MemoryWhole

A Distributed Deleted Tweet Discovery Service

By Rob Gevorkyan, Michal Guzek, and Ai Enkoji

Introduction/Motivation

- Twitter is a major player in mass communication today.
- Certain tweets can carry significant political or social weight and may be suppressed by various parties.
  - Political discussions globally.
  - Controversial tweets from various public figures including candidates for office, celebrities, etc.
- We introduce MemoryWhole, a scalable service for archiving, detecting, and presenting deleted tweets.

Architecture

- Hadoop framework
  - Hadoop common: utilities and libraries
  - HDFS: clustered storage for high throughput access
  - YARN: Manages resources and schedules jobs
  - MapReduce: splits, maps, and reduces data

Tools and Techniques

- For each user, we collect a growing sequence of batches 1, 2, … , t containing all historical tweets from some beginning time frame to current time. Over time, t grows to reflect the up-to-date tweets made/deleted.
- A tweet is considered deleted if it appears in some batch i, is present in all tweets i, i + 1, … , j, and then is missing for all batches k > j
- Our program proceeds as follows
  - Fill a HashMap data structure with all previously processed data (with tweet IDs as keys) or initialize to empty if it is the first batch for a user
  - Reducer performs a hash table lookup to determine tweets that were seen before. These are deleted from the HashMap
  - The tweets that remain are those that are deleted

Evaluation

- Data simulation
  - Use a Python script to generate an arbitrary number of batches for one or more users
  - Each batch contains a JSON array of tweet data closely modeled after the real format of Twitter API responses
  - The collection of batches is guaranteed to have m missing tweets, where m is specified as a parameter
  - Tweet text indicates whether a tweet is deleted later or not. So it was simple to simply check whether
    - All m deleted tweets are detected as deleted.
    - No non-deleted tweets were falsely detected
  - Our solution successfully distinguished all cases

Conclusion and Future Work

- Higher velocity data
  - Shorter time frames, larger data loads among many more users
- Classification
  - Categorize deleted tweets by hashtags (basic) or topic modeling (advanced)
- Generalize solution to other social media
  - There is nothing in our system that requires Twitter strictly. The same techniques could be applied to Facebook, Instagram, and other social media.