of the hashtag over a period of time. This tells us the importance of hashtag. E.g 100 people discuss a topic in a day. Next day 1000 people involved in the discussion. Then the following day 500 likewise and so on.
2. To retrieve location(more specifically latitude and longitudes) of all the tweets for a certain hashtag. This gives us a worldwide view of the discussion happening at particular time.
3. Based on the language of the tweets retrieve no of tweets for all the languages.
4. Retrieve the hot topics of the day and visualize in a word cloud.
5. For each unique Tweet find out aggregation of number of retweets. Each tweet is given a color and find the aggreation for top 10 retweeted users.

**d)Division of work:**

We worked as a team through out the project till now. As it is required we segregated the development work among ourselves in the following way.

* Rahul Ponnada: came up with 2 queires 1&2
* Satish Chowdary:came up with 2 queires 3&4
* Gori Mehaboob: Nothing

Decision and division of modules is done as a team.