TweetSmart
(Smart Twitter Notification System)

Aishwarya Sundararajan, Saurabh Shanbhag, Sayali Godbole

DEFINITION:

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. We will be building a data intensive solution which will store the historical as well as streaming data and we will query this distributed database to calculate those interest values with a few parameters. TweetSmart will notify the users only about the updates of interest for them. The data-store will contain terabytes of historical data along with high velocity streaming data. This framework will help to analyze topics of interest over time, locations, age of users, etc. and enable smarter targeting of advertisements.

JUSTIFICATION:

- **Mining Streaming Data:**
  This project would be an excellent learning opportunity to learn the process of mining high velocity data, storing, and processing the same to retrieve interests and querying notifications of the Twitter pages of interest.
- **Distributed Computing:**
  This project would also enable us to use Amazon DynamoDB/Apache Cassandra, a NoSQL database for storage of data. We would use Hadoop/Zookeeper to implement distributed storage service for the huge Twitter DataSet. This could help us enhance our knowledge about all these technologies.
- **Druid Architecture:**
  Since a large database is being queried for real-time notifications, we take advantage of the druid architecture for deep storage and persistence of historical data. The architecture also uses Zookeeper for balancing and coordination amongst nodes that are indexed to store data.
- **AWS:**
  We would get the opportunity to use AWS platform and all its capabilities in the project for deploying our application. It will be a great learning curve.

OVERVIEW:

The goal of this project is to build a data intensive system that delivers efficient storage of user profiles and activity, determines top interests for each profile and notifies that user for only those interesting topics. This will ensure that the user will not get notifications which are not really important to him. TweetSmart will analyze huge quantity of data (Twitter profiles, likes, shares, retweets, etc) which will keep changing dynamically and ensure scalability, reliability and efficiency of the system.

The various tasks involved are:

- Designing the components of system, determination of the amount of data to be processed in-memory and stored in persistent storage. This step also helps determine the number of nodes that need to be installed for distributed data storage and retrieval.
- Mining streaming data from Twitter and storing the data in the Amazon Dynamo/Cassandra database; Using Hadoop/Zookeeper for distributed computing.
- Linking the Database with the Druid API.
- Querying the Distributed database for dynamically computing top interests for users.
- Displaying results to the user and notifying user on new post (watcher).

VALIDATION:

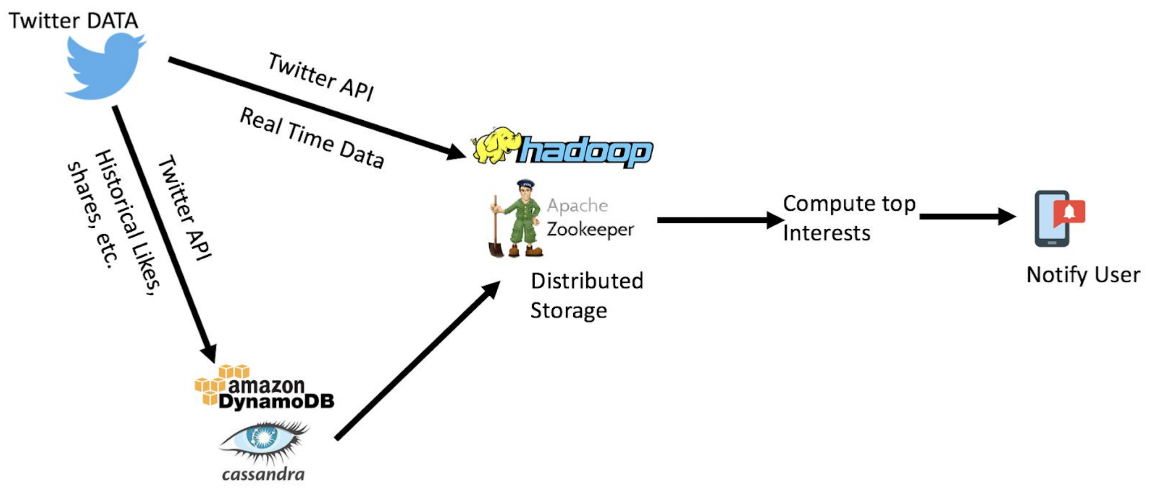
We will check the application against historic and real-time data to check whether the interests are correct for any given user. We will test the notifications based on our interests.

ARCHITECTURE:

We are planning to have architecture similar to Druid paper.**[1]** We'll have a persistent storage database and a in-memory cache (Redis) to answer queries on recent data. The datastore will have a streaming input from Twitter API. In the first run we'll have some users and calculate their top five interests on historical data and this will be done one time. After that, once the application is live we'll keep updating the interests of users as per feed from Twitter. In the background, whenever a new tweet arrives, we'll find all users with relevant interest and push the notification to them.

Components:

1. Twitter API - Receiving Data
2. NoSQL DB (DynamoDB/Cassandra) - Storing Data
3. Hadoop/ZooKeeper - Distributing Data
4. Notification System - Notifying users



TIMELINE:

TASK	**TIMELINE**
Design Discussion	09/30/2018
Mining streaming Twitter Data, Persist historical Data	10/15/2018
Linking DynamoDB with Druid	10/30/2018
Create Framework for determining interests, Querying the DB for notifications	11/15/2018
Testing the prototype, Scale up	11/30/2018

BACKGROUND:

Twitter has 336 Million monthly active users as per statista.com**[2]**. These users spend most of their time tweeting about daily incidents. Most celebrities and political figures are followed by several hundred thousands of people. Determining the political interests of people without the media bias has become a much needed use case today**[3]**. These followers do not want to miss out on notifications from these popular profiles/pages and this is exactly the problem that TweetSmart attempts to address.

RESOURCES:

Data: Twitter API
Database: Amazon DynamoDB
Scripting Language: Python 3.0
Real-Time Data Analysis: Druid

The amount of data stored in memory can be quantified to GBs and the amount of deep storage data can be quantified to terabytes.

FUTURE SCOPE:

The future scope of this project would involve mapping the Twitter profile of users to their corresponding LinkedIn profiles by building a Lambda stream between the output of interest from Twitter API and the LinkedIn API. Adequate filters could be applied to achieve this mapping with a reasonably high accuracy of 70%. This would aid in displaying the professional interests of users along with their personal notifications from Twitter and serve as an excellent framework for targeting users with right notifications in the long run. Furthermore, the accuracy could be improved to ~90% by applying advanced machine learning and data mining techniques.

REFERENCES:

[1] Fangjin Yang, Eric Tschetter , Xavier Léauté, Nelson Ray, Gian Merlino, Deep Ganguli.Druid A Real-time Analytical Data Store

[2]https://www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-united-states/

[3] Jennifer Golbeck & Derek Hansen. A method for computing political preference among Twitter followers. Social Networks Volume 36, January 2014, Pages 177-184