Barcelona GSE Data Science Center

Agrius

November, 2019

Motivation

Researchers from the Reuters institute asked us to collect Tweets regarding the Indian election.

They would provide a set of users and hashtags they would like to follow and they wanted collection to run during the 5-week-long Indian election.

Barcelona GSE Data Science Center

Problem

We needed to collect streaming tweets and store them somewhere reasonable.

Suprisingly, there is not an off-the-shelf open source tool for doing this.

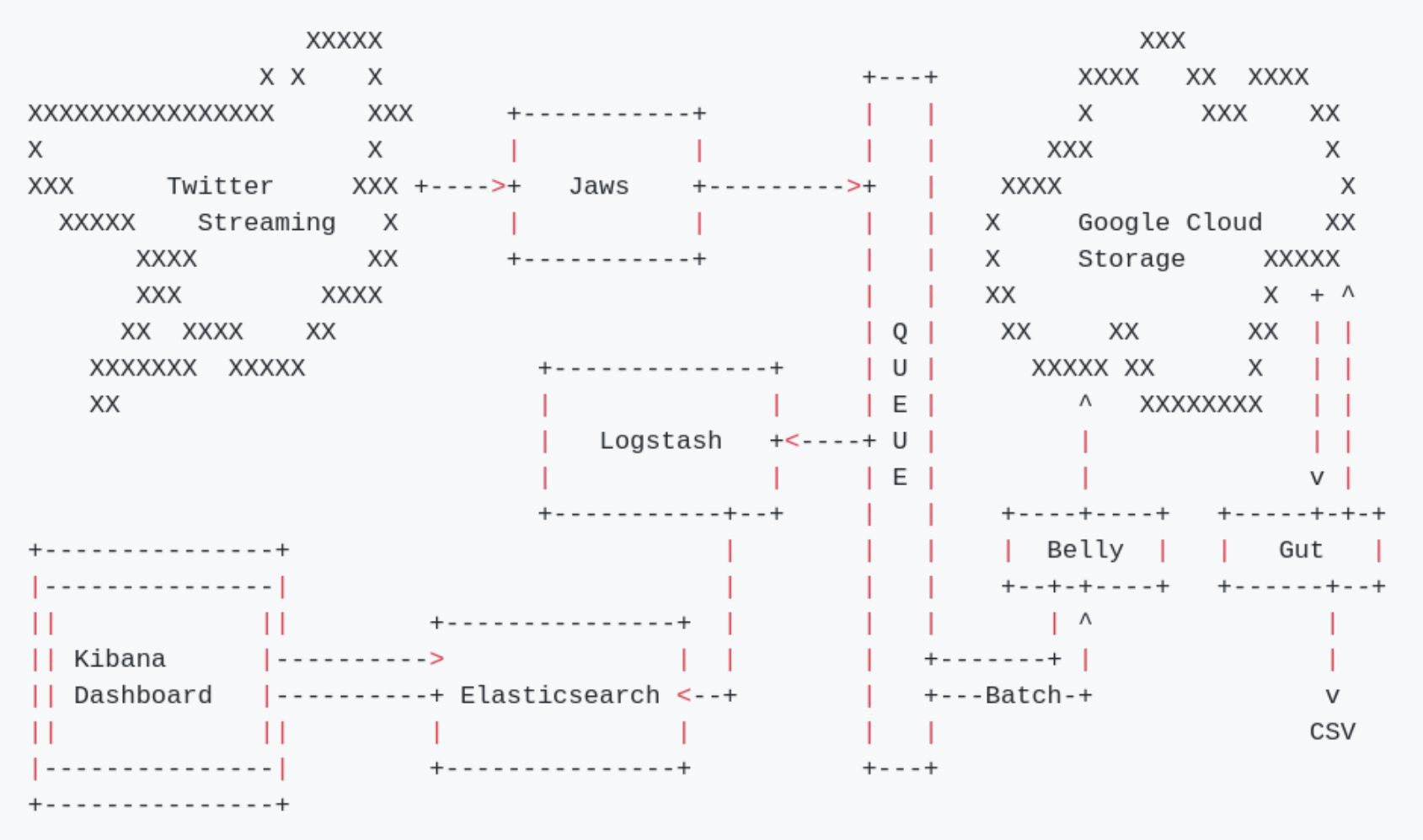
So we built one.

Challenges

1. Should work with free tier of Twitter streaming (no redundancy allowed implies a single point of failure).
2. Highest possible uptime despite that single point of failure.
3. Highly durable storage layer required.

And: it should be fully reusable!

Architecture



Barcelona GSE Data Science Center

Jaws

Jaws is our application that actually connects to the Twitter streaming API and recieves tweets.

By design, this is our single point of failure when on the free tier (Twitter doesn't allow multiple connections). Thus, it needs to be rock solid.

Barcelona GSE Data Science Center

Jaws + Clojure

It is written in Clojure and built to use Twitter's own Hosebird (Java) library.

Hosebird puts tweets on an in-memory queue. Jaws uses Clojure's go loops to create a high number of lightweight asynchronous workers on a moderately-sized thread pool. The workers read from the queue and write to our Durable Queue.

Barcelona GSE Data Science Center

Stress Testing

3 weeks spent testing Jaws on Twitters hottest topics.

Memory profiled to ensure no memory leaks under heavy load and tuned to have a small, constant memory footprint (VisualVM, G1GC vs. Parallel)

CPU usage also optimized (go loops!) to keep CPU usage down under load.

All of this also ensures that the project is usable for larger loads!

Barcelona GSE Data Science Center

Durable Queue

Used Google Cloud Pub/Sub for a completed hosted solution.

In theory, this should be the most reliable option.

In practise, our only major downtime was due to an outage in the Pub/Sub service.

In the future, maybe just use Kafka?

Barcelona GSE Data Science Center

Dashboard

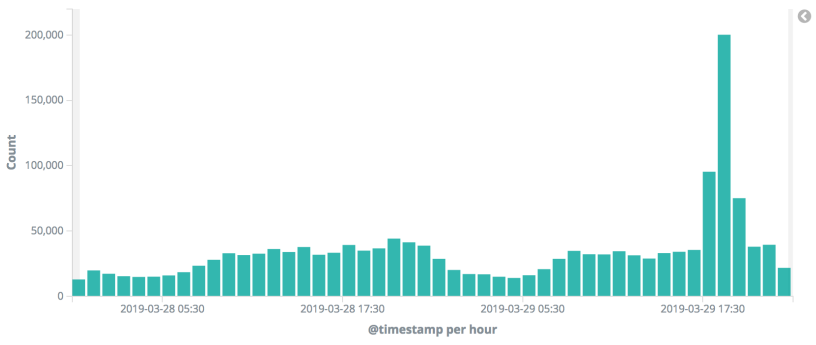
Dashboard / dash_v01_b

Full screen | Share | Clone | Edit | Auto-refresh | March 28th 2019, 00:00:00.000 to March 29th 2019, 23:11:52.681

Search... (e.g. status:200 AND extension:PHP) | Options | Refresh

Add a filter +

Tweet volume over time | Indian Standard Time (UTC+5:30)

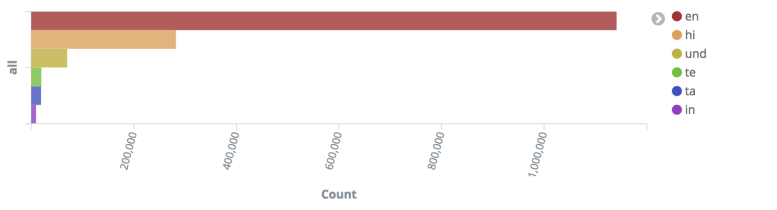


Total tweet count

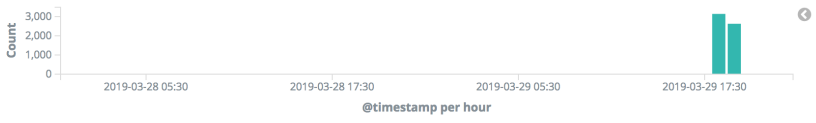
1,628,055

Tweet counts by type

Tweet counts by language



Rate Limited (1%)



50 most retweeted tweets

id_str	retweet count	screen_name	text
1111119886762364929	18,847	narendramodi	Dear friends, Over the next few days, I would be travelling across the country seeking your support for the upcoming Lok Sabha polls. Today, I would be addressing rallies in Meerut (UP), Rudrapur (Uttarakhand) and Jammu (J&K). Watch the rallies live on the NaMo App.
1110854289579216897	16,547	guardian	Congresswomen Alexandria Ocasio-Cortez gave an impassioned speech during a committee hearing in response to Republicans push-back on her climate change policy, the Green New Deal https://t.co/piPwE6ftd5
1110953416480735232	15,378	TIME	RT @TIME: The "comeback trailer," featuring group leader RM, is the first video in @BTS_twt's new musical era https://t.co/G3Cf8SC3Xs
1111287189323735040	11,061	narendramodi	बचपन से सुनते आ रहे हैं– गरीबी हटाओ, गरीबी हटाओ। लेकिन कांग्रेस के राज में गरीबी सिर्फ बढ़ी। आज देश का गरीब कह रहा है- "कांग्रेस हटाओ, गरीबी अपने-आप

Top Hashtags

hashtags	Count
ModiSpeaksToBharat	73,726
IndiaWithNaMo	61,577
MissionShakti	45,125
PhirEkBaarModiSarkar	40,679
LokSabhaElections2019	21,651
Elections2019	19,782
BJP_भगाओ_देश_बचाओ	11,490

Kibana dashboard built on ELK stack: Elasticsearch, Logstash, Kibana.

Kibana + Elasticsearch combo makes it very easy to:

1. Write fully custom, complex queries that are performant over millions of tweets.
2. Tweets kept in JSON form, (few) decisions made as to what is needed.
3. Easily add new widgets to the dashboard.
4. Drill down and use for diagnostics, see problems.

Logstash

Logstash loads tweets in realtime from queue into Elasticsearch.

Custom logstash filter (Ruby) to prepare tweet (extended tweet, retweet, etc.)

Fast and solid.

Barcelona GSE Data Science Center

Belly + Cloud Storage

Belly is a super simple service, reads a batch from the queue, writes it in original format to Cloud Storage (JSON), acks the messages when everything is written.

File storage, in this case Cloud Storage, is highly durable and scales infinitely.

No reason for JSON in the future (Parquet/Avro).

Gut

Simple Python framework to perform computations and summary statistics on tweet warehouse.

Consists of two parts:

1. Reads tweets from storage, deduplicates, writes to temp Redis (ArDB).
2. Performs aggregations across entire tweet corpus and outputs CSV's (engagement counts, network edges between users if shared mention)

Barcelona GSE Data Science Center

Kubernetes + Helm

Kubernets is used to handle container orchestration. In practise this gives us:

1. One place to capture errors, metrics, send alerts on downtime (Google's Stackdriver).
2. Supervisor to restart containers and run containers multi-availability-zone (3-zone cluster)
3. Helm allows us to package the app for single-command deploy and puts all configuration in one place. Helps make this feasible as open source project!

Open Source!

https://github.com/agriuseatstweets

Summary

Open Source framework for collected streaming Tweets: github.com/agriuseatstweets. Instant deploy with Kubernetes/Helm to multi-zone cluster. Highly durable file storage. Optimized for high throughput (Hosebird + Clojure, PubSub). Live, flexible dashboarding framework provided by Elasticsearch/Kibana.