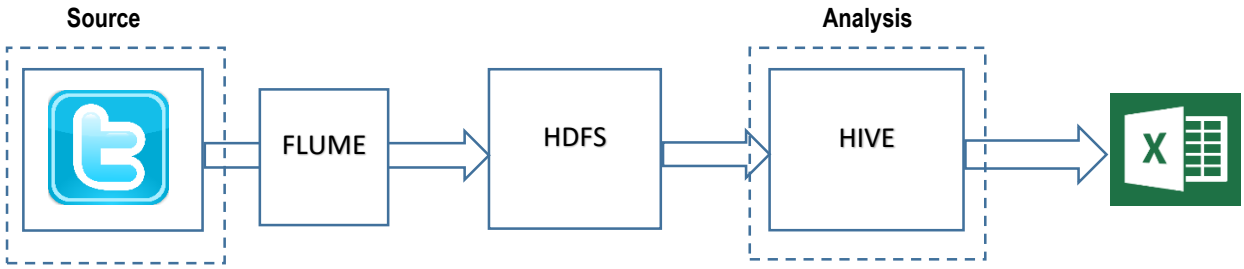
DISTRIBUTED COMPUTING PROJECT PROPOSAL (BATCH F2S)

1. **Title:** Opinion mining for gauging UB's online presence and popularity using keyword processing.
2. **Team:** Hope to leverage this project experience to dive deeper in to the field of social network analytics. We would like to learn Hadoop and develop a thought process to help us apply this knowledge to other real world problems. The team comprises of,
Pranay Rao: I will be particularly focusing on 'Obtaining twitter feed and creating a data dictionary' & 'Data Ingestion into HDFS using FLUME' & 'Metadata and table creation using Hcatalog'
Sourav Mukherjee: I will be mainly working on 'Data analysis using Hive' & 'Data visualization'
3. **Motivation:** In today's world social media plays an important role is building and maintaining a brand. Negative buzz generated online can have a huge impact. Similarly, positive buzz largely affects the way people perceive an organization/product. The goal of this project is to prepare UB for a PR crisis and provide enough information for it to monitor and analyze its online presence. Opinions are free of bias as Twitter users are not a part of a study. They mostly communicate freely online. Therefore, the information obtained would help make a bigger impact.
4. **Problem Statement:** To understand how UB performs in terms of popularity in the social media, namely twitter and also how people feel about our university.

- How many times does UB get mentioned on Twitter?
- In what context does UB get mentioned on Twitter?
- Can results obtained be correlated to real life events?

5. **General Approach:**
• Data sources: Twitter API
• Architecture:



- Relevant tweets would be obtained using a Twitter API.
- We would use a sentiment dictionary that would label a list of words in the English language as positive, neutral or negative.
- Twitter data that would be in JSON format would be ingested into HDFS using Apache Flume.
- Using Hcatalog (sub-component of Hive) we would create a relational structure for the data (metadata and table management).
- Apache Hive would then be used to query and analyze the data.
- Excel Powerview would be used for data visualization in the form of bar graphs, histograms, pie charts etc.

• Implementation Plan:
Implementation would be done on a single node installation of Hadoop. We considering the options of using a third party Hadoop distribution (like Cloudera) or installing the individual Hadoop projects as per our needs. The following packages would be used as a part of our implementation:

• Apache FlumexApache	• Hive/ Hcatalog	• Excel Powerview
• SQL	• Java	

• Other Resources:

1. http://www.cloudera.com/content/cloudera/en/resources/library/whitepaper/Ten_Common_Hadoop-able_Problems_Real-World_Hadoop_Use_Cases_White_Paper.html
2. http://www.ijcsit.com/docs/Volume%205/vol5issue03/ijcsit2014050393.pdf
3. https://cwiki.apache.org/confluence/display/Hive/StatisticsAndDataMining
4. http://blog.datumbox.com/10-tips-for-sentiment-analysis-projects/

6. **Tentative Schedule:** Obtaining twitter feed and creating a data dictionary – 5th October
Data Ingestion into HDFS using FLUME – 21th October
Metadata and table creation using Hcatalog – 30th October
Data analysis using Hive – 20th November
Data visualization – 30th November