Scaling Marketplaces at Thumbtack

QCon SF 2017

Nate Kupp

Technical Infrastructure

Data Eng, Experimentation, Platform Infrastructure, Security, Dev Tools

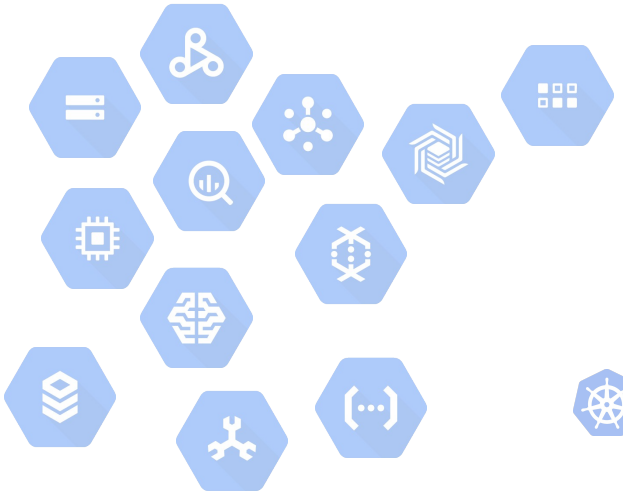
Infrastructure from early beginnings

FINE WOODWORKING
SINCE 1997 KUPP & BUILDERS
GENERAL CONTRACTING

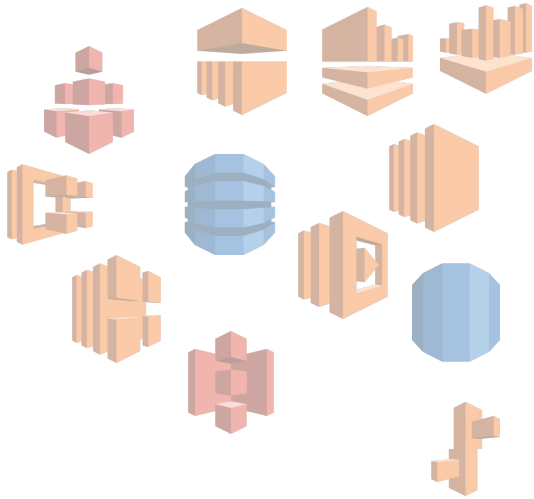
“You see that?”
my dad would say,

“That won’t last more than a few years.

They’re going to have to rip the whole thing out within a decade”.



kubernetes



You have choices. A lot of them.

snowflake

presto

scikit learn

cloudera IMPALA

HIVE

docker

TensorFlow

APACHE Spark

There's no right answer. Sorry.

ME: HARDWARE → **SOFTWARE**

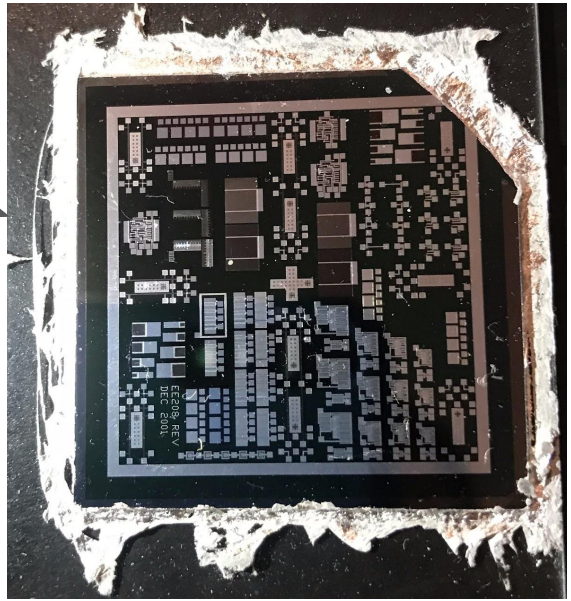
Semiconductor manufacturing research

↳ Apple iOS battery life

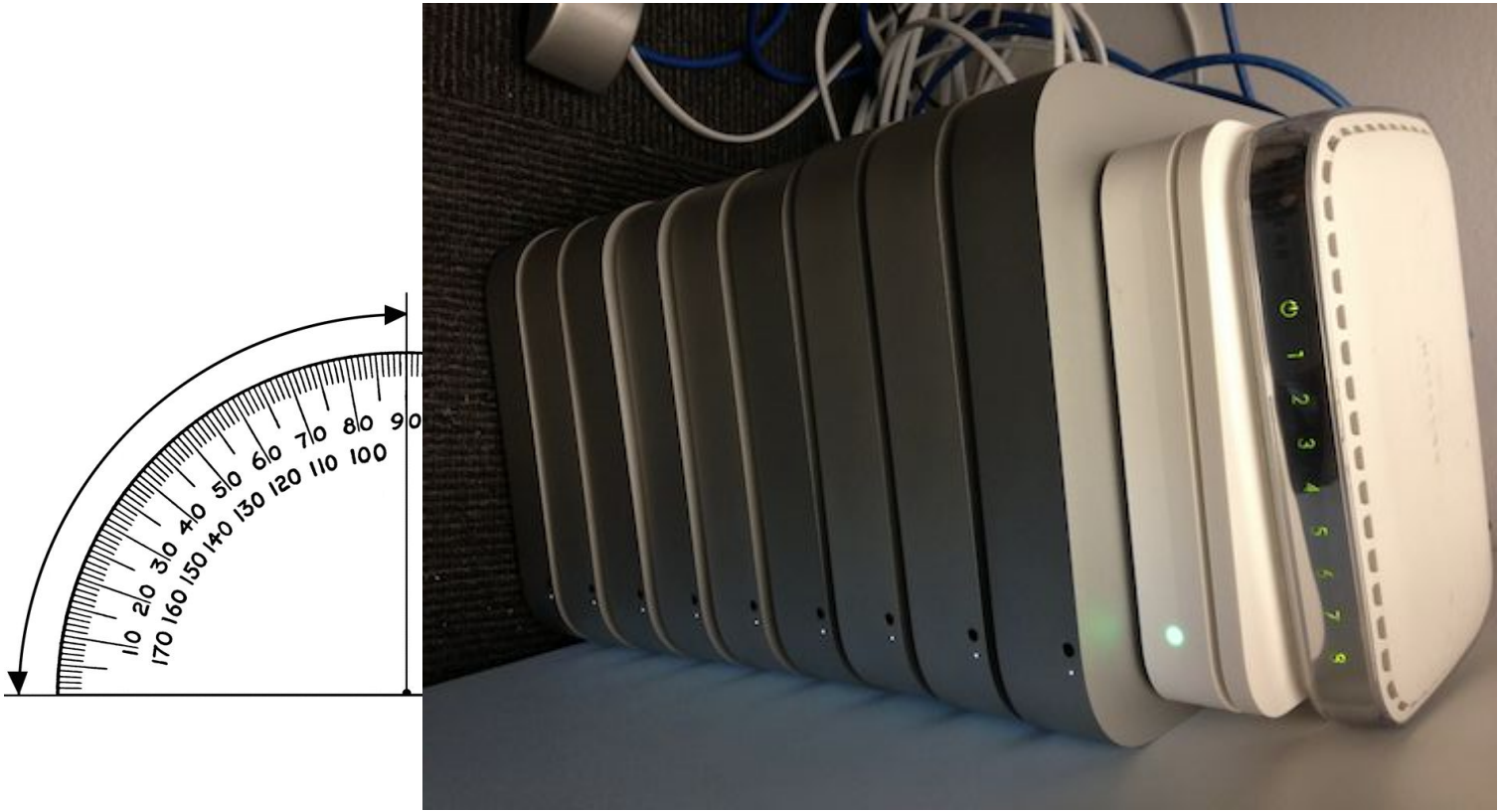
↳ Data Infrastructure @ Thumbtack

↳ Technical Infrastructure @ Thumbtack

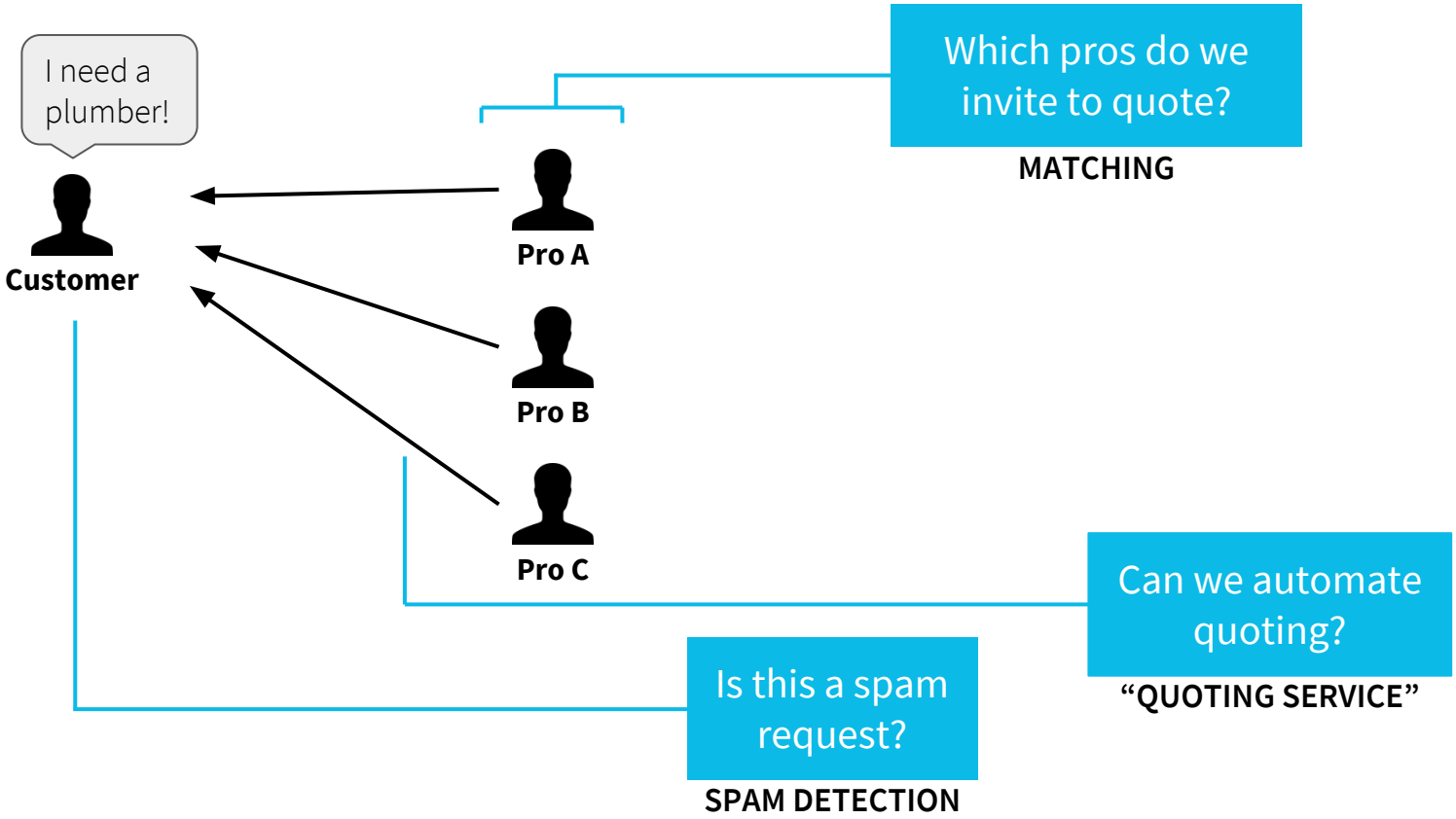
Data, core platform, A/B testing & dev tools



Yup, those are working transistors



Growth Hacking with Team Philippines

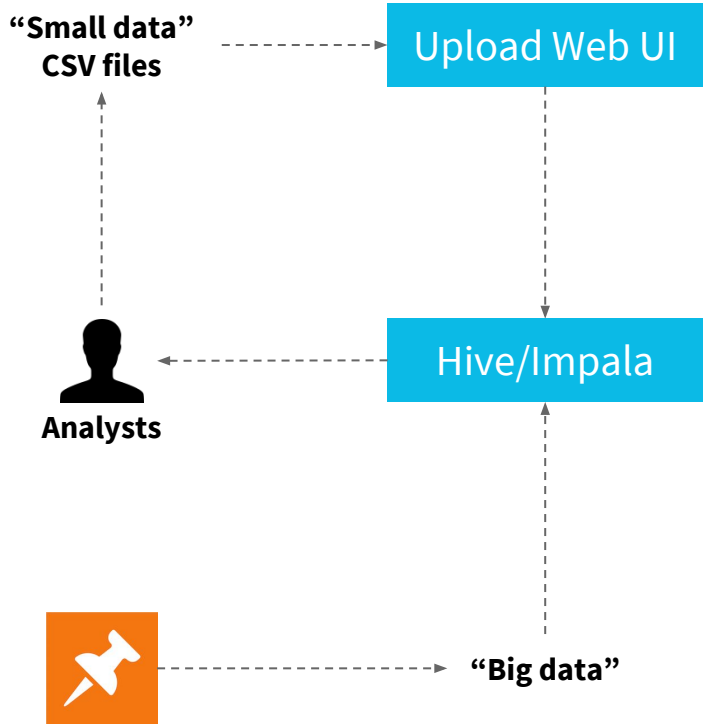


Hacky (especially in the early days) is fine, even necessary

Hacky is going to happen. With or without you.

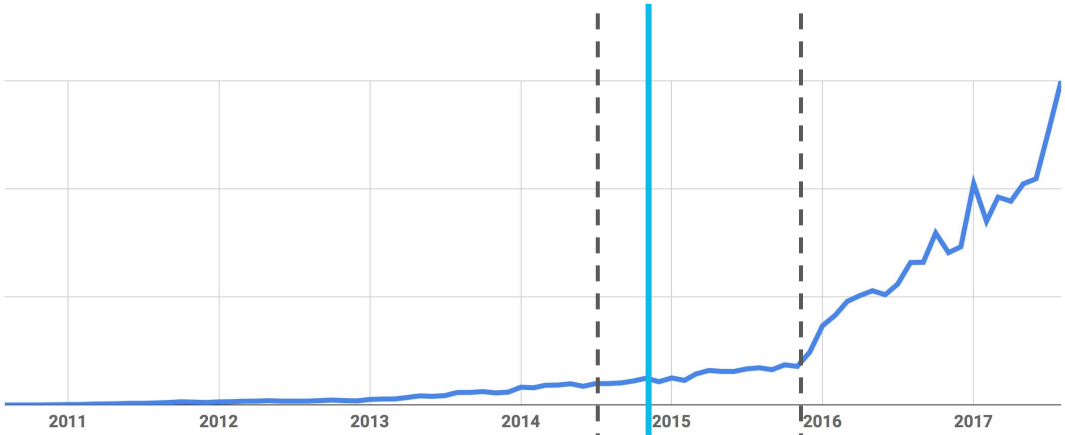
*“It has long been an axiom of mine that the **little things** are infinitely the most important.”*

- *Sherlock Holmes*



Change doesn’t happen in a vacuum.
Give people a way to get things done while you build or migrate.

Growth @ Thumbtack



2009
Thumbtack founded

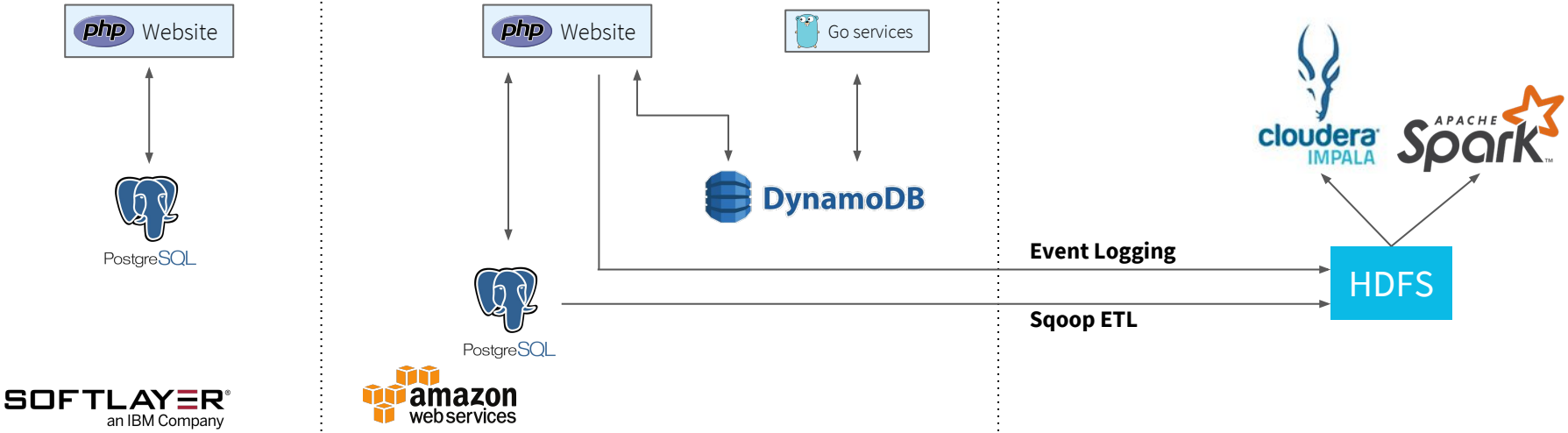
2014
Raise $130mm

Dec. 2014
I join Thumbtack

Late 2015
Starting to hit "real" scale

Now
Continued growth

The Evolution of Thumbtack's Infrastructure

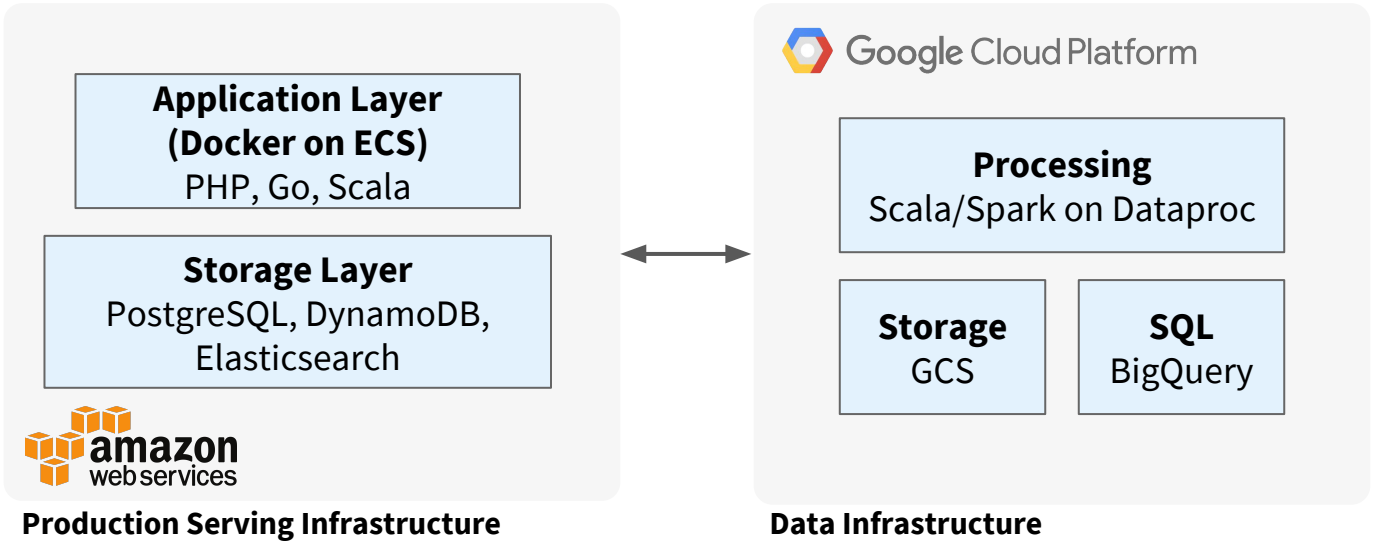


2009 - 2014
The monolith

2015
Finally on AWS! Started using DynamoDB, Golang microservices

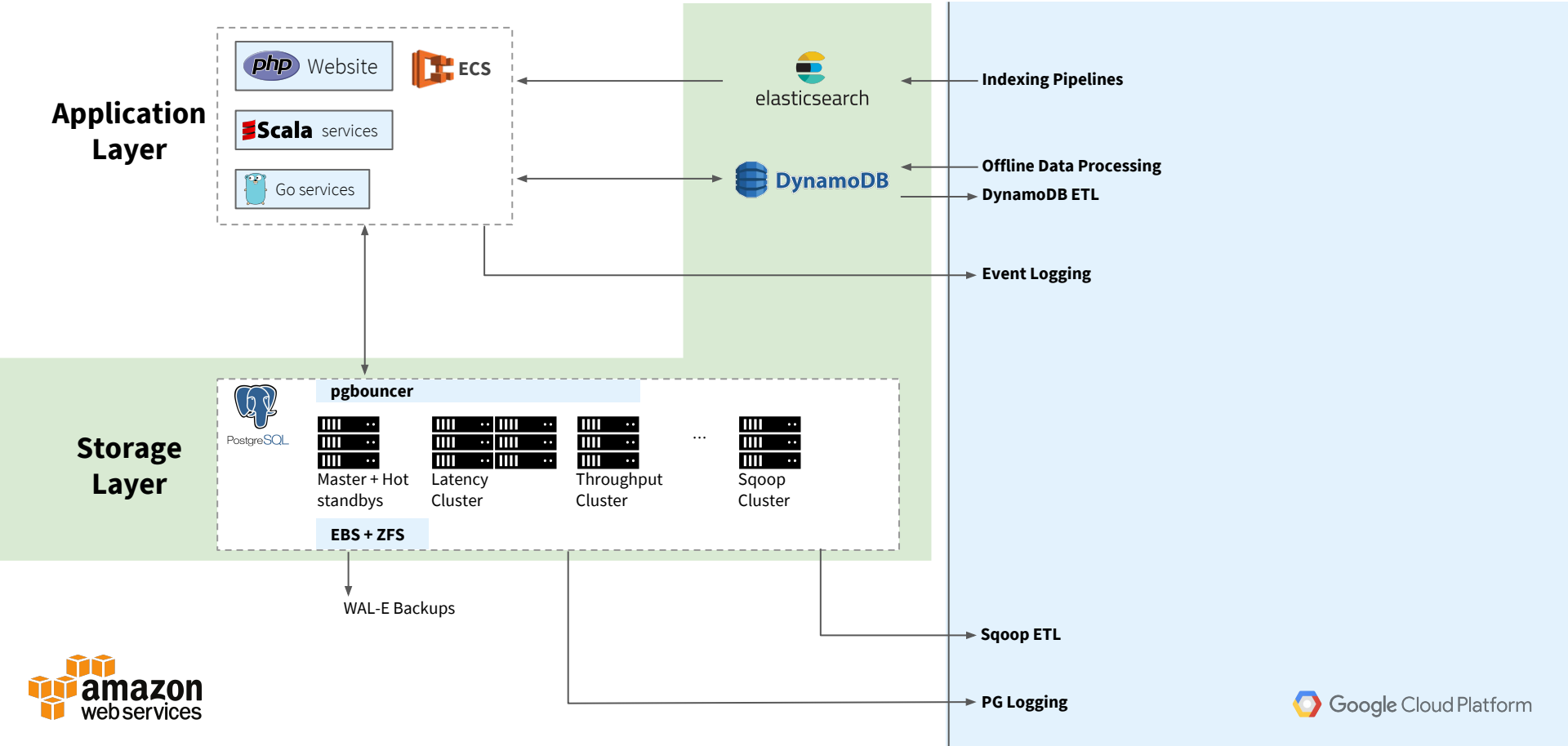
Mid-2015
Early data infrastructure

The Evolution of Thumbtack's Infrastructure

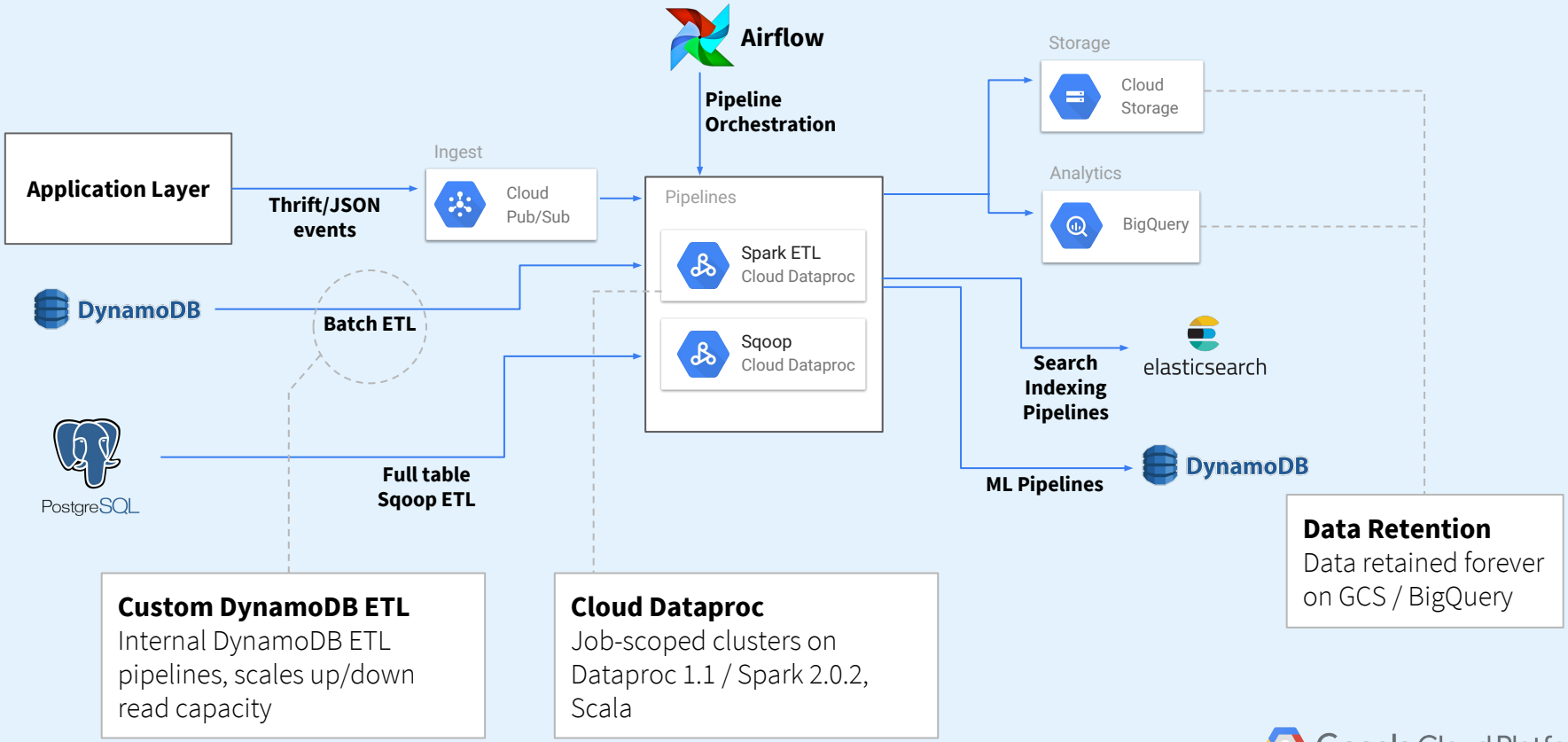


Today
Mixed clouds,
fully-containerized
infrastructure, terraform

Production Serving Infrastructure

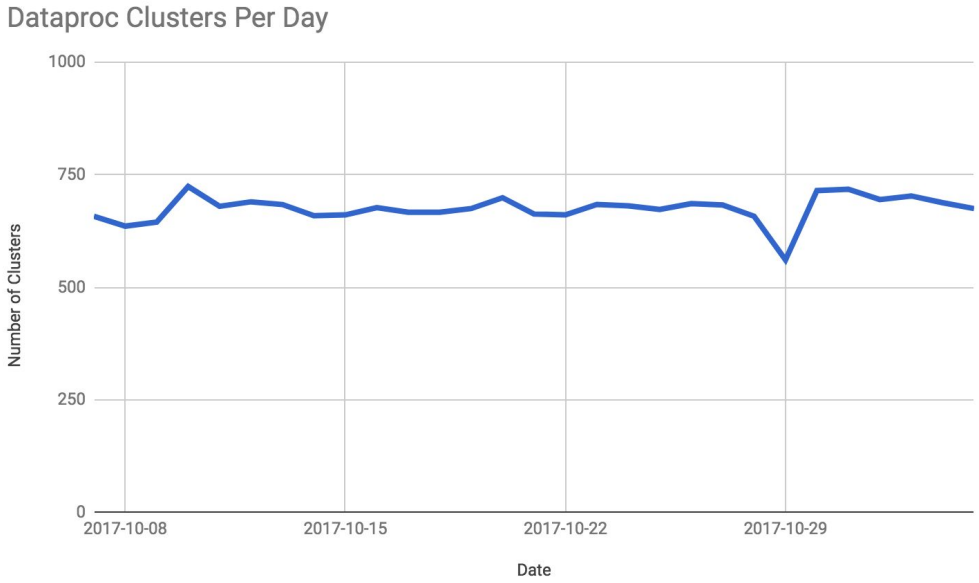


Data Infrastructure



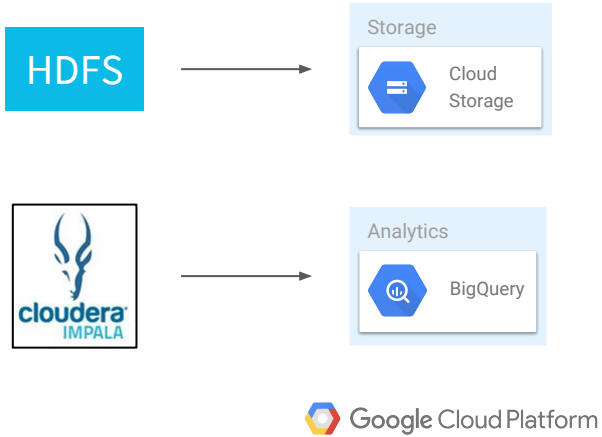
Google Cloud Platform

And it was great!



- Costs comparable to monolithic cluster
- Creating and destroying **> 600** clusters per day
- **> 12,000** BigQuery queries per workday

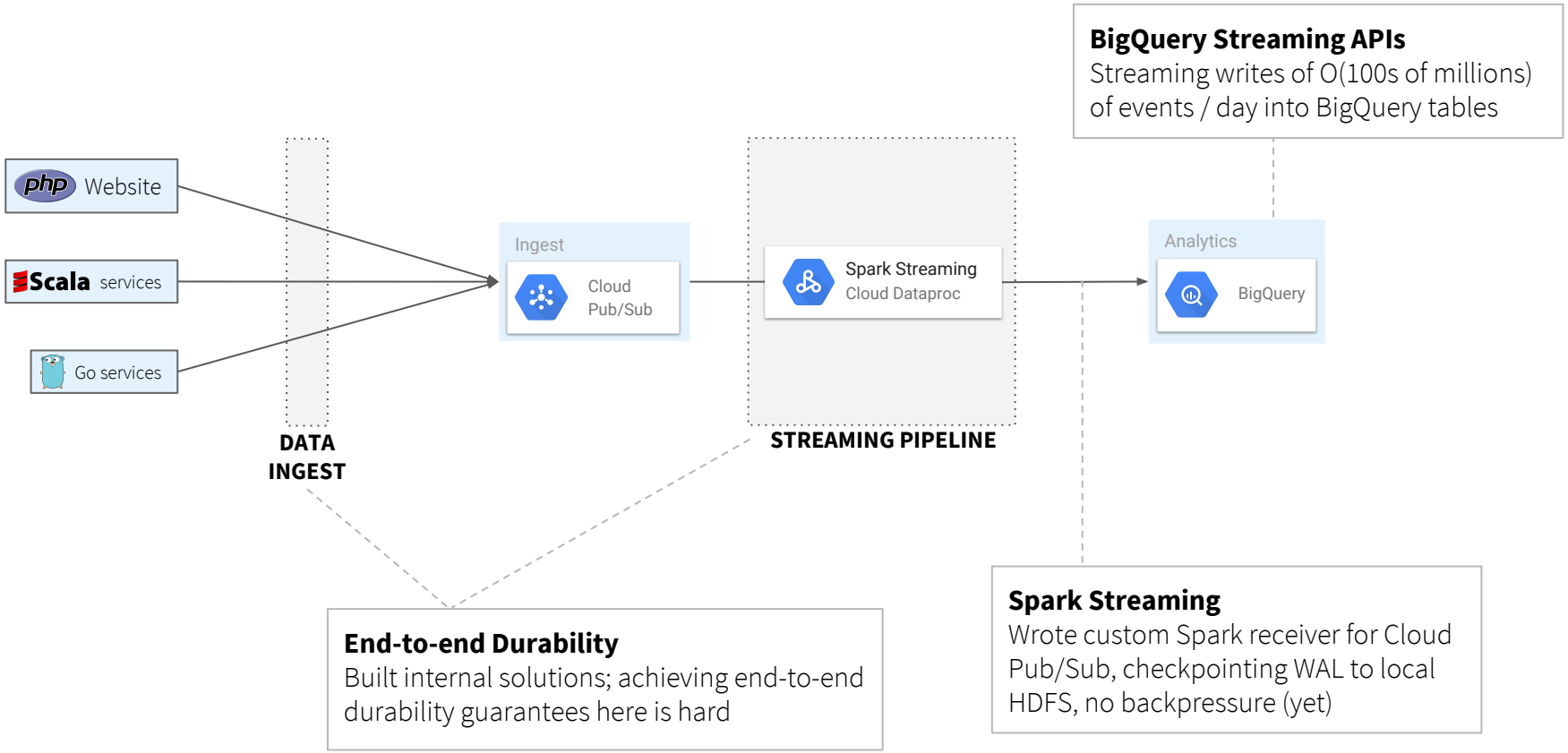
But, getting there took some *real* work



Code Changes

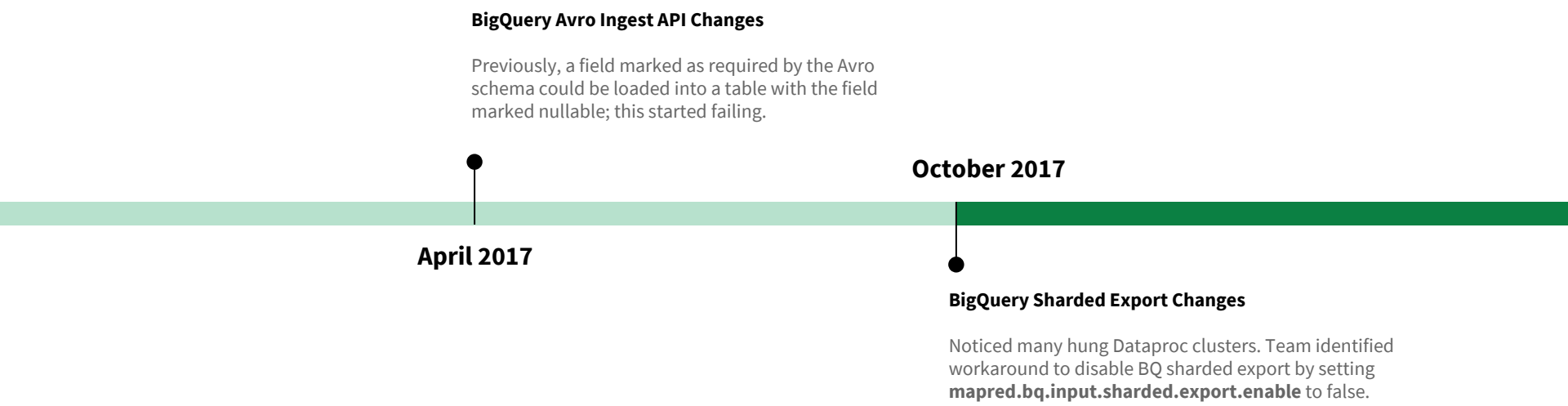
- Add retries everywhere (many 500/503 error codes)
- Rewrite HDFS utilities
- Change all Parquet to Avro
- Educate team (100s of users) on new SQL dialect and web interface

And, still needed to build interfaces to managed services



“Here be dragons”

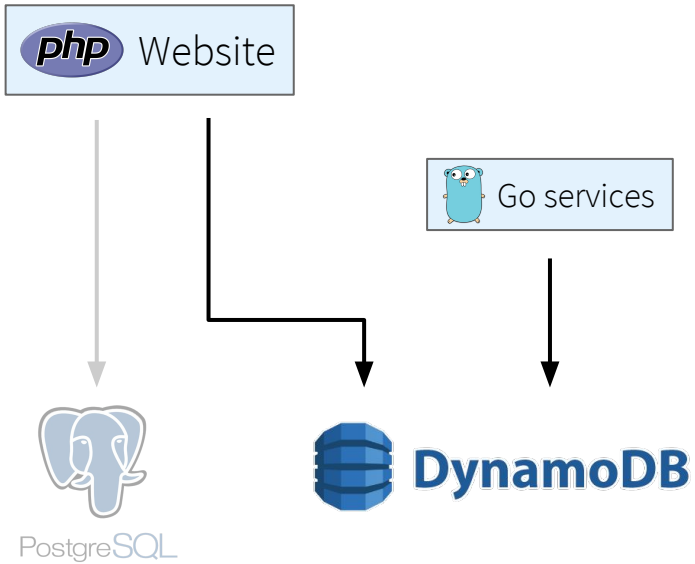
+ **Uptime:** GCP uptime has actually been great
- **Unexpected BigQuery API Changes:** Google occasionally makes production changes which break us. When these happen, we can’t really do anything until fixed by Google or a workaround is identified.



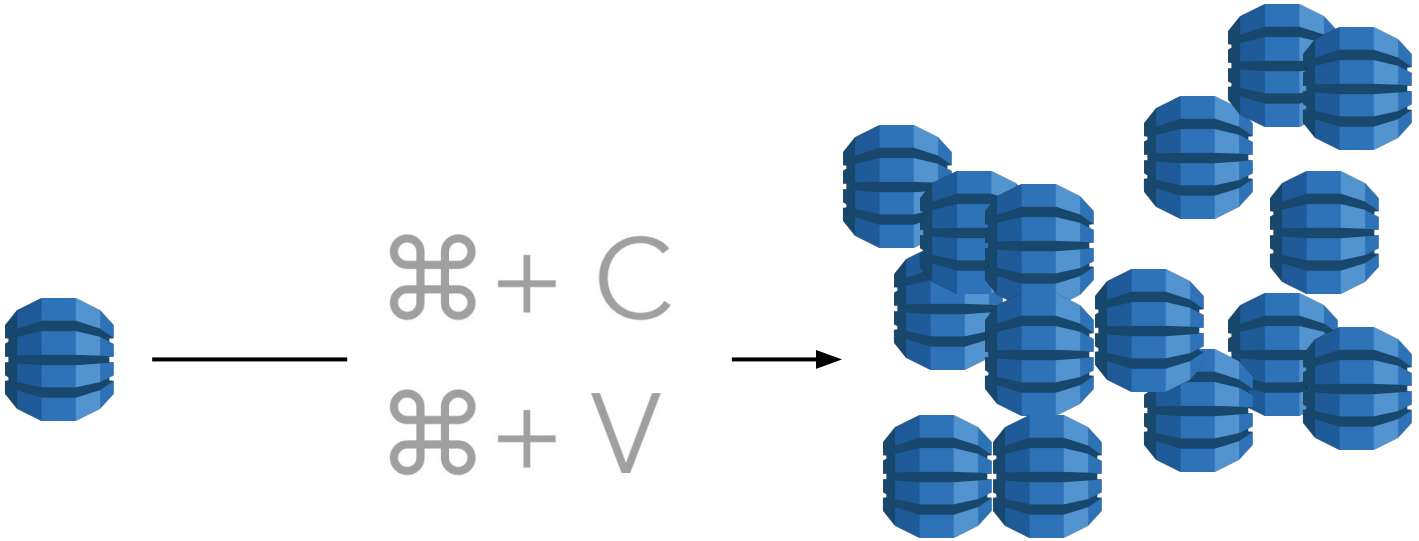
Managed services make some things easier…
But only some things.

Go in with eyes wide open.

Our first managed service: DynamoDB



DynamoDB table proliferation



GCP Migration & BigQuery

Why? A few reasons...

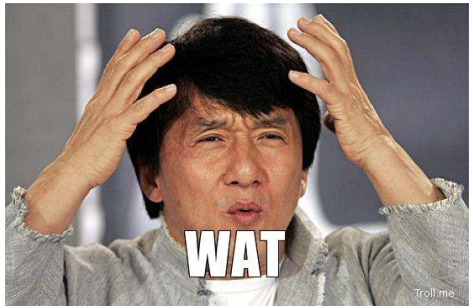
1. Fully-managed, serverless, petabyte-scale data warehouse
2. Nested/complex data types
3. Security & access controls
4. Self-serve interface to Google Sheets

Surprise dependencies

PagerDuty APP 4:59 AM

Triggered #75942: Airflow alert: <TaskInstance: gcp.data_platform.refresh.imports.a.quotes 2017-08-31 09:00:00 [failed]>

```
Exception: BigQuery job failed. Final error was:
{
  'reason': 'invalid',
  'message':"Could not parse '.' as int for field category_id
(position 0) starting at location 340 ",
  'location': '/gdrive/home/Category taxonomy - master file'
}
```



Product (EPAD) / ... / Postmortems

2017-08-31 a.* tables stale due to google sheets backed table

Last modified Sep 01, 2017

Summary

On 2017-08-30 updates to BigQuery table a.quotes failed due to issues in an underlying google sheet that was being used as reference. a.* tables data on BigQuery/GCS was stale between 20:00 PDT and 08:58 PDT (+1). The issue also caused most Mode dashboards and Prospect experiments not up-to-date during the time. At 08:50 PST the error was fixed by reorder the tab order in the underlying google sheet.

Timeline

- 07:03 PDT - rvp@ mentioned on #growth-analytics that the data in go/daily-tracking was out of date since the previous evening.
- 07:15 PDT - nate@ mentioned in #ti that a.quotes was failing to refresh due to an issue with the reading the go/category-taxonomy google sheets file.
- 08:30 PDT - andrewlam@ suggested in #growth-analytics that the error could be related to tab order in the google sheet. @Sku confirmed that this is likely the issue.
- 08:33 PDT - venky@ emailed teamsf@ to inform that a.quotes data was stale
- 08:45 PDT - sam@ mentioned in #growth-analytics that he swapped the order of tabs back to the original.
- 08:58 PDT - venky@ emailed teamsf@ to inform that the issue had been resolved, a.quotes and a.requests had been updated, and other dependencies should be refreshed at their next import schedule

Root Cause Analysis

When you make it really easy to deploy infrastructure...

*...you get **A LOT** of random infrastructure*

What does that mean for Thumbtack & Serverless?

Hacky is going to happen. With or without you.


Empower the team while limiting the wild west.

QUESTIONS?