Online Social Network Data Analytics using the Hadoop/MapReduce Framework

By: Karan Kaushik
Abhisht Arkalgud Sathyanarayana
Outline

- Goal: Mining and obtaining data from the web which can be qualitatively analyzed via the quantitative analysis of the generated outputs.
- Our Dataset: Tweets from a user's Twitter account.
- Distributed Computing: MapReduce operation is being performed on HDFS, the Hadoop Distributed File System.
Challenges

- Getting data from Twitter user accounts
- Loading data onto the Hadoop Distributed File System (HDFS)
- MapReduce Operation on Hadoop
- Drawing the Graphs or Plotting on Maps
- Statistical Analysis and Interpretation of generated output
Accessing and Fetching Twitter Data

OAuth:
'Open Authentication'- a 3 legged authorization scheme without use of user-name or password

Constraints and Issues:
API rate limits and Fail Whale

Redis:
Capability to perform operations on large sets of data to generate key-value pairs.

Infochimps:
Data catalog which archives historical Twitter Data

JSON Object:
Representation of the fetched Twitter data
Hadoop

- Can store large volumes of tweets in the HDFS (Hadoop Distributed File System) and perform Map/Reduce analytics on the data set.
- Easy to scale and distribute workload across nodes.
- We now construct a Map and Reduce operation to parse out the entities from tweets and count their frequency.
- Entities are - keywords, urls, user mentions
- Map operation emits <keyword><1>, reduce operation aggregates across these key value pairs to return <keyword><count>
Future Work

Map of USA
Related Work
Applications and Utility

- Faster processing of available data
- Understanding current trends
- Forecasting of future trends
- Pattern Recognition
- Audience Targeting
Thank you