**Problem Description:**

We need to build an Interactive dashboard, which can help the Business users to better understand the perceived presence of IonIdea/Client business in the marketplace.

The dashboard will help to capture and curate content & conversations based on social insights analysis that can become digital assets based on external messaging.

**Proposed Approach:**

**Data Retrieval:** Social networking sites and blogs like Twitter, Facebook, YouTube etc. can be targeted as a data sources to download data related t specific industries using the targeted keywords.

{MR program to filter the data/exclude words}

**Data cleaning/Analysis:** The data can be analyzed for their sentiment to get the information about the view of users in marketplace based on the geographical locations. Also these data can be divided in terms of major industries and subclasses.

Retail (Class)

a. E-Commerce

b. RetailCustomerCentricity

c. MobileRetailing

**Visualization:** The data can be visualized using various tools to display important information retried based on classes and sub classes. Timeline plots, bar chars and heatmaps can be used for this purpose.

**Technical Specifications:**

**Data Retrieval**:

Flume – can be configured to retrieve data from various sites and JSON objects can be stores in flat files in HDFS.

Hive – can be used to parse the JSON objects and store the important information in tabular format in HDFS. {JSON UDF from Hive}

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sr | TweetID | PrimaryTweetID | Text | Latitude | Longitude | TweetTime | UserName | UserId | NumberOfRetweets | NegativeStrength | PositiveStrength | MajorCalss | MinorClass |

**Data Analysis:**

Java programs – can be used to process information for sentiment analysis (Sentistrength java classes)

Pig – can be used to filter and analysis of the data, like quick merge of the data based on specific information like locations, states etc.

**Visualization:**

Tableau – can be used for visualization purposes. (Pentaho)

**Design:**

**Twitter Source:**

1. Get the data from the twitter for Retail, Healthcare and Telecom Keywords using Flume into HDFS.
2. Create a table in Hive with one column Tweets to load the raw data.

create table poc.raw\_twts (tweets string);

1. Use a Text Loader in Pig to load the Flume data with HCatStorer option into the table created in step 2 as below:
   1. a1 = LOAD 'hkotha/tweets/2015/11/18/21/FlumeData.1447902880291' using TextLoader() as (tweets:chararray);
   2. STORE a1 into 'pocdb.raw\_twts' using org.apache.hive.hcatalog.pig.HCatStorer();
2. Using Json\_tuple function in Hive to extract the required columns from the raw\_twts table and store in another table fmt\_twts as below:

create table pocdb.fmt\_twts as select t1.id,t1.text,t1.created\_at,t1.retweet\_count,t2.\* from pocdb.raw\_twts a

lateral view json\_tuple(a.tweets,'id','text','created\_at','retweet\_count','user') t1 as id,text,created\_at,retweet\_count,user\_info

lateral view json\_tuple(t1.user\_info,'name','location','time\_zone') t2 as name,location,time\_zone;

1. UDF developed in Java to classify the data as below:
   1. Build each file for the 3 industries with keywords and read those files from Distributed Cache in UDF into 3 different maps.
   2. Check the tweet text for any matches with the keywords from each of the file and set the respective indicator if found.
   3. Return the three indicators concatenated value from the UDF Ex: YYN/NYY/NYN..