Varnish & Kafka

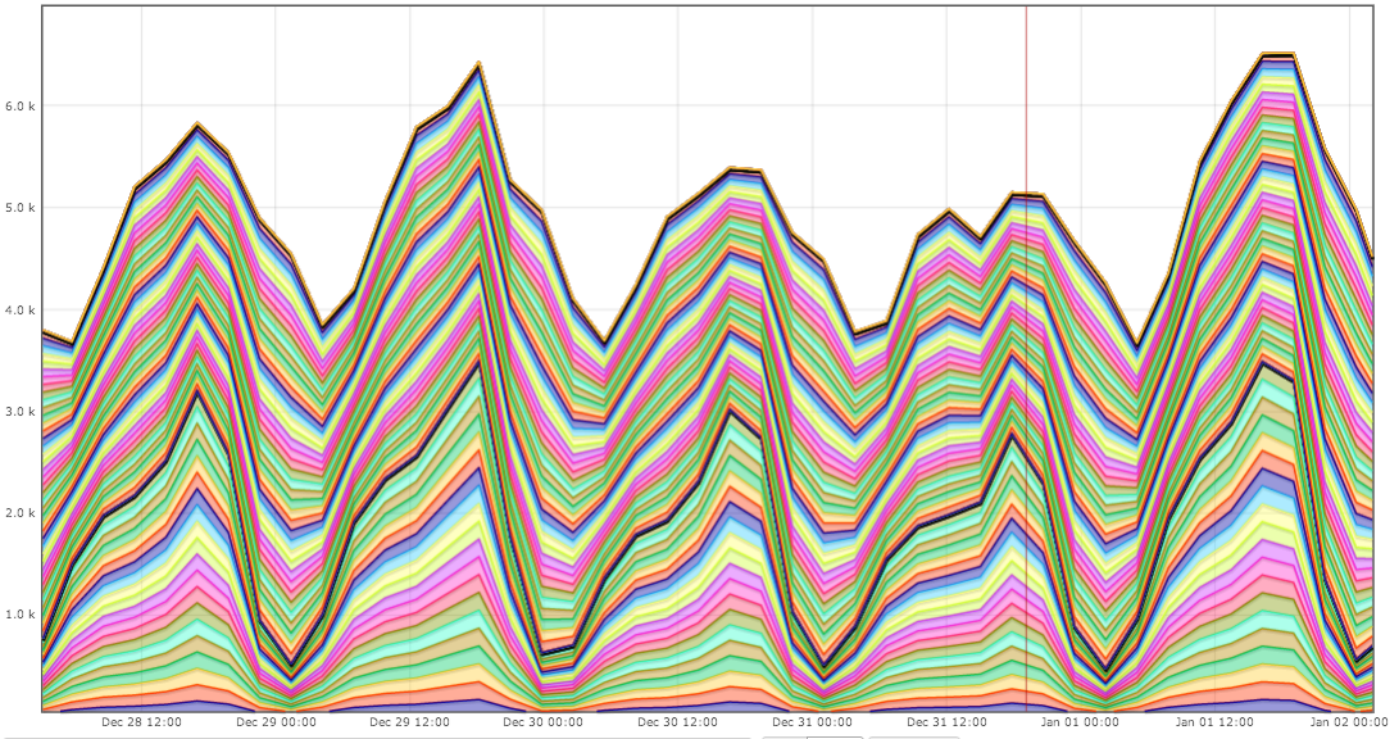
Quick overview:

- varnish is our web caching frontend. Serves web requests
- Currently, we collect sampled web request logs via custom software called udp2log
- Kafka is a reliable and scalable replacement

Varnish & Kafka

- Magnus Edenhill (author of Kafka C library) wrote varnishkafka software to format and send web request logs to Kafka
- We are now using this for all mobile webrequests

Varnish & Kafka



- Pushing 6K mobile webrequest messages / second now.

- Total webrequest rate is around 100K messages / second

- Kafka will scale way beyond this.

Hive

- Web request logs are imported into Hadoop from Kafka
  - These requests will be anonymized before this system is made available
- Hive tables are mapped onto this data hourly, making the web request logs queryable via SQL
- Eventually will be able to join with mediawiki database tables too

(Caveat! Automated imports + hive mapping is WIP not 100% production-ized yet!)

Hive Query Example

Top 10 mobile wiki projects for which Accept-Language header is different than project's language on January 7th.

```
select
  uri_host,
  substring(ltrim(accept_language), 0, 2) as language,
  count(*) as cnt
from
  webrequest_mobile
where
  year=2014 and month=01 and day=07     and
  http_status between '200' and '299'   and
  uri_host not in ('meta.m.wikimedia.org', 'commons.m.wikmedia.org') and
  accept_language != '-' and
  substring(uri_host, 0, 2) != substring(ltrim(accept_language), 0, 2)
group by
  uri_host, substring(ltrim(accept_language), 0, 2)
order by
  cnt desc
limit 10;
```

Hive Query Example

uri_host	language	cnt
id.m.wikipedia.org	en	768964
en.m.wikipedia.org	es	700844
es.m.wikipedia.org	en	494325
en.m.wikipedia.org	de	457552
en.m.wikipedia.org	zh	408308
en.m.wikipedia.org	fr	358336
ar.m.wikipedia.org	en	355334
en.m.wikipedia.org	sv	321009
en.m.wikipedia.org	nl	297289
ru.m.wikipedia.org	en	265518

Cool! Indonesian mobile wikipedia gets more mismatched language requests than any other wiki!* (on January 7th).

* Totally unscientific query run by a non-data scientist.