# Twitter's Challenge: User Base and Monthly Active Users
# Technology Solutions

Prepared by Group 4: Haoyou Li, Qiutong Xu, Fan Yang, Yawen Wu, Xiaorui Zhu

**Examine the Business Requirements**

The challenge of Twitter's shrinking share in the North American market is escalating, challenging Twitter's leading position in the global social media industry. The core objective of this project is to increase the user base, monthly active users, and advertising revenue of Twitter by analyzing the structured and unstructured data currently owned by the company. By doing so, we DI team can distill the characteristics of target users, benefiting marketing team to explore new markets, seize more users, and attract more investment.

**Examine the nature of the data**

The data that can be utilized in this project can be classified as two kinds: structured and unstructured data. Structured data like DAU(daily active users), MAU(monthly active users), id, date, age, daily tweets, monthly tweets and number of followers are essential indicators to measure the success of a social media platform, and they are the prime data that required by the marketing team and need to be analyzed in this project for marketing strategies design. Unstructured data include text, photo, videos, and other forms of contents that are tweeted or forwarded in every second. They are data of higher level that can be used to analyze trends, keywords, and target groups of certain advertisements. Technical requirements of fulfilling unstructured data analysis will be higher as well, asking more expensive data engine like Spark, to achieve the expected effects.

**Database Engines Evaluation**

| | SQL | NoSQL | Hadoop | Spark |
|---|---|---|---|---|
| Strength | Digital; easy to use; use long-established standard adopted by ANSI & ISO; power and flexible for raw data manipulation; SQL Queries can be used to retrieve large amounts of records from a database quickly and efficiently | Easy to manage due to table-less; Flexibility, able to handle structured, unstructured data; Agility; Elasticity and scalability; expand transparently and easier to scale out; open-sourced | Open source; Low cost; Efficient; Storage flexibility; fault tolerant due to the availability of backup data; Computation spread on endless servers; Hadoop cluster can speed up large data process time | High speed, can be 100x faster than Hadoop in dealing with large quantities of data; Some APIs are easier for Twitter to use in large datasets; Unified Engine: Some higher level libraries in Spark can effectively increase the productivity of developers and build complicated workflows |
| Weaknes s | High cost; Unable to handle unstructured data; difficult to scale as a database grows larger; | Data consistency and language consistency problem; Immaturity, less stable | Security issue; Slower speed to process data through hard disks; Not suitable for small volumes of data | Expensive to keep data in in-memory; No mainstream file storage module so it is necessary to |

| | | | | |
|---|---|---|---|---|
| | Sharing is quite problematic | | | combine Spark with Hadoop or other data platform |

**Primary recommendation / Configuration 1**

Based on the specific needs of Twitter, we decide to choose MySQL and NoSQL as our main engines. The data input will be the specific variables that required by the analytics team, which include the exact number of the daily and monthly active user, customer ID, total tweets, topic, followers, text etc.(images and videos are not included). Since the nature of these required inputs are highly structured and aggregated, the indicators will be able to be extracted in a standard format, therefore, we can use the relational database to handle these data. The indices are loaded when imported so we can apply present algorithms or models to analyze and get results constantly. In addition, the features of open-source engines enable us to easily optimize the function as Twitter's business involves. For more flexible data model that MySQL is hard to overcome, such as key-value stores, document databases, wide-column stores. we would use NoSQL to deal with them.

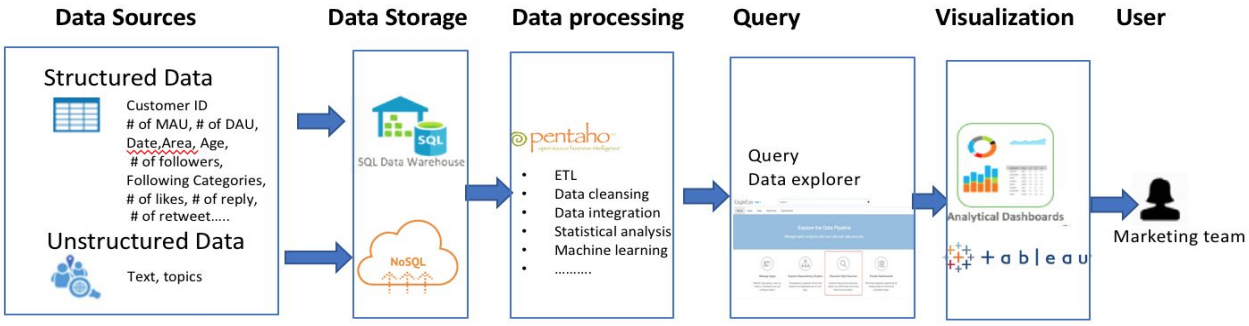


***Figure 1: The flow and logic of Configuration 1***

First, for the data inflow, all the required data will be gathered, ingested and stored within the data warehouse( MySQL), or the data lake (NoSQL) depending on the structured or unstructured feature of data. Next, the processing tool of NoSQL database such as Pentaho will proceed with the extract, transform and loading processes. Myriad of third-party plugins of Pentaho Data Integration (PDI) enables us to orchestrate many data manipulations with Pentaho such as data cleansing, data integration, statistical analysis, machine learning, as well as other advanced operations (Database zone). Lastly, we will develop a data explorer User interface(UI) which will integrate all the metadata, the output provided by the processing layer. The UI can serve as an in-house ETL tool for data transformation across different backends such as SQL or NoSQL. The analytical team can use it to create dashboard and import data into Tableau for visualization. This system will be handy and suits the users of marketing team well, especially those lacking coding experience.

According to the acceptance criteria of the business requirement, this project needs to be delivered within 3 months and under 1 million dollars budget. Therefore, the approach we addressed above is the maximum work we can do in 3-4 months. However, we believe that in

order to successfully expand the user base of Twitter, more unstructured data need to be analyzed. Insights from a huge amount of daily bulky text, videos, and images are also extremely important to portray users and thus to tailor marketing strategy and boost advertising income. Considering this, we developed our secondary recommendation.

**Secondary recommendation/ Configuration 2**

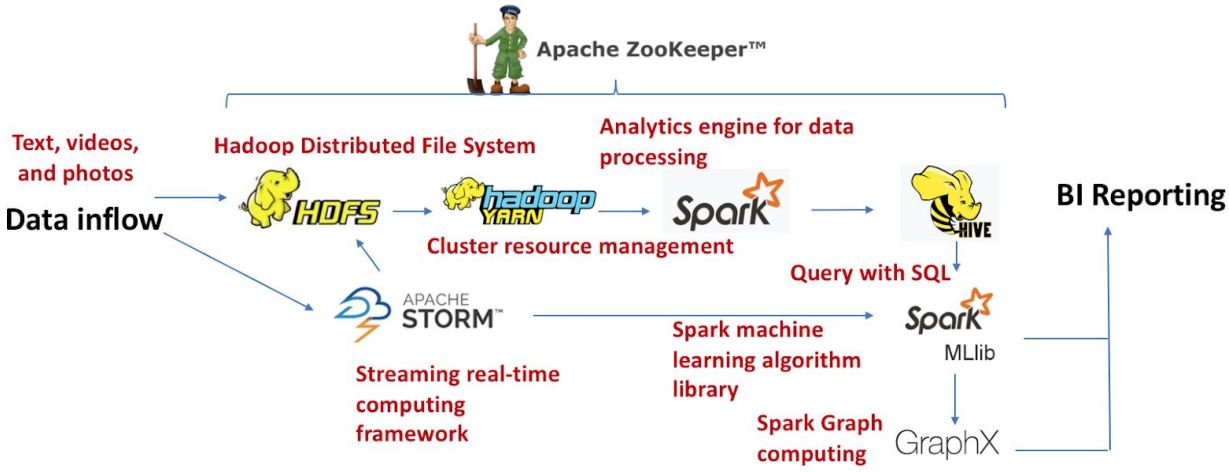


***Figure 2: The flow and logic of Configuration 2***

Configuration 2 mainly consists of three layers.

First, in terms of storage, data lake is appropriate for unstructured data. We decide to use Hadoop Distributed File System (HDFS) to deal with the exploding and highly unstructured data of Twitter. We find HDFS a good fit because the data can be stored on multiple devices and we have paths as the reference to obtain data in need.

Second, for computing and processing, we decide to choose Spark. Spark, as the second generation of MapReduce computing engine (computing core of Hadoop), is able to distribute a significant amount of workload to different machines and summarize the results efficiently. Its standalone cluster mode on Hadoop can help us access data in any Hadoop data source while supporting with other data analysis and database management functions. Compared with the heavy bulky initial MapReduce, Spark unifies the Map/Reduce model and makes data transfer even more flexible. Thus, we choose Spark to handle the large volume of unstructured throughput of Twitter.

The business requirement also mentioned analysis of real-time topics so we choose Storm to provide real-time computing to serve the special need. Since the platform computes the target index such as vocabulary frequency when the data flows through, the marketing team can take agile movements based on insights generated by the analytical team. Third, in terms of query, we will use Spark SQL, a built-in library, to deliver SQL queries.

**Data Lake Architecture Recommendation**

For this project, NoSQL and HDFS are two kinds of data lake we would like to use for primary and secondary recommendation respectively. As we discussed above, our primary recommendation will mainly focus on structured data - text will be the only format of unstructured data addressed. Since NoSQL is scalable, relatively cheap, open-sourced, and abundant in storage system options, we find it appropriate to handle the business needs of

Twitter. However, when it comes to unstructured data like videos and photos, NoSQL is not strong enough. In addition, NoSQL is less efficient in data backup than HDFS, an important function for social media companies like Twitter to initiate deeper analysis afterward. However, the risk of no storage or network level encryption is still there, security will be an issue (Koirala, 2013).
In order to ensure the sustainability of the project and minimize the risk of data loss, system crash, and transformation error, we developed our data governance framework.

**Data Governance / Recovery Continuity of Business**
Our main objective of data governance is to assure data quality regarding accuracy, accessibility, consistency, completeness, and promptness.

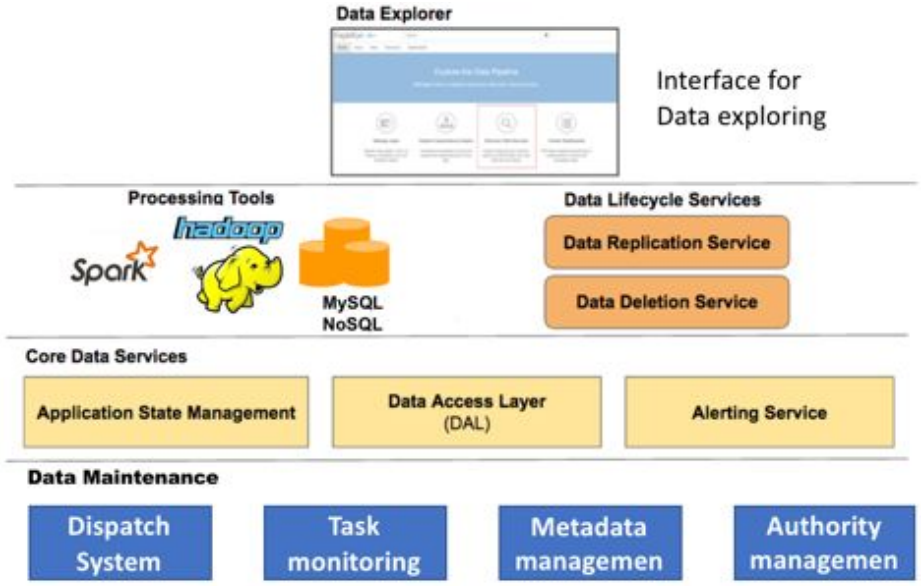


***Figure 3: the flow of Data governance***
At the bottom of our governance are the four maintenance blocks. Dispatch system manages the big picture and coordinates multiple lines. Task monitoring center oversees the status and progress of each task and reports anomaly. Metadata management keeps updating the context of data storage. Authority management confirms the responsibility of and applies mutual restraints to each user. Above the principles are core data services. The three segments address the assess of database, management of ongoing tasks, and alert when abnormal detected respectively. These procedures serve as checking points that alert when job delays (Krishnan,2016). Data Lifecycle Service includes Data Replication Service and Data Deletion Service, where data are replicated to multiple copies for backup, and deleted after three years retention. Besides, aligning policy and accountability are the two foundations of our data quality control.

**References**
**Please refer to the reference by clicking on the embedded icon.**

7 Pros and Cons of NoSQL. GreenGarage. Retrieved from https://greengarageblog.org/7-pros-and-cons-of-nosql

Buckler, C. (2015). SQL Vs NoSQL: How to Choose. Sitepoint. Retrieved from https://www.sitepoint.com/sql-vs-nosql-choose/
Acodez. (2017). An Overview of SQL and NoSQL with It's Pros and Cons. Acodez. Retrieved from https://acodez.in/sql-and-nosql-an-overview/
Harrison, G. (2010). 10 things you should know about NoSQL databases. TechRepublic. Retrieved from https://www.techrepublic.com/blog/10-things/10-things-you-should-know-about-nosql-databases/
Tozzi, C. (2016). The Limitations of NoSQL Database Storage: Why NoSQL's Not Perfect. ChannelFutures. Retrieved from http://www.channelfutures.com/cloud-services/limitations-nosql-database-storage-why-nosqls-not-perfect
MongoDB. (2018). Advantages of NoSQL. Retrieved from https://www.mongodb.com/scale/advantages-of-nosql
NoSQL. NoSQL Definition. Retrieved from http://nosql-database.org
What is Apache Spark? (n.d.). Retrieved March 25, 2018, from https://databricks.com/spark/about
(n.d.). Retrieved March 25, 2018, from https://mapr.com/ebooks/spark/04-hadoop-and-spark-benefits.html

Lo, F. (2017). What is Hadoop and NoSQL? Retrieved March 25, 2018, from https://datajobs.com/what-is-hadoop-and-nosql

SQL Advantages & Disadvantages,Retrieved March 25, 2018, from www.cs.iit.edu/~cs561/cs425/VenkatashSQLIntro/Advantages & Disadvantages.html
Vijayalakshmi, Konduru. "What Are the Advantages and Disadvantages of SQL?" *Quora*, 31 July 2017, www.quora.com/What-are-the-advantages-and-disadvantages-of-SQL

Borysowich, Craig. "Some Pros & Cons of Relational Databases." *Tech*, ToolBox, 1 May 2008, it.toolbox.com/blogs/craigborysowich/some-pros-cons-of-relational-databases-050108
Tweet Object - Twitter Developers. (n.d.). Retrieved February 18th, from https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object

Fiegerman, S. (2017, JUly 27). Twitter Now Losing Users in the U.S. Retrieved February 10, 2018, from http://money.cnn.com/2017/07/27/technology/business/twitter-earnings/index.html
Krishnan, S. (2016, June 29). Discovery and Consumption of Analytics Data at Twitter. Retrieved April 18, 2018, from https://blog.twitter.com/engineering/en_us/topics/insights/2016/discovery-and-consumption-of-analytics-data-at-twitter.html
Fowler, A. (2017, December 26). Why NoSQL Needs Schema-Free ETL Tools - DZone Database. Retrieved April 18, 2018, from https://dzone.com/articles/why-nosql-needs-schema-free-etl-tools
Brathwaite, C. (2015, December 11). Pentaho Data Integration (Kettle) Tutorial - Pentaho Data Integration. Retrieved April 18, 2018, from https://wiki.pentaho.com/display/EAI/Pentaho%20Data%20Integration%20(Kettle)%20Tutorial
T. (n.d.). The DGI Data Governance Framework. Retrieved April 18, 2018, from http://www.datagovernance.com/the-dgi-framework/
Koirala, P. (2013, December 13). Top Big Data Technologies and Tools- Hadoop and NoSQL Ecosystem. Retrieved April 18, 2018, from http://blog.venturesity.com/top-big-data-technologies-and-tools-hadoop-and-nosql-ecosystem

Bopardikar, Y. (2016). TWITTER DATA ANALYSIS USING SPARK. Retrieved April 17, 2018, from https://csus-dspace.calstate.edu/bitstream/handle/10211.3/182694/Bopardikar_Yash_Masters_project_Report.pdf?sequence=1