Team12

- Title: **Twitter Stream Categorizer**
- Description: 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 & technology,finance and sports

- Team members:

Shriya Sharma	ssharm25
Sonal Patil	sspatil4
Sai Sameer Tirumalasetti	stiruma

Deliverables

1. Implement an end to end infrastructure pipeline that takes in twitter data stream into Kinesis, processes it and stores it into S3.
2. Perform analysis on the captured data independent of the volume of data being received from Twitter using Kinesis Analytics.
3. Giving out near real-time results for the analysis performed updating results of analysis with the changing input data by monitoring the traffic.
4. Define classifiers that categorize the data obtained from stream in application and build models based on them.
5. Generate graphical visualisation in the model to compare the different categories of the data captured.

Status

1. Implemented sample Kinesis stream with 4 shards that is fed with the twitter data of New York area and a consumer that consumes the data.
2. Created an S3 bucket that holds the tweet data.
3. Implemented Firehose delivery stream that stores the data stream onto S3
4. Implemented helper functions to update shards of Kinesis Stream, describe stream, list existing streams, create roles

Issues

1. Handling processing of multiple shards efficiently
2. Appropriate partitioning of data
3. Determine the optimal number of shards for the kinesis stream.
4. Create role policies dynamically through api.
5. Distribute shards efficiently through regions when the volume of tweets is scaled up to be received from a larger area.