TWITTER'S TECHNICAL SOLUTIONS

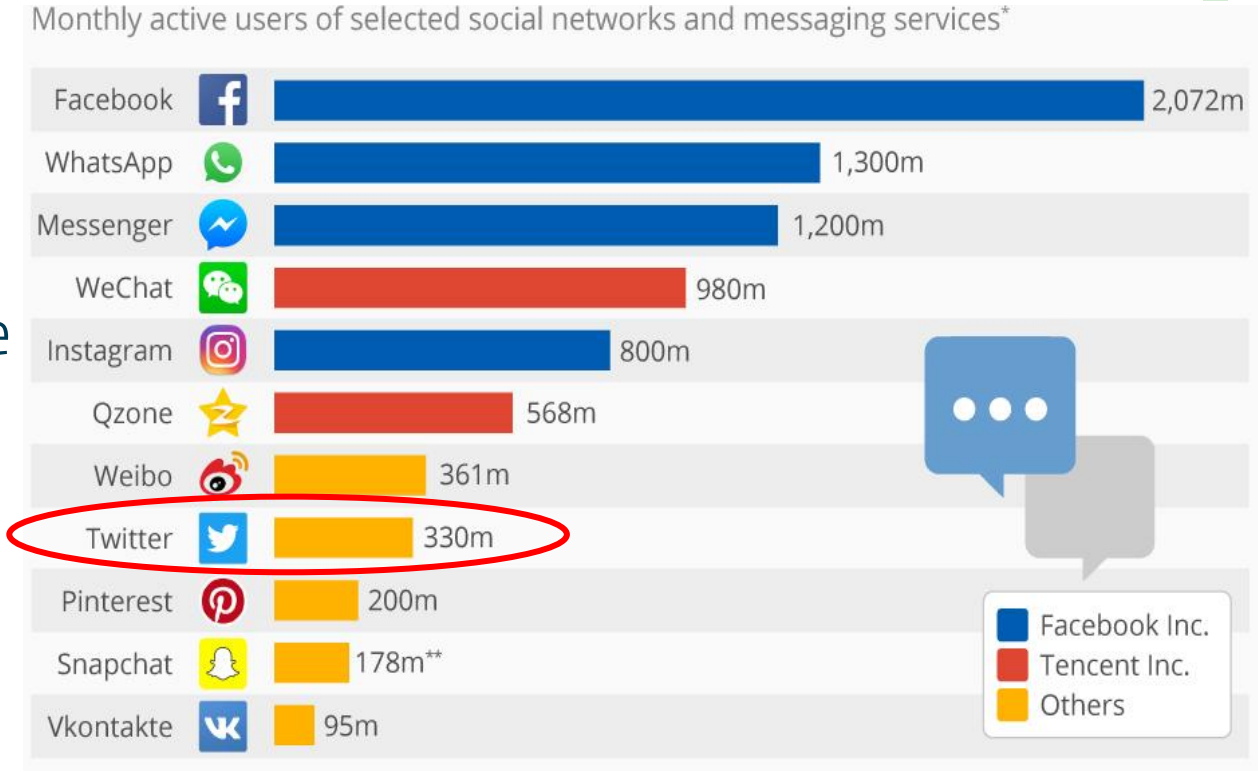
-Expand User Base & Monthly Active Users

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

COLUMBIA UNIVERSITY
School of Professional Studies

Overview of Business Requirements

Increase the user base, monthly active users, and advertising revenue of Twitter



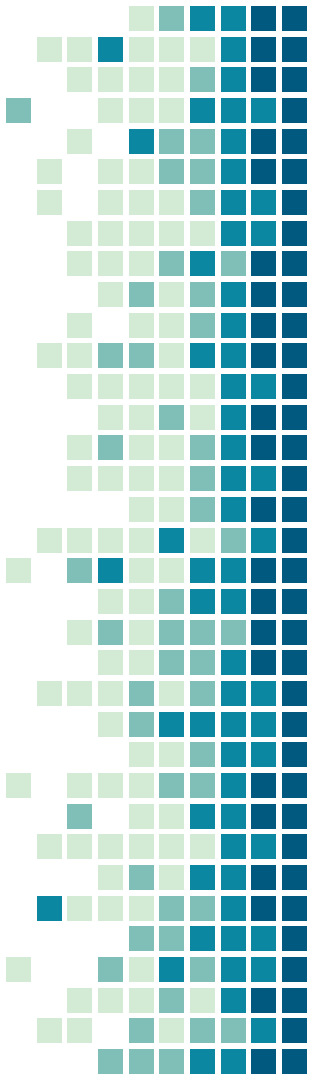
Source: Statista.com

Executive Summary

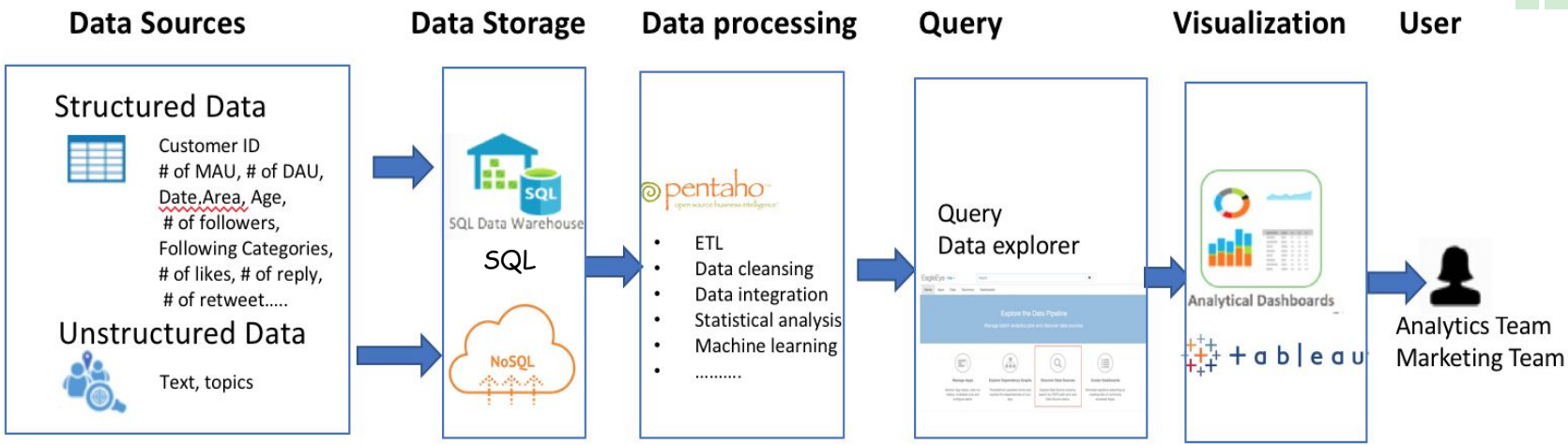
- Nature of data: structured and unstructured
- Primary plan: MySQL & NoSQL
- Secondary plan: HDFS & Spark
- Governance: Data maintenance, Data Alerting service, Data lifetime service

Nature Of Data Requirement

Abbreviation / Variable Name	Type	Description
DAU	Integer	Daily Active Users
MAU	Integer	Monthly Active Users
ID (id)	String	ID represents the unique identification for each Twitter user.
Initiation (created_at)	String	Age represents the time when the account is registered.
Date (date)	String	The date of the tweet.
Area(area)	String	The geographic location of each Twitter user.
Topics(topic)	String	The topic of a tweet, or the topic that a tweet is related to.
Age(age)	Integer	The age of each Twitter user, which is estimated by the user's date of birth.
Daily Tweets (day_tweets)	Integer	The number of Tweets the user posts on a daily basis.
Total Tweets (total_tweets)	Integer	The total number of Tweets that the user has posted since the account was created.
Number of Following (following)	Integer	The number of other users each Twitter user follows.
Number of Followers (followers)	Integer	The number of followers each user has.
Following Categories (categories)	String	The trends or categories that the user is following.
Text (text)	String	The actual text/content of each Tweet.
Likes (likes_count)	Integer	The number of liked Tweets for each user.
Topics of Likes (likes_topics)	String	The type of the topics each user has liked.
Reply (reply_count)	Integer	The number of times the user has replied to other Tweets or comments.
Retweet (retweet_count)	Integer	The number of times the user has retweeted other Tweets.



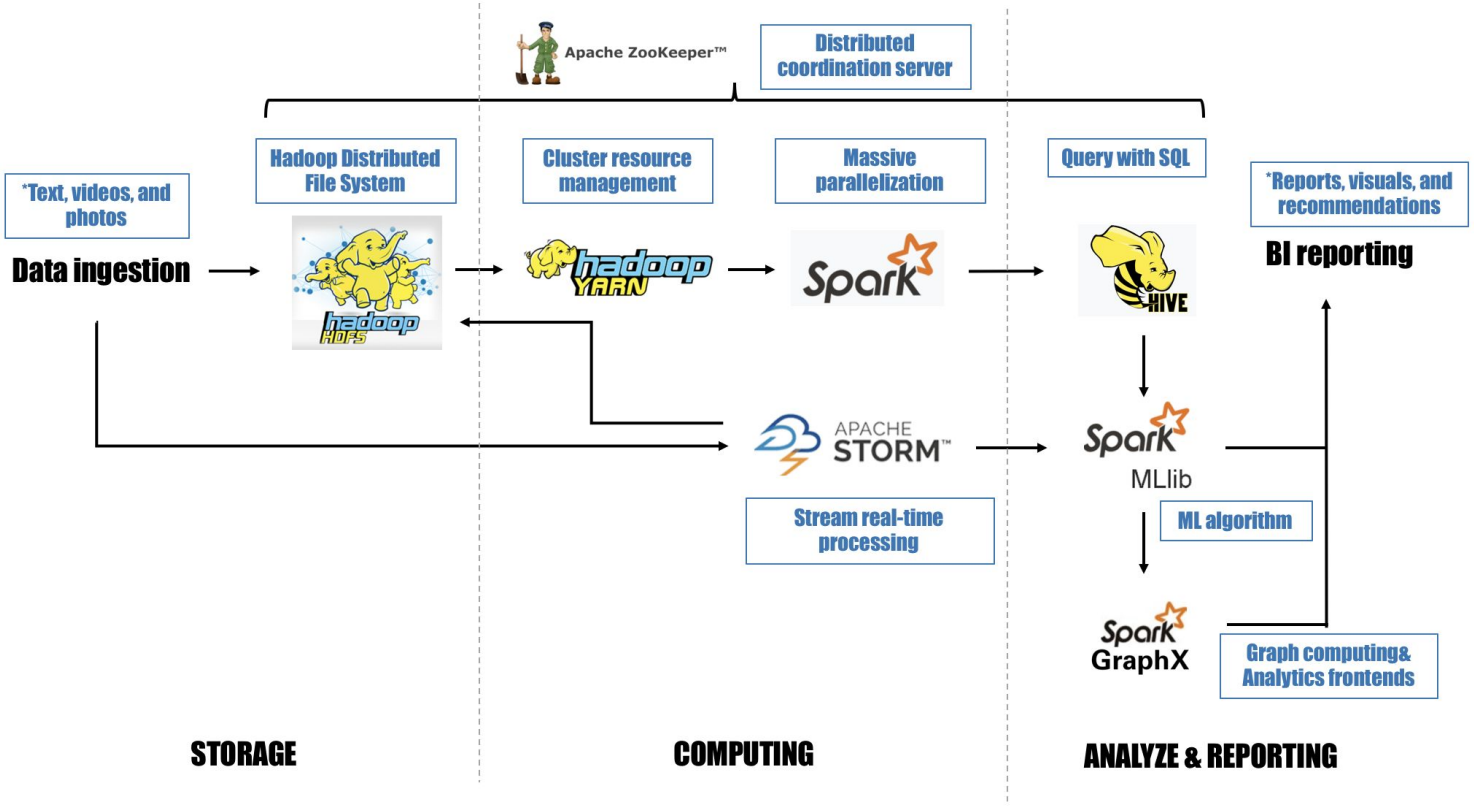
Primary Configuration



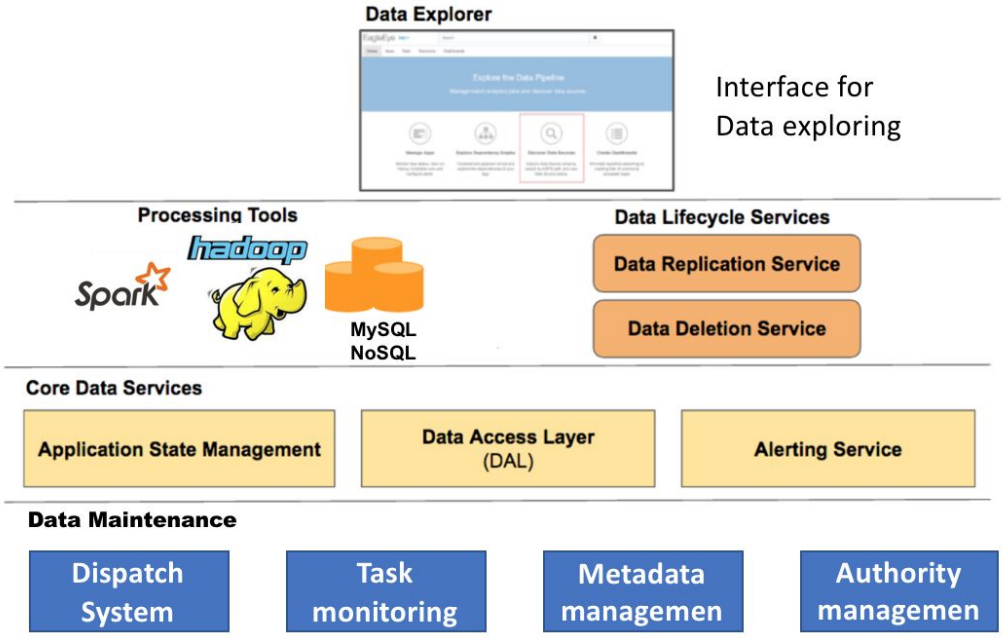
- Exactly focus on the variables required by Twitter Analytics team.
- Text is the only format of unstructured data addressed by primary configuration.

- In-house ETL tool
- Searching for data
- Create dashboard
- import to Tableau

Secondary Configuration



Data Governance



- 3 Layers
  - Data Lifecycle Services
  - Core Data Services
  - Data Maintenance

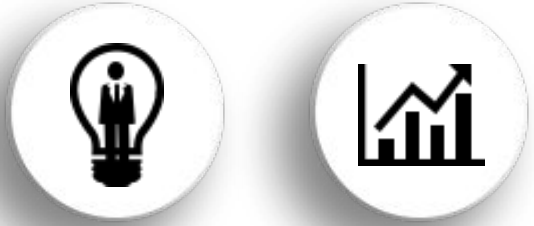
Human Layer

Objective

Understand key characteristics of Twitter's user ⇨ **Descriptive Analytics**
Anticipate future user growth points ⇨ **Predictive Analytics**

Functionality

- Analyze data to get customer portrait
- Identify trends of social media



Refresh Rate

Weekly basis for Descriptive Analytics
Monthly basis for Predictive Analytics

Conclusion

	Configuration 1	Configuration 2
Input	Mostly structured	Mostly unstructured
Storage	MySQL+ NoSQL	HDFS of Hadoop
Processing	NoSQL processing tool (Pentaho)	Spark
Query	Developed an in-house data explorer UI	SQL
Objective	Analyze the basic usage pattern	Deeper customer portray analysis
Budget & Time	Can be finished within the time restriction	Need more negotiation


THANK YOU
Q&A