**Project #2:** Analyze Twitter Data with Apache Hadoop

Industry: Social media

Data: Publicly available twitter data set filtered on few keywords

Problem Statement : Which Twitter users get the most retweets? Who is influential within our industry? We need to determine whose tweets are getting heard by the widest audience, and also determine whether these people are communicating on a regular basis or not. We can use this information to more carefully target our messaging in order to get them talking about our products, which, in turn, will get other people talking about our products.

Social media has gained immense popularity with marketing teams, and Twitter is an effective tool for a company to get people excited about its products. Twitter makes it easy to engage users and communicate directly with them, and in turn, users can provide word-of-mouth marketing for companies by discussing the products. Given limited resources, and knowing we may not be able to talk to everyone we want to target directly, marketing departments can be more efficient by being selective about whom we reach out to.

In this project we will use [Apache Flume](http://flume.apache.org/), [Apache HDFS](http://hadoop.apache.org/hdfs/), [Apache Oozie](http://incubator.apache.org/oozie/), and [Apache Hive](http://hive.apache.org/) to design an end-to-end data pipeline that will enable us to analyze Twitter data. We will see how each component is involved and how the custom code operates.

