Analyzing Twitter data with 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.

We'll learn how we can use Apache HDFS and Apache Hive to design an end-to-end data pipeline that will enable us to analyze Twitter data.

Process:

First of all we will acquire data that we want to process. The data we have here is twitter feeds data in JSON format. It has different types of information about feeds like text of tweet, sender of tweet, if tweet has been resent? Etc.

The data is very much in structured form so we can store it in hive and use hive QL queries to process it.

We will first create hive tables which have same schema as JSON data of twitter. We will use JSON SerDe to read data from JSON file and show it as Hive table data.

Then we will run our Hive QL queries to process that data.

1. We will try to identify what topics are most discussed on twitter?
2. Who is famous celebrity by identifying which user gets most number of tweets?