DeZyre

How-to: Analyze Twitter Data with Apache Hadoop

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.

You have to use Apache Flume, Apache HDFS, Apache Oozie, and Apache Hive to design an end-to-end data pipeline that will enable us to analyze Twitter data.

Who is Influential?

To understand whom we should target, let's take a step back and try to understand the mechanics of Twitter. A user – let's call him Joe – follows a set of people, and has a set of followers. When Joe sends an update out, that update is seen by all of his followers. Joe can also retweet other users' updates. A retweet is a repost of an update, much like you might forward an email. If Joe sees a tweet from Sue, and retweets it, all of Joe's followers see Sue's tweet, even if they don't follow Sue. Through retweets, messages can get passed much further than just the followers of the person who sent the original tweet. Knowing that, we can try to engage users whose updates tend to generate lots of retweets. Since Twitter tracks retweet counts for all tweets, we can find the users we're looking for by analyzing Twitter data.

Now we know the question we want to ask: Which Twitter users get the most retweets? Who is influential within our industry?