

Trials and Tribulations of Scaling Data Engineering @BuzzFeed

BuzzFeed all about the share

- BuzzFeed creates content for the web and social platforms
- Editors look to make a connection with our audience
- People share content with which they've had a connection

TOP POST

2,386,629 VIEWS



27 Signs You Were Raised By Asian Immigrant Parents

Specifically East Asian (Chinese, Korean, Vietnamese, etc.) See these other posts for [Indians](#) and [Persians](#).

posted on Apr. 17, 2013, at 1:21 p.m.



Dao Nguyen
BuzzFeed Publisher



Kevin Tang
BuzzFeed Staff



1. Before prom, your parents had stern words for your date:



Data Infrastructure

- Part of the Infrastructure Group
- Tasked with building foundational technology to collect, build, and store data
- Team of 5
- Founded in late 2014



Data Infrastructure Core Values

- We view our major systems as products
 - Iterative development
 - Stakeholders
- Optimize for other teams' independence
- We build tools that other engineers build on top of

What We'll Talk About

- The Journey
 - Where we started: Wild West
 - What Should We Build First?
 - Making the impossible possible
 - The world changes
- Takeaways
- The Future

Let's start our journey

Where We Started: Wild West!

- One big tech stack
- One tech team, no dedicated data engineers
- Data Scientists made it work
 - Own databases
 - Lots of unchecked code
 - Lots of manual work



Where We Started: Wild West!

- Data is hard to access
- Collection systems have >15 minute latency and are unstable



Where We Started: Wild West!

- Data is hard to access
- Collection systems have >15 minute latency and are unstable
- Small number of people understand and can access data
- Engineering lacked autonomy and agency



What should we build first?

- Patch existing systems
- Build a data warehouse to combine and make data accessible
- Tools to query existing data stores

What should we build first?



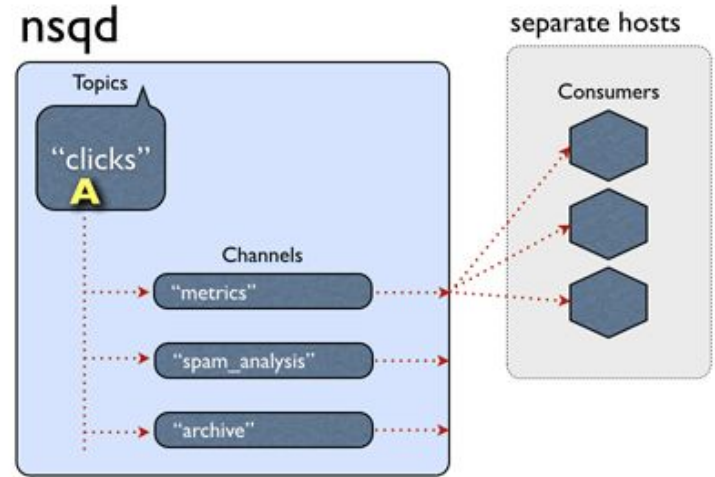
What should we build first?

- Fixed data collection first
 - Foundational
 - Focus on high quality, dependable data. The rest can come later

Making the impossible possible

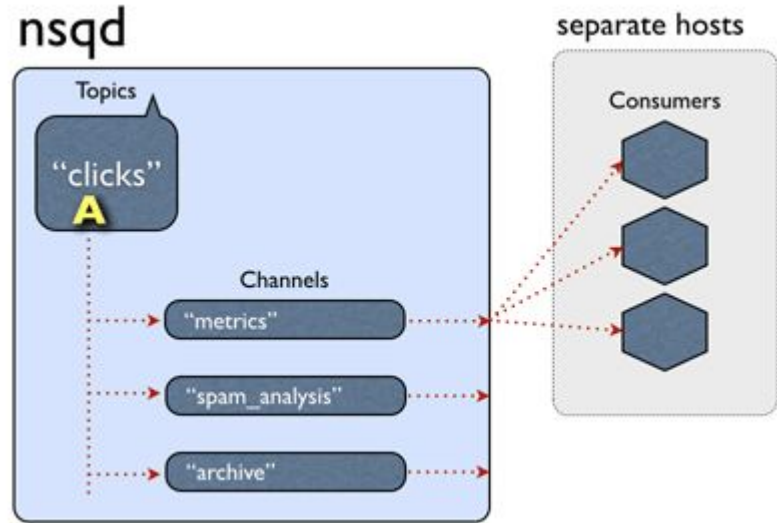
Making the Impossible Possible: NSQ

- Distributed, scalable message queue
- Producers send messages to a topic
- Consumers subscribe to a topic, which creates a channel
- Data published to a topic goes to all consumers



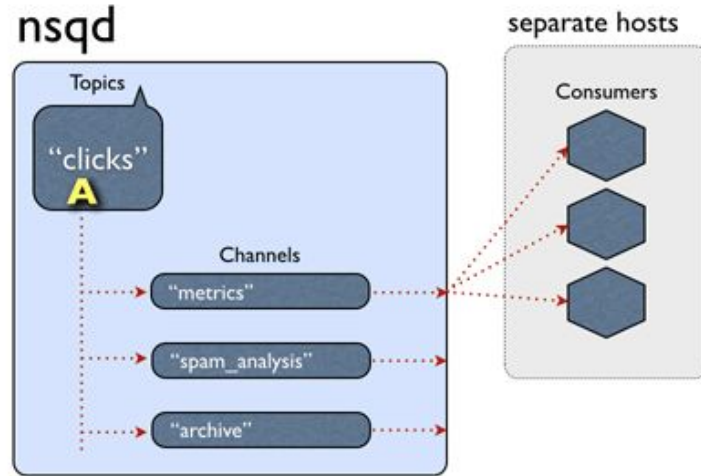
Making the Impossible Possible: NSQ

- Every message archived to s3
 - Gzip'd archive of all messages



Making the Impossible Possible: NSQ

- Simple Python API on top
- JSON over the wire



Making the Impossible Possible

- We now have access to realtime data!
- Cool things are now possible
 - Trending posts by region
 - Realtime dashboards

TRENDING
136,811 VIEWS



PROMOTED BY
FALLING WATER
TONIGHT 10/9c | u-a

Dashbird

Region: USA

Period: Launch



You Guys, Somebody Supposedly Invented A Dress ...

juliegerstein

Published 9 hours ago

1.8X

SOCIAL LIFT

85.7K

VIRAL VIEWS

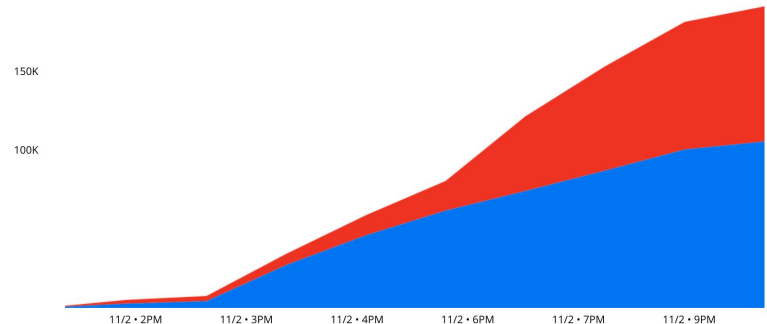
106K

SEED VIEWS

191K

TOTAL VIEWS

Viral Views  Seed Views 



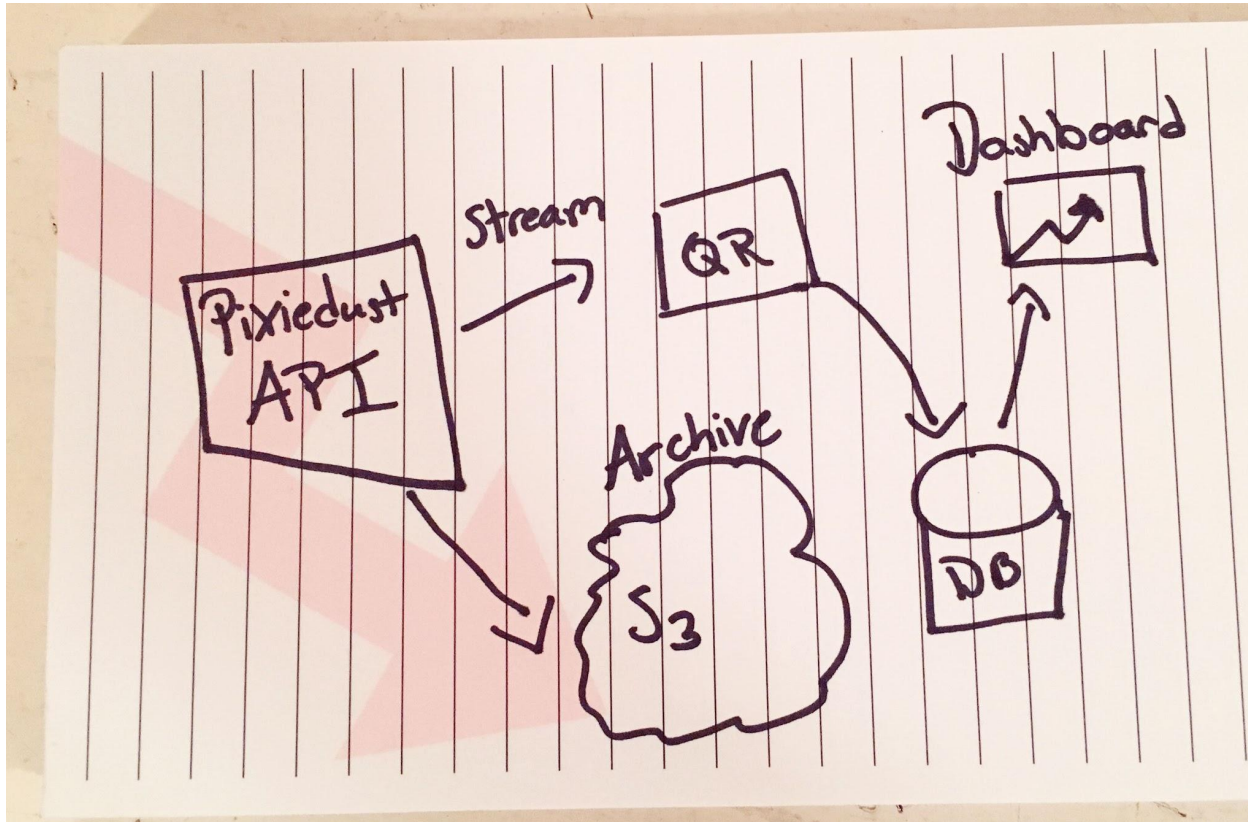
Making the Impossible Possible

- Calling our archive of messages the source of truth for all tracking simplified follow-on architecture
- All batch process done off this data set (and not a derived dataset)
- Adds resilience to downstream systems
 - Can rebuild/repair data

Making the Impossible Possible

- Collect everything. Store everything.
- Flexibility to change priorities later and calculate for all-time

Data Ecosystem circa 2015



Making the Impossible Possible: Phase 2

- Foundations are laid
- We turn our focus to building tools

And Then, The World Changed

The World Changes

- Video is now huge
- People consuming content off site
- Proliferation of platforms
 - Several buzzfeed apps
 - Apple watch and apple tv app
 - Facebook and Youtube
 - Snapchat is a thing



The World Changes

- Is share still relevant for video?



The World Changes and it's ok

- No fundamental changes to our data infrastructure needed
 - New data sources are just another producer
 - Additional apps have access to data in realtime



The World Changes and it's ok

- We experiment and learn
- Tasty
 - 74 million likes



The Try Guys Go Bald

BuzzFeedVideo ✓

1 month ago • 5,590,328 views


I look like a respectable dad." Check out
<http://bit.ly/YTbuzzfeedvideo> Did you ki

CC

The World Changes and it's ok


- Yes, shares still count

Like · Reply · 3 hrs

 **Victoria Harris** Stephanie Hull. We have to make this. Screw that diet. Screw it all. We have tooooooooo. This isn't one of those "hey this looks good" tags this is a "WE FREAKING HAVE TO!!!!" tag!!!!!! 🤔🤔😊😊

Like · Reply · 1 hr

👉 1 Reply

 **Domi Moreno** Too bad I am not making the thabksgiving dinner this year boys....look what could have been the dessert **Sujan Badal Joel T Pj Mohammed Al-Suraih**

Like · Reply · 🍌 1 · 5 hrs

👉 2 Replies · 4 mins

 **Tasty**
8 hrs · 🌐

Caramel Apple Upside-Down Cake (via [Bien Tasty](#))
FULL RECIPE: <http://bzfd.it/2eV8rvO>



0:09 🔊 ⚙️ ↗️

Takeaways

Simple, Composable Systems

- Establish patterns that work, repeat
 - NSQ + Queue Reader = realtime systems
 - EMR + Spark = bulk data processing
 - Big Query or Redshift = exploratory access

Optimize for Independence

- Data Engineering builds tools that other teams can use
 - Ops bits abstracted away
 - Service level monitoring for free
 - Layers of alerts

Optimize for Independence

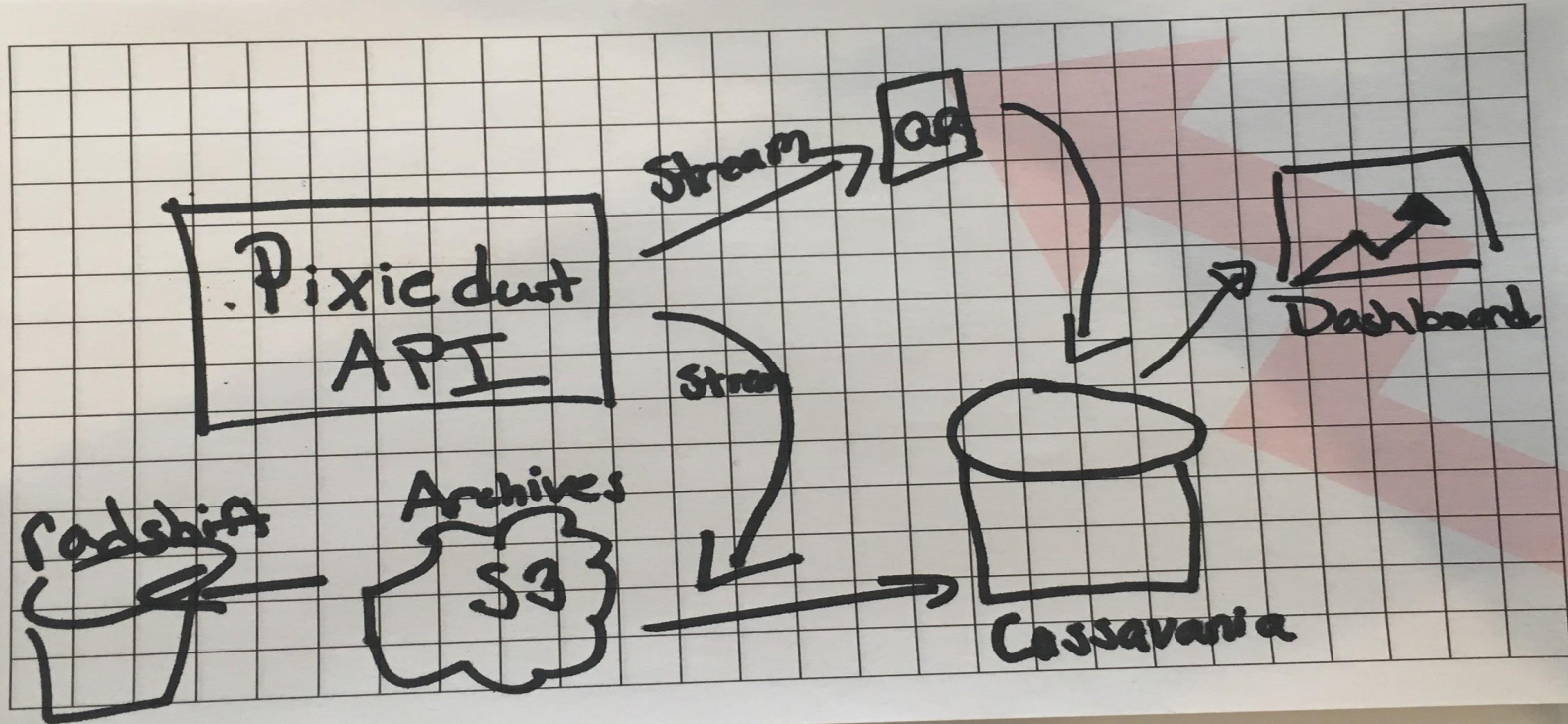
- Aggressively track down bottlenecks and eliminate
 - Slow process is as damaging as slow code
- Understand stakeholder's perspectives
 - Talk constantly with data scientists to understand their needs and approach

Simple, Composable Systems

- Small code footprint
- Lots of small services



Data Ecosystem circa 2016



Simple, Composable Systems

- Intentionally decoupled services
 - Can spin up something new
 - Replace something old in parallel
- Horizontally scalable

Simple, Composable Systems

- Components fit together but aren't explicitly tied together
 - Allow experimentation
 - Supports new platforms and formats
- Systems build on top of one another
- Teams can leverage what others have done



Simple, Composable Systems

- Data Scientists write production code
- For example, a data scientist can use app data to send data quality alerts
 - Only have to write the logic to determine when to send an alert
 - Collection has been taken care of by another team and broadcast over NSQ

stats-monitoring BOT Yesterday, 2016-10-30, there were 5,182,775 in Radshift and 1,824,775 in Google Analytics. The Radshift:GA ratio was 2.8. Summary: 25 ad users were analyzed

The Future

- Provide stronger guarantees for consumers of our streams
 - Validation at the edges w/ schemas
- Multi-tenancy
 - AWS and Google Cloud
- Abstracted data access
- Better monitoring



Ashley Miller

@csprite