**Interim Report**

**A) NO-SQL tool - HIVE:**

Apache Hive is a [data warehouse](http://en.wikipedia.org/wiki/Data_warehouse) infrastructure built on top of [Hadoop](http://en.wikipedia.org/wiki/Hadoop) for providing data summarization, query, and analysis. While initially developed by [Facebook](http://en.wikipedia.org/wiki/Facebook), Apache Hive is now used and developed by other companies such as [Netflix](http://en.wikipedia.org/wiki/Netflix). Amazon maintains a software fork of Apache Hive that is included in [Amazon Elastic MapReduce](http://en.wikipedia.org/wiki/Apache_Hadoop#Amazon_Elastic_MapReduce) on [Amazon Web Services](http://en.wikipedia.org/wiki/Amazon_Web_Services). (Source: **Wikipedia**)

Hive is a data warehousing package infrastructure built on top of Hadoop. It provides an SQL like language, called Hive Query Language (HiveQL) for querying data stored in a Hadoop cluster. HQL is the Hive query language like all SQL dialects in widespread use. It is perhaps closest to MySQL’s dialect, but with significant differences. Hive offers no support for row level inserts, updates, and deletes. Hive adds extensions to provide better performance in the context of Hadoop and to integrate with custom extensions and even external programs. It is well suited for batch processing data like: Log processing, Text mining, Document indexing, Customer-facing business intelligence, Predictive modeling, hypothesis testing etc. It provides indexing, different storage types, Built in user defined functions (UDFs). Hive can be used within applications written in Java, PHP, Python and Ruby.

**B) Modules:**



Fig 1.0

Data Collection:

We collect the tweets using twitter API and Java code. We are collecting the tweets in JSON file.

Data Processing:

In this module, tweets collected in the JSON file will be loaded into HDFS. Required data will be extracted from HDFS using HQL (Hive Query Language) queries.

Data Visualization:

Graphical representation of the Extracted data will be done using IBM Bluemix.

**C) Queries:**

We will collect tweets based on hashtag fast and furies movie. Then we will process and analyze the data to project below results by using queries.

* Top countries talking about Fast and Furious.
* Most engaged actor in the movie Fast and Furious.
* Most retweeted tweet about movie and about actor.
* Gender and age group that are interested in Fast and Furious

We are currently working on the queries and may add additional queries if possible.

**D) Work Distribution:**

Data Collection: Rakesh Vistarakula and Keerthi Yanda

Data Processing and Analysis: Sandeep Ballu and Rakesh Vistarakula

Data Visualization: Keerthi Yanda and Sandeep Ballu