Case Study

Sai Hemanth Nirujogi
G01065588

Facebook

Facebook is undoubtedly the world's most popular social network with over 2 billion users every day. This is the best example to show NoSQL databases used to store all the data. "Every 60 seconds, 317,000 status updates, 147,000 photos are uploaded, 510,000 comments are posted, and 293,000 status updates are posted." [1] The data is stored in several thousands of servers across the world. Facebook once used Cassandra for storage, one of the most used NoSQL databases which are used to store the massive amounts of data across many machines developed by Facebook.

Challenges

1. Data Sources

- Facebook collects user data from all the user activity in each login session. The user activity data includes likes, comments, shares, status updates, and messages exchanged between multiple users on Facebook. "According to stats released by Facebook, the system processes 2.5 billion pieces of information and more than 500 TB of data every single day. Facebook is collecting 2.7 billion likes information and 300 million images each day. With all this, Facebook's system scans approximately 105 terabytes of data every 30 minutes." [2]
- 60 million businesses have a Facebook profile, 39% the Facebook users follow other Facebook pages, 32% of the users participate with brands regularly and 42% of the page responses occur in the first hour. [2]
- This massive data is processed with 150 petabytes disk space in a single HDFS cluster, 70,000 queries executed, 105 terabytes of data processed via Hive every half hour and 900 terabytes new data consumed. [2]
- Facebook has a set of terms and conditions that have information about its data protection regulations. Few other social media platforms like Twitter, Facebook, Instagram etc. have some open access user data. Out of these platforms (Instagram) sell their data to other big companies for ads.

- Facebook has a strict privacy policy that always assures users that the personal information only shared with other marketers but the sold data is anonymized.
- Users always face issues with the privacy policy because of the privacy setting are too complex and not explained really clearly about it.

2. **Organizational Challenges**
   - The biggest challenges Facebook has faced over the years are:
     - *General data protection regulation and potential regulations (GDPR)* are one the bigger problems Facebook has faced in the EU. Europe issued a new privacy policy to the users, by making some changes to the operations.
     - Facebook has been facing *Ad load saturation* in its user news feeds, which made the application slowdown. Facebook also opened the ad inventory to Instagram too.
     - *User engagement and growth*: Since Facebook has reached over 2.2 billion monthly active users, it gets hard to let the user base grow. Even then it managed to consistently let the users grow 13 percent for the past several years. [3]
3. **Technical Challenges**

   Facebook just relies so much on one technology when there is so much information to handle. They have a massive and highly scalable Hadoop framework that uses inexpensive servers to solve any issues. Facebook also designed their own hardware to work with Hadoop. It has a 300 PB data warehouse to analyze the information every minute, by pulling out the required information into tables. They even built a search engine to find the required data easily by using indexing.

Stakeholders

In recent years, Facebook has been creating a buzz around the online privacy and data security. The enormous success of Facebook caused the company many privacy problems. Privacy to every user is a human right and monetizing the data that everyone cares about left the users with complex decision of deleting their accounts by giving up the comfort of finding connections to their old relatives and friends through it. Fixing the privacy issue and demonetizing the data is the main job of Facebook for the past couple of years. Users are really bothered about how their information is being used by the marketers without their consent.

After a massive security breach at Facebook and affecting a lot of users, Facebook has worked on its privacy issues. Cambridge Analytica, a U.K based political data analytics industry, illegally obtained the information of over 50 million Facebook users without their permission. The obtained information is then used to create the voter-targeting techniques for the presidential campaign of Donald J Trump. Facebook and Cambridge Analytica stated that the researcher who collected the data used a Facebook app that accesses the information on people who uses it. [4] Facebook has faced a data-privacy crisis that has led investors to panic and users angered. The big data securities are transparent to users and investors, because of that only few understand the risks.

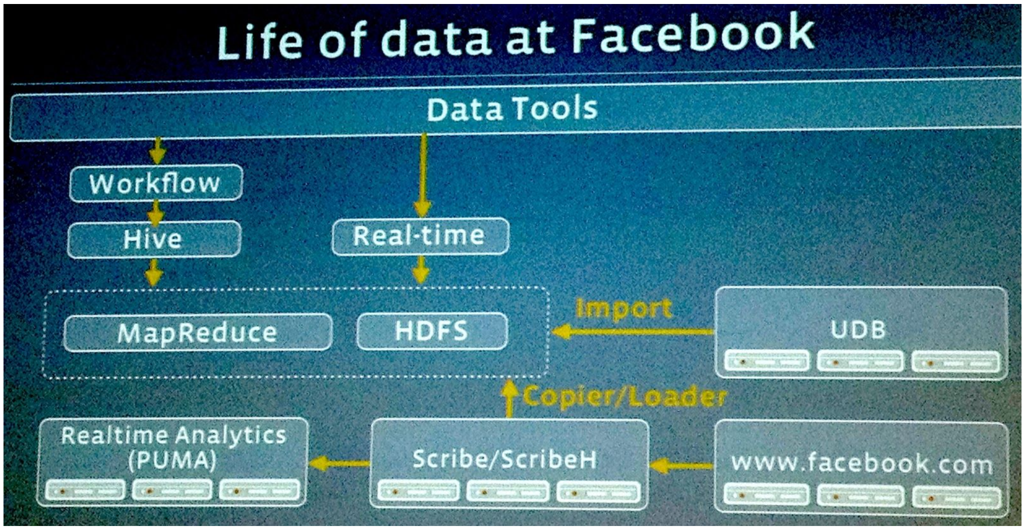
Requirements

At Facebook, the data warehouse clusters have been growing massively over the past few years. Facebook uses many open source software to keep the data processing efficiency in its data warehouses like Apache Hadoop, Apache Hive, Apache HBase, Apache Thrift, and Facebook Scribe.

Apache Hadoop is used by Facebook as a warehouse for data analysis, a storage for a distributed database and for backups from MySQL. Facebook has the worlds largest Hadoop cluster which runs on 4000 machines and stores thousands and millions of gigabytes of information. This also provides the developers with some new ability to analyze the data:

- Creating map-reduce programs efficiently in any programming language.
- Integration of NoSQL for processing massive datasets has been made possible because most of the data is stored in the Hadoop file system in tables. By pulling the subsets of data into tables helps developers to easily analyze and work on the data.

Hive is implemented at Facebook after Yahoo created the search engine based on Hadoop. Hive improved the query capability with Hadoop by using tables of data pulled from the SQL database and later gained acceleration in Unstructured databases. Processing many pieces of data and running multiple jobs at a time in the system has made quicker and easier using Hive.

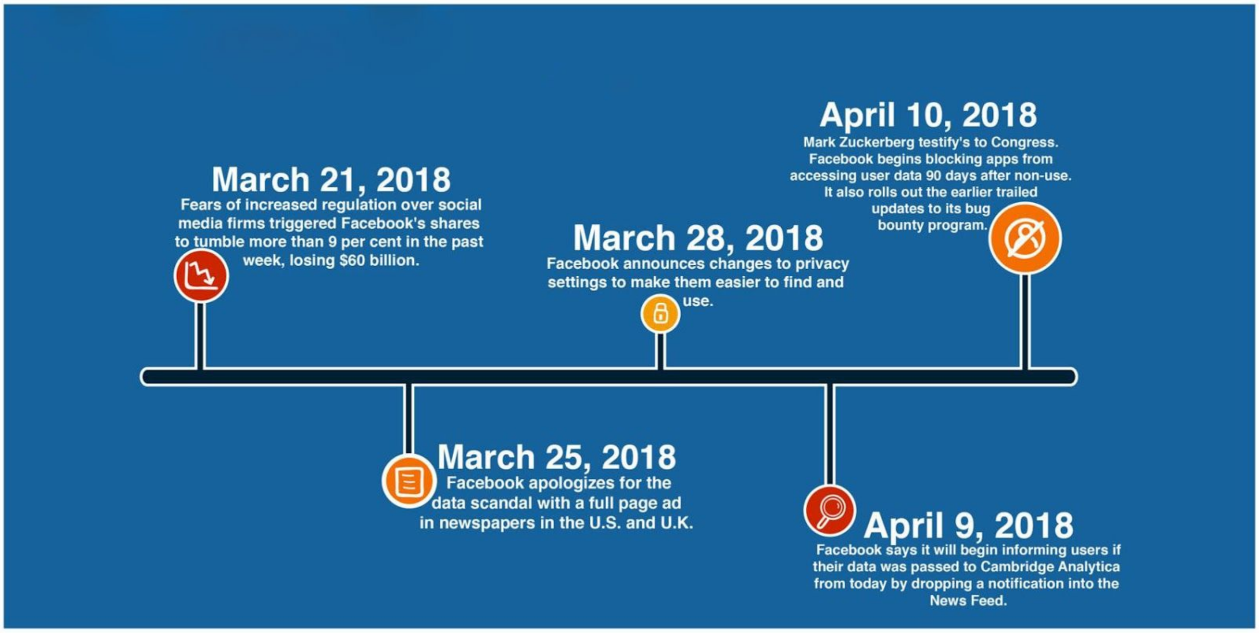


I. Life of Data at Facebook [2]

Corona is a software tool which allows multiple jobs to run on a single Hadoop cluster at once. When the Hadoop framework started misbehaving, developers at Facebook started to implement Corona. With such massive amounts of growing data every minute, it has got difficult for Facebook developers to manage all the cluster resources and task tracking applications. Carona's implementation and development helped in forming a new scheduling framework which separates the cluster resource management from job coordination. Facebook always pushes the limits of Hadoop and relies heavily on Hive, which helped them to use the software with standard BI tools. [5]

Facebook stores the entire live generated data in a single data center. When Facebook implemented Hadoop, it was not created to function across multiple data centers. "Prism is a platform, which brings out namespaces instead of the single one governed by the Hadoop, which in turn helps to implement many logical clusters".[5]

The Cambridge Analytica Scandal has started in the summer of 2014, the United Kindom affiliate of the United States political consulting firm Cambridge Analytica gathered the user profile information along with "Like" data of 300,000 Facebook users. Cambridge Analytica hired an American researcher, Alexandr Kogan to create an app called "This is Your Digital Life", which provides the Facebook users with surveys. This app collected data of users and their Facebook friends, only when the privacy settings allowed it. People who have used this app went to 50 million users and then grew to 87 million. It gathered the data to target voter with specific appeals, on Facebook and other apps. This has a name called "psychographic targeting".


March 21, 2018
Fears of increased regulation over social media firms triggered Facebook's shares to tumble more than 9 per cent in the past week, losing $60 billion.
March 28, 2018
Facebook announces changes to privacy settings to make them easier to find and use.
April 10, 2018
Mark Zuckerberg testify's to Congress. Facebook begins blocking apps from accessing user data 90 days after non-use. It also rolls out the earlier trailed updates to its bug bounty program.
March 25, 2018
Facebook apologizes for the data scandal with a full page ad in newspapers in the U.S. and U.K.
April 9, 2018
Facebook says it will begin informing users if their data was passed to Cambridge Analytica from today by dropping a notification into the News Feed.

II. Timeline of Cambridge Analytica Scandal [6]

Results/Findings

- With so much invasion in the privacy of over 87 million users, by revealing data to 3rd party marketers and Facebook started letting all the users know that their profile is affected. They started sending alerts to users who got affected by the Scandal and users whose information is used for the “This Is Your Digital Life”. A facebook help center page with an option to tell the users how the app used the personal information.
- Facebook changed all the privacy settings by centralizing the app settings. This provides a link to the page with all the third-party apps that have permission to the information and can toggle to stop permitting the apps to use the information.
- Facebook announced that the platform stopped allowing third-party apps to access the information from the friends of friends. Facebook removed the API which used to allow the third party app developers to access the information on a user’s friends.

Mark Zuckerberg has made clear that there are several steps to further protect the user information. The third-party apps will only be linked to the account for three months and allow only for one-time use apps from checking the information in the background. In April, they made a lot of changes to the API which allows the third-party app access, limits access to Groups, and pages from events API.

Facebook will be sending notifications to the users if their information is being misused by any third-party applications. [7]

- Facebook disabled some features that helped users to find their friends easily using their email or phone number. They made users find other people hard to find by disabling that feature to block others to acquire public information to multiple accounts.
- In a previous couple of months, Facebook ended the ad program called "Partner Categories" that enabled marketers that pairs information to offline and online accounts to the Facebook account. [7]

Critique

Even though Facebook has fixed the bug and created an alert service that notifies users if their account is impacted by the CA scandal, it is still prone to project the information to all the information to other ad marketers. To prevent this from happening again, Facebook should provide users with less complex and more reliable privacy settings. Zuckerberg blocked apps from accessing user data 90 days after no usage. But the problem is, how many of the developers accept these changes and rolls with Facebook and their changes.

Political ads are not regulated like they are on Tv and on the radio even after the "Fake News" issue after the elections in 2016. The internet life has become like a part of living in a society with battered messages and ads that we don't like in our feed.

References

1. Monnappa, Avantika. “How Facebook Is Using Big Data - The Good, the Bad, and the Ugly.” *Simplilearn.com*, Simplilearn, 6 July 2018, accessed on 6 Oct. 2018, www.simplilearn.com/how-facebook-is-using-big-data-article.

2. Constine, Josh. “How Big Is Facebook's Data? 2.5 Billion Pieces Of Content And 500 Terabytes Ingested Every Day.” *TechCrunch*, TechCrunch, 22 Aug. 2012, accessed on 17 Oct. 2018 techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/.

3. Levy, Adam. “The 3 Biggest Challenges Facing Facebook.” *The Motley Fool*, The Motley Fool, 2 June 2018, accessed on 17 Oct. 2018 www.fool.com/investing/2018/06/02/the-3-biggest-challenges-facing-facebook.aspx.

4. Spangler, Todd. “Facebook Under Fire: How Privacy Crisis Could Change Big Data Forever.” *Variety*, Variety, 3 Apr. 2018, accessed on 6 Oct. 2018, variety.com/2018/digital/features/facebook-privacy-crisis-big-data-mark-zuckerberg-1202741394/.

5. Chakraborty, Shravani. “How Is Facebook Deploying Big Data? - DZone Big Data.” *Dzone.com*, 6 Feb. 2017, accessed on 19 Oct. 2018 dzone.com/articles/how-is-facebook-deploying-big-data.

6. Goh, Zhi Jiao Danielle. “Facebook's Cambridge Analytica Data Scandal.” *Los Angeles Loyolan*, 27 Apr. 2018, accessed on 21 Oct. 2018 www.laloyolan.com/news/facebook-s-cambridge-analytica-data-scandal/article_702b3c7a-f972-539c-a7f8-f202fd478634.html.

7. Ivanova, Irina. “8 Promises from Facebook after Cambridge Analytica.” *CBS News*, CBS Interactive, 10 Apr. 2018, accessed on 21 Oct. 2018 www.cbsnews.com/news/facebooks-promises-for-protecting-your-information-after-data-breach-scandal/.