CSC:591 - Data Intensive Computing:
Project Overview

GROUP NUMBER 11:

TweetSmart
(Smart Twitter Notification System)

- TweetSmart is a data-intensive computing system for dynamically determining the top 5 interests of each user and notifying them about the most recent updates regarding those interests.
- A data intensive solution is implemented to query this distributed database containing historical and streaming data to calculate those interest values with a few parameters.
- This framework will help to analyze topics of interest over time, locations, age of users, etc. and enable smarter notifications and targeting of advertisements.

Team members:

Saurabh Shanbhag, sshanb2
Sayali Godbole, ssgodbol
Aishwarya Sundararajan, asundar2

DELIVERABLES

1. **Design:**
   - Designing the components of system
   - Determination of the amount of data to be processed in-memory and stored in persistent storage.
   - Determine number of nodes for distributed data storage and retrieval.
2. **Getting and Storing Twitter Data:**
   - Mining streaming data from Twitter and storing it in the Amazon Dynamo database.
3. **Using Hadoop, ZooKeeper for Computation:**
   - Distributing the data efficiently and computing top interests (Mapreduce) for each user.
4. **Linking Database with the Druid API:**
   - Querying the Distributed database for dynamically checking user interests for notifications.
5. **Displaying results/Notification:**
   - Notifying user on new interesting post (watcher).

DEPENDENCIES

- **Dataset:** Twitter API json response
- **Storage:** Amazon DynamoDB.
- **Distributed Computing:** Apache Hadoop
- **Coordination:** Apache Zookeeper
- **Server Hosting**: AWS

ANTICIPATED ISSUES

- Twitter Data access (Continuous API requests)
- Fault Tolerance
- Load Balancing
- Accuracy (We can never be sure about the result - Not much room for validation tests)