CS644 ASSIGNMENT 1

Report of Big Data Application on Social Media

Group members:
LIN TANG (31400221)
YI WU (31392055)
JIAYU ZHANG (31395853)
JINZHEN WANG (31374073)

Lecturer: Prof Chase Q. WU

Table of Contents

APPLICATION DOMAIN 2

FOUR V'S OF FACEBOOK 3

BIG DATA-RELATED PROBLEMS 4

EXISTING AND PROPOSED SOLUTIONS 5

REFERENCES 6

Application Domain

Big Data of Social Media

A report from McKinsey & Co. stated that by 2009, companies with more than 1,000 employees already had more than 200 terabytes of data of their customers' lives stored.

Now, adding this startling amount of stored data to the rapid growth of data that has seen in social media over the last four years, there are trillions of tweets, billions of Facebook likes, and an even higher number of check-ins on Foursquare. Instagram and Pinterest are only adding to this social media data deluge. Picture the buckets of data that has been gathered by social media sites alone.

Social media guarantees the acceleration of innovation, the drive of cost savings, and the strengthening of brands through mass collaboration. Across every industry, companies are using these platforms to market and hype up their services and products, along with monitoring what the audience are saying about their brand.

The convergence of social media and big data gives birth to a whole new level of technology. [1]

Big Data of Facebook

As a social network that has been popular for the last five years with over 1.2 billion users worldwide, Facebook stores a gigantic amount of user data, making itself a massive data wonderland.

The Social Media Marketing Industry Report 2015 states that Facebook is the #1 social platform for marketers.

Every day, we feed Facebook's data beast with mounds of information. 10 billion Facebook messages, 4.5 billion hits on the 'like' button, 350 million new picture uploads, all on a daily basis.

Based on this mounds of information of Facebook, we choose Facebook as the delegate of social media of our report. [2]

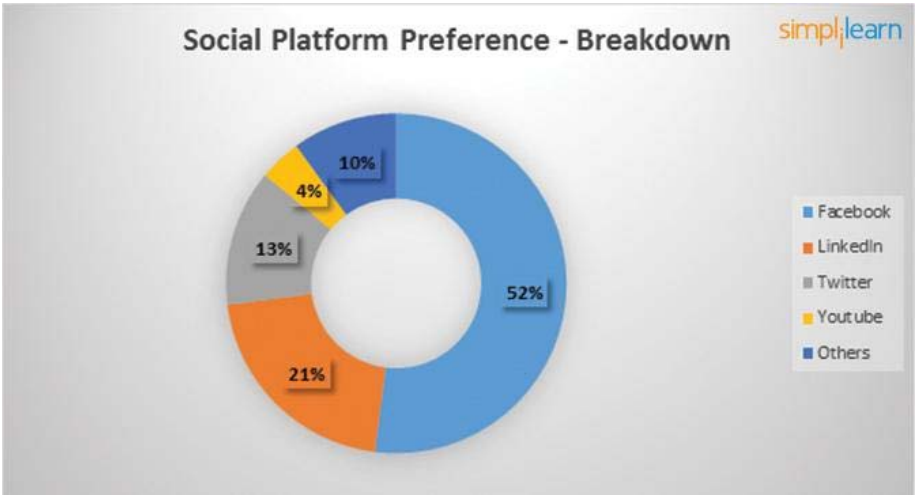

Figure1. Social Platform Preference Breakdown

Four V's of Facebook

- **VOLUME**

  Simply, the Volume stands for the quantity of generated and stored data. The size of the data determines the value and potential insight and whether it can actually be considered as big data. [3]

Figure2. the Big Data of Facebook

Facebook revealed some big stats to a few reporters at its HQ, including that its system processes 2.5 billion pieces of information and 500+ terabytes of data each day. It's pulling in 2.7 billion Like actions and 300 million photos per day and it scans roughly 105 terabytes of data every half hour. Plus, it is also providing the first details on its new "Project Prism". [4]

- **VARIETY**

  The type and nature of the data. This helps people who analyze it use the resulting insight effectively.

Figure3. Variety of Facebook Data

As we mentioned in Volume part, the data type of Facebook contains 'Like' action, graph. It also has video, text, search, cookies, audio and so on.

- **VELOCITY**

  In this part, the speed at which the data is generated and processed at Facebook is presented and discussed. As is known that the data generating and data processing speed has to meet the demands and challenges that lie in the path of growth and development.
  Below we present several examples to show the Velocity of Facebook.

  Example 1: 'Celebrate Pride'
  Following the Supreme Court's judgment on same sex marriage as a Constitutional right, Facebook turned in a 'rainbow drenched spectacle', called 'Celebrate Pride'. Within the first few hours, more than a million users had changed their profile pictures.

  Example 2: "I Voted"
  Facebook successfully tied political activity to user engagement when they came out with a social experiment by creating a sticker allowing its users to declare "I Voted" on their profiles. This experiment was run during the 2010 midterm elections and seemed effective. Out of a total of 61 million users, 20% of the users who saw their friends voting also clicked on the sticker.

- **VERACITY**

  The quality of captured data can vary greatly, affecting accuracy analysis.

Figure4. Fake Profile on Facebook

In Facebook, there are fake profiles. The fake information will affect precision analysis. For example: Facebook Bot. A Facebook bot is an automated software program that is designed to create and control a fake Facebook account. A Facebook bot is completely automated program that generates a profile by scraping images and information from other sources. After setting up a fake profile, it spreads by friending other Facebook users. [5]

Big Data-related Problems

Problem 1: (Analytics) Facebook Bot

As we mentioned above, Facebook Bot will create some fake profiles, social bots present an interesting security challenge given the trust factor in social networking as well as the emotional factor of users' motivation to have as many "friends" as possible. This kind of fake information will lead to an unprecise analysis result and information security concern.

Problem 2: (Database/Storage) How does Facebook manage such a huge amount of data?
Every day there are 2.5B content items shared and 2.7B Likes. We care less about GiGo[6] (Garbage-in Garbage-out) content itself, but metadata, connections, relations are kept transactionally in a relational database. The above 2 use-cases generate 5.2B transactions on the database, which means it takes over 60000 write transactions per second averagely only to process the transaction operations generated by Sharing and Like.

Problem 3: (Transfer) Reading, not storage, is an issue [8]
One of the issues Facebook has to face is how to deal with the time series data. In a time-series database, usual queries are a comparison of data points across different time intervals. “Is the number of users served today more or less than a year ago?”
Producing a lot of data is easy and producing a lot of derived data is even easier. Solution? Compress all the data. But how do you answer the queries then? Scan through the data. Problem is here. Reading data may need a long querying time.

Existing and Proposed Solutions

Problem 1
Existing Solution: Facebook has strict systems in place to prevent social bots.
Proposed Solution: New user must provide their ID numbers when they sign up Facebook. Such as, SSN, driver ID.

Problem 2

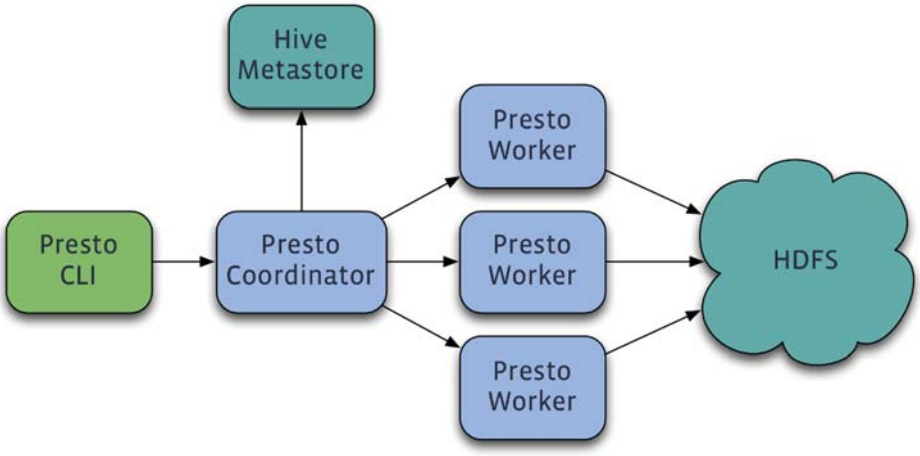

Figure5. Presto [9]

Existing Solution: Facebook offered some insight into how it handles the more than 300 petabytes of data it stores for its 1.19 billion monthly active users, providing some details on

Presto, an interactive query system it created and is open-sourcing, in a note on the Facebook Engineering page. Presto was designed to help Facebook process queries for data with lower latency — in other words, quicker from an end-user standpoint — and a "small team" in the social network's data infrastructure group (Martin Traverso, Dain Sundstrom, David Phillips, Eric Hwang, Nileema Shingte, and Ravi Murthy) launched the project in the fall of 2012. The interactive query system is now open-sourced, and developers can access the code and other information via the Presto site or GitHub. [7]

Proposed Solution: The problem might be solved by creating a monster RAM layer, kept in memcached, with minimum latency where constant updates from a single user to many others can be performed efficiently and in a timely manner.

Problem 3

Existing Solution: To reduce the querying time, the goal should be to minimize IO time by reducing the number of records read each time to answer a query.

The solutions are a special kind of time-series databases based on open-source technologies and a smart data model to overcome said deficiencies. We suggest Parquet as the file format. Apache Parquet is a columnar storage format available to any project within the Hadoop ecosystem, regardless of the choice of data processing framework, data modeling or language. Columnar storage has several advantages. Firstly, organizing data by columns allows better compression, as the data is homogenous. Secondly, IO is considerably reduced because we can effectively scan only a subset of the columns. Thirdly, as data of the same type are stored in each column, it allows effective encoding techniques. Lastly, Spark is known to work better with Parquet.

Proposed Solution: The problem can be solved by creating a cache to store the latest record of profiles of each user.

References

[1] Avantika Monnappa: https://www.simplilearn.com/how-facebook-is-using-big-data-article
[2] MICHAELA.STELZNER: http://www.socialmediaexaminer.com/SocialMediaMarketingIndustryReport2015.pdf
[3] Wikipedia: https://en.wikipedia.org/wiki/Big_data#Characteristics
[4] Josh Constine: https://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/
[5] https://www.techopedia.com/definition/27812/facebook-bot
[6] Wikipedia: https://en.wikipedia.org/wiki/Garbage_in,_garbage_out
[7] David Cohen: http://www.adweek.com/socialtimes/presto/429910
[8] Sankalan Prasad: https://techcrunch.com/2016/02/19/optimizing-analytics-on-time-series-databases-a-supply-chain-perspective/?sf44687389=1
[9] Presto: https://prestodb.io/overview.html