ABSTRACT

Ashish Deole : 201250880
Juhi Kulshreshtha : 201205534
Meenal Goyal : 201101145
Pankhuri Goyal : 201101174

Title : *Twitter with Cassandra*

Problem Definition: To build an online microblogging website (like Twitter) over Cassandra.

Problem Description/Approach

Current issue related to social networking sites is their inability to handle large scale data since they have been developed over RDBMS. Our task is to build an online social networking and microblogging platform over Cassandra which is a hybrid non-relational database. It provides scalability and high availability without compromising performance.
Cassandra is an ideal runtime database for web-scale domains like social networks. It trades off strong consistency in favour of high partition tolerance and availability but provides eventual consistency(CAP Theorem). It is chosen among other nosql databases because its write performance is excellent.

Why cassandra?
With MySql, it becomes very difficult to build a high performance, write intensive application on a data set that is growing quickly, with no end in sight. Whereas, Cassendra which is NoSQL distributed database management system, efficiently uses Map/reduce with Apache Hadoop, provides fast and reliable service and has the ability to handle read and write requests at the same time.(writes are much faster than reads).

Tools/ Framework used

- **Cassandra :** Apache Cassandra is an open source NoSQL distributed database management system designed to handle large amounts of data across many commodity servers, providing scalability and high availability with no single point of failure.
- **Pycassa :** It is an open source Python client library for Apache Cassandra having following features:
    - Auto-failover for normal or thread-local connections

- Batch interface
- Connection pooling
- Method to map an existing class to a Cassandra column family

Modules

- **Front end User Interface :**
  Provide a GUI for the users to interact with the application.
- **Data Model Layer :**
  To design and develop an efficient data model to organize and store large quantities of structured and unstructured data.
- **Abstraction Layer over Cassandra :**
  Build a wrapper over the Cassandra API to exploit its functionalities as per the requirement of the project.
- **Rest API :**
  We would be using REST calls as an interface between the front-end and the back-end.

Features

- **Adding a new user**:
  This will allow a user to sign up for an account.
- **Tweeting**:
  A user could post about any topic which could be viewed by other users.
- **Following a user**:
  A user could choose to follow tweets of a particular user.
- **Retweeting**:
  A user could repost another user's tweet.
- **Marking favorite tweets**:
  A user would have an option to mark some tweets as his favorite.
- **Hashtags**:
  A user could insert the tags by putting '#' before important words.This will provide a means of grouping messages, allowing a user to search for the hashtag and get the set of messages that contain it.
- **Current Trends**:
  A user would be able to view tweets about immediate popular topics based on who the user follows and his location.
- **Search**:
  A user could search for other users/ friends using keyword searching.
- **View profiles/tweets**:
  A user would be able to view profiles, tweets and activities of other users.

Plan of Action

Phase I :

- ➔ Understanding Cassandra
- ➔ Design of Data Model

Phase II :

- ➔ Develop a basic end-to-end complete application with minimal functionalities like user registration, following/unfollowing a user, tweeting etc.
- ➔ Functioning application over REST calls

Phase III :

- ➔ Provide a front-end interface
- ➔ Add more functionalities like hashtags, current trending topics, etc.

References

https://github.com/AboutUs/Cassandra-SF-2011-Notes/blob/master/cassandra-at-twitter.md

https://github.com/AboutUs/Cassandra-SF-2011-Notes

http://www.rackspace.com/blog/cassandra-by-example/

https://github.com/twissandra/twissandra

http://wiki.apache.org/cassandra/ArticlesAndPresentations

http://planetcassandra.org/blog/post/using-cassandra-for-real-time-analytics-part-1

http://widwebway.com/en/blog/?p=22

http://www.informationweek.com/software/enterprise-applications/twitter-drops-mysql-for-cassandra/223100894

http://www.computerworld.com/s/article/9161078/Twitter_growth_prompts_switch_from_MySQL_to_NoSQL_database

http://highscalability.com/blog/2011/12/19/how-twitter-stores-250-million-tweets-a-day-using-mysql.html