Claims

Team12:

Team Members:

Shriya Sharma	ssharm25
Sonal Patil	sspatil4
Sai Sameer Tirumalasetti	stiruma

The project aims at capturing and analyzing streaming data from Twitter to compare the different categories that come up in tweets. The various categories we want to take into consideration are:

- Political
- Social
- Entertainment
- Environmental
- Science and technology
- Finance
- Sports

Infrastructure required:

Separation of data collection, data processing and its analysis. This infers that the volume of data received from the data producers should not affect the processing speed of data or its analysis. Amazon Kinesis would be a good fit for the requirements as it separates data collection and processing. It provides flexibility according to the changing data needs as the Twitter feed is a stream of data and changes according to intervals. Streamed data will be loaded using Firehose and stored on S3 prior to further processing.

Claims:

1. Implement an infrastructure pipeline to handle large data volumes
2. Perform analysis on the captured data independent of the volume of data being received from producer (i.e. Twitter Stream).
3. Giving out near real-time results for the analysis performed over the data that updates the results of the analysis with the changing input data.
4. Categorize the data stream being received from producers based on the predefined classifiers in the application.
5. Generate graphical visualisation to compare the different categories of the data captured.