Twitter Analytics and Hadoop
Sarah Silverman @SarahKSilverman
20 Sep
When ur relatives drive you crazy just close your eyes & pretend it’s dialogue in a woody allen movie
Details

mia farrow
@MiaFarrow
@SarahKSilverman tried that. Didn’t work
RT When ur relatives drive u crazy just close yr eyes & pretend its dialogue in a woody allen movie

WHAT IS TWITTER?
Twitter usage over time

% of internet users who use Twitter

Source: Pew Research Center’s Internet & American Life Project Winter 2012 Tracking Survey, January 20-February 19, 2012. N=2,253 adults age 18 and older, including 901 cell phone interviews. Interviews conducted in English and Spanish. Margin of error is +/-2.7 percentage points for internet users (n=1,729).
About Twitter

• The fastest, simplest way to communicate

• More than 140M active users
  • Majority (also) mobile; 60% out of U.S.

• More than 400M twitter.com visitors

• More than 400M tweets/day (peak: 25K/sec)

• 1,000 employees (majority in San Francisco)
  • 50% engineers
Twitter data: text

- Caleb Howe @CalebHowe
  Glass tubes recovered from Solyndra have been turned into an art exhibit at Berkeley. RECYCLING. “Now” I feel like it was money well spent.

- Todd Kincannon @ToddKincannon
  What a bunch of misogynists! RT @JosWillard And those folks at UC Berkeley - Completely Anti-Women!

- Cal @cal
  The irony of Todd Akin bit.ly/PBubxF #cal

- MISS TOPNOTCHSKILLZ @TOPNOTCH
  @EarPeaceRecords presents END OFcapture at Queen's Berkeley. Performing Live Kansho Kuroki of @PatienceSFC, & More pic.twitter.com/... View photo

- Ken Goldberg @ken_goldberg
  Suggest a question for William Gibson on stage sept 4: opinion.berkeley.edu

- Carlo Mazzocco @Mazzocco26
  We get it. Everyone not from Berkeley party... Now stop.

- richlyons @richlyons
  Berkeley among the top 5 university bs... bit.ly/SjI/Erq

- UroToday.com @urotoday
  Beyond the Abstract - N-acetylimidazole and prostate cancer ... BERKELEY, bit.ly/SjI/Erq

- ...
Twitter data: time series

#Euro2012
A summary for the action on Twitter during the European football tournament.

The Streamgraph below shows volume of Tweets during the #Euro2012 period. Click on a team’s name to see details.

Countdown to Super Tuesday
Daily new follower growth for the four leading GOP candidates – January 1, 2012 to March 5, 2012
Follow @gaw for other interesting data during the campaign.
Combined: the pulse of the world
Examples of Analytic Tasks

Search

Ads

Recommendations

Anti-Spam
What happens when you “Tweet” this message?
Older Twitter Model:

Probably Worlds largest Ruby on Rails Website

Everything was being written on a big monolithic stack. Doesn’t scale, 400 engineers work on same code base, no independence to team, too much time spent in co-ordination.
Typical Twitter Events

- Handles about 500k/sec events and 300 QPS from a pool of about 150 Million Users
- Tweets, follow, share, re-share, unfollow
- Move the monorail app from a monolithic RoR / MVC to event driven infrastructure
- Event driven framework allows distribution and de-centralization of processes/infrastructure
- Something goes wrong in some part of monorail, the whole website goes down
what are the goals?

- evolve from being solely a web stack
- isolate responsibilities and concerns
- site speed and reliability
- developer innovation speed

Everything has to happen in Real-Time.
Event driven programming model to understand when a Tweet was posted, when someone replied etc.
This is needed to push tweets in user’s timeline. The timeline is replicated three times.

Not saved to disk..!
Stored in RAM, Allows fast recovery: 45 ms
Only active users in past 28 days (LRU) stored in RAM.
Rest goes on Disk.

800 tweets per home timeline.
Rest is stored on Data-centers

Use Write API to Write tweet in DB
Fan out is delivering the Tweet to every single person who is following that person

Redis Cluster: user-id = key, Tweet = value...!
(Map part of)
Map Reduce Programming

Timeline Service figures out where person’s Timeline lies in Redis cluster.
(Reduce part of Map-Reduce)

The goal is avoid hitting the disk as much as possible...!
Twitter Fanout

• Suppose you have 20K followers.
• Fanout daemon will look up location of all 20K users inside Redis Cluster. Then it will start inserting the Tweet ID of the tweet into all those lists through redis cluster
• What is being stored is the tweet ID, the user ID of the originator of the tweet and a marker to mark if it’s a retweet or a reply
Anatomy of a User

An example of a Twitter user's profile data:

- **id**: 6253282
- **id_str**: "6253282"
- **name**: "Twitter API"
- **screen_name**: "twitterapi"
- **location**: "San Francisco, CA"
- **url**: "http://dev.twitter.com"
- **description**: "The Real Twitter API. I tweet about API changes, service issues and happily answer questions about Twitter."
- **protected**: false
- **followers_count**: 1217031
- **friends_count**: 31
- **listed_count**: 10784
- **created_at**: "Wed May 23 06:01:13 +0000 2007"
- **favourites_count**: 25
- **utc_offset**: -28800
- **time_zone**: "Pacific Time (US & Canada)"
- **geo_enabled**: true
- **verified**: true
- **statuses_count**: 3336
- **lang**: "en"
- **status**: {
  - **created_at**: "Thu Sep 06 17:55:54 +0000 2012"
  - **contributors_enabled**: true
  - **is_translation**: false
  - **profile_background_color**: "CODEED"
  - **profile_background_image_url**: "http://a0.twimg.com/images/themes/theme1/bg.png"
  - **profile_background_image_url_https**: "https://si0.twimg.com/images/themes/theme1/bg.png"
  - **profile_background_tile**: false
  - **profile_image_url**: "http://a0.twimg.com/profile_images/2284174872/7df3h38zabcviylnyfe3_normal.png"
  - **profile_image_url_https**: "https://si0.twimg.com/profile_images/2284174872/7df3h38zabcviylnyfe3_normal.png"
  - **profile_link_color**: "0084B4"
  - **profile_sidebar_border_color**: "CODEED"
  - **profile_sidebar_fill_color**: "DDEEF6"
  - **profile_text_color**: "333333"
  - **profile_use_background_image**: true
  - **show_all_inline_media**: false
  - **default_profile**: true
  - **default_profile_image**: false
  - **following**: null
  - **followers_request_sent**: null
  - **notifications**: null

---

**Twitterapi** Twitter API

The Real Twitter API. I tweet about API changes, service issues and happily answer questions about Twitter and our API. Don’t get an answer? It’s on my website.

http://dev.twitter.com

---

**Followers 988,982**  **Following 33**
Use a separate pipeline for search. It allows to index a tweet only once (unlike fanout process). This helps searching through tweets of followers/retweets.

This architecture allows to blend in copied/followed tweets into respective users timeline.

The pattern is searched across datacenter and a software called Blender searches, sorts, merges and reranks the results.
Push Me Pull Me Architecture

Pull Me

• Targeted timeline: Tweets are delivered to you because you asked to see them. Pull based delivery: you are requesting this data from Twitter via REST API call
• Query timeline (Search API): Return all tweets that match particular query

Push Me

• Open a socket to twitter and they will push all public tweets in 150msec
• At any given time there are about 1M sockets open
• Goes to firehose clients. All public tweets go out these sockets
• User Stream Connection: When you login they look at your social graph and only send messages out from people you follow, recreating the home time-line experience
Search Operation

• It is opposite of pull / write operation.
• As tweet comes in, the ingester tokenizes and figures out everything they want to index against.
• It is stored in Early Bird machine (a modified version of Apache Lucene). It allows to index a tweet only once (unlike fanout process). This helps searching through tweets of followers/retweets. Hence, write and search pipeline are different
• The pattern is searched across datacenter and a software called Blender searches, sorts, merges and reranks the results.
• Search process is based on what they know about you. And they know a lot about you because you follow lot of people, click of shared links. All this information is stored and used to provide personalized search
Analyze influence

Read Path: when you are trying to search for a tweet

Write path

Search Operation

Regular tweet

Access via web/mobile
Twitter Analytics data flow

Servers in production to handle incoming traffic
Analyzing Machine Generated Data

• Searching, monitoring and analyzing machine generated big data via web interface
• Allows real-time response model when servers/clusters fail
• Allows trend detection/understanding unpredicted events
• Widely used in web-analytics
Lessons Learned

• Refactor architecture for flexibility/maintenance ease and not just performance.
• Modify existing programming methodology to suit your business/application needs (in case of Twitter store in RAM).
• Modular teams/infrastructure better suited for scale than trying to grow a big monster.
• Separate Data Generation process and searching process.