# SCALING A DATA PIPELINE:
## MYSTERY TO MASTERY

Dan Goldin

@dangoldin

triplelift

# AGENDA

Introduction

AdTech and Data

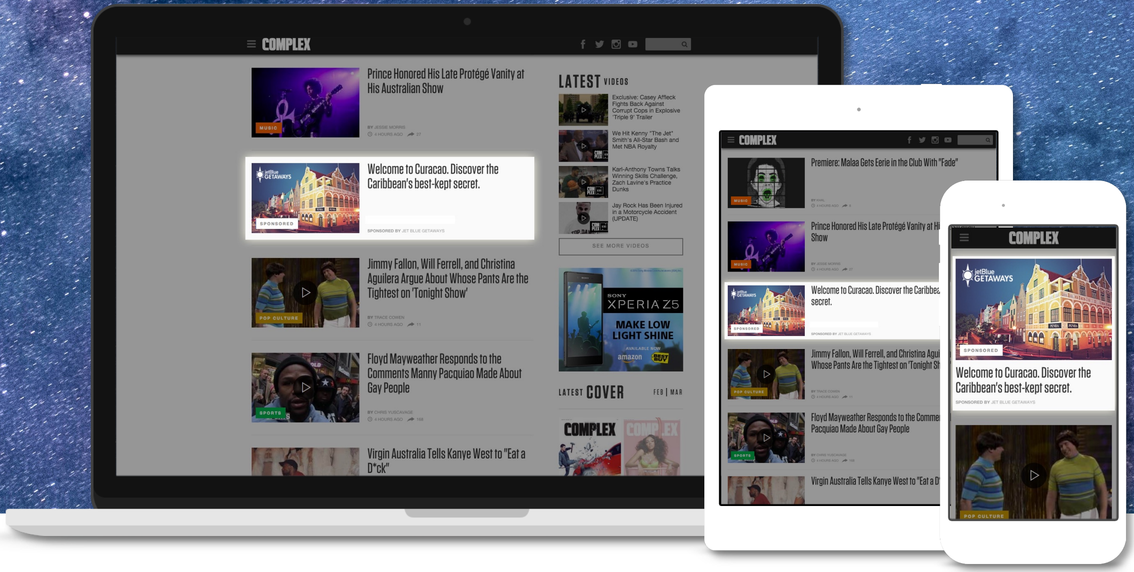
The Evolution

Current State

Lessons Learned

Q&A

# INTRODUCTION



Prince Honored His Late Protégé Vanity at His Australian Show

Welcome to Curacao. Discover the Caribbean's best-kept secret.

SPONSORED BY JET BLUE GETAWAYS

Jimmy Fallon, Will Ferrell, and Christina Aguilera Argue About Whose Pants Are the Tightest on 'Tonight Show'

Floyd Mayweather Responds to the Comments Manny Pacquiao Made About Gay People

Virgin Australia Tells Kanye West to "Eat a D*ck"

LATEST VIDEOS

Exclusive: Casey Affleck Fights Back Against Corrupt Cops in Explosive 'Triple 9' Trailer

We Hit Kenny "The Jet" Smith's All-Star Bash and Met NBA Royalty

Karl-Anthony Towns Talks Winning Skills Challenge, Zach Lavine's Practice Dunks

Jay Rock Has Been Injured in a Motorcycle Accident (UPDATE)

Premiere: Malaa Gets Eerie in the Club With "Fade"

## SIMPLE

Render brand's assets to match the unique look and feel of the publisher

## SCALABLE

Bringing scale to high performing consumer friendly formats

## EFFECTIVE

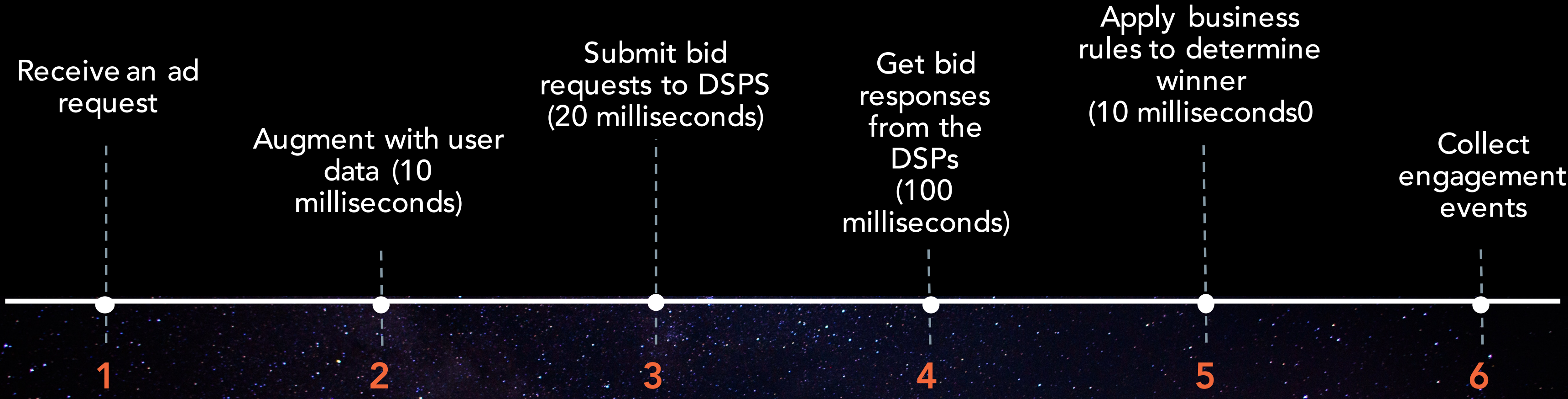
Integrations into the world's largest DSPs – Google, The Trade Desk, Turn, MediaMath, AppNexus and more

# DATA & ADTECH

High volume across many dimensions
that needs to be handled in real time

# VOLUME

# REAL TIME BIDDING AUCTION TIMELINE



1 Receive an ad request

2 Augment with user data (10 milliseconds)

3 Submit bid requests to DSPS (20 milliseconds)

4 Get bid responses from the DSPs (100 milliseconds)

5 Apply business rules to determine winner (10 milliseconds0

6 Collect engagement events

# TRIPLELIFT TODAY

**25K+**
Auctions/sec

**2B+**
Auctions/day

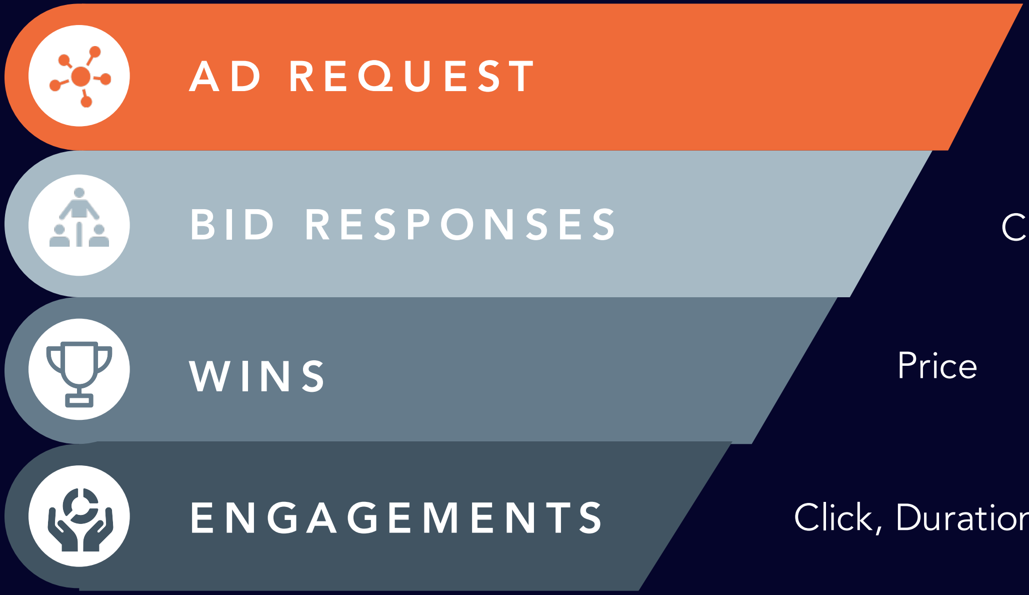
**~300**
Bids/auction

**~600B**
Bids/day

**~21B**
Events/day

# DIMENSIONS

# LOTS OF DIMENSIONS



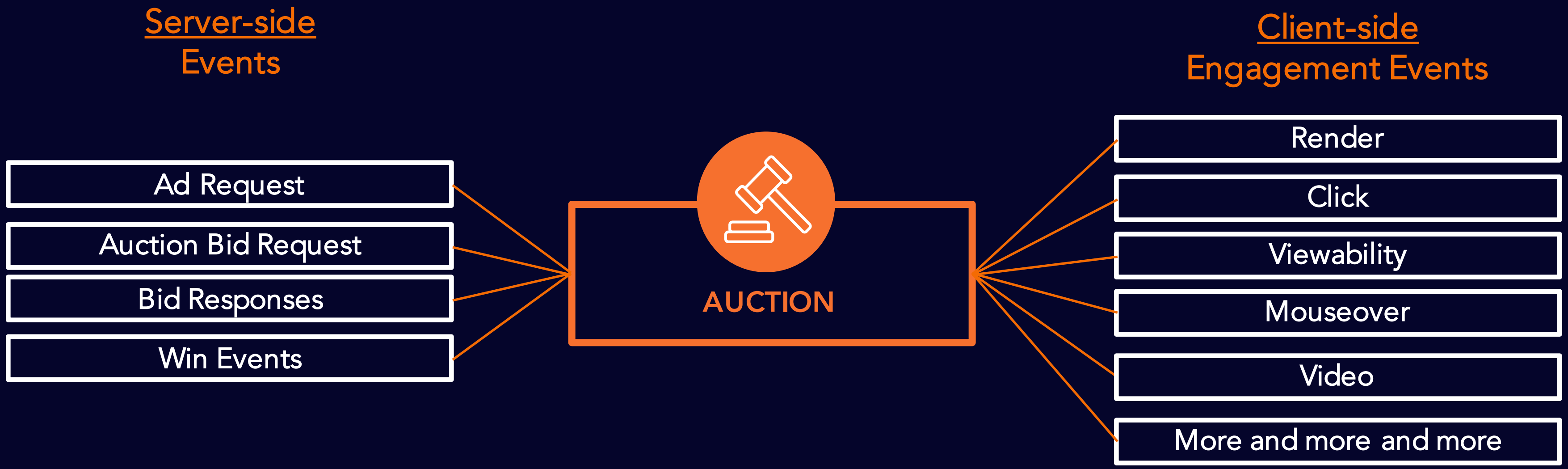
AD REQUEST — Browser, Device, OS, Geography, Domain, Time

BID RESPONSES — Creative, Format, Brand Buyer, Bid Amount

WINS — Price

ENGAGEMENTS — Click, Duration, Render, View

# AUCTION EVENTS



Server-side Events

Ad Request

Auction Bid Request

Bid Responses

Win Events

AUCTION

Client-side Engagement Events

Render

Click

Viewability

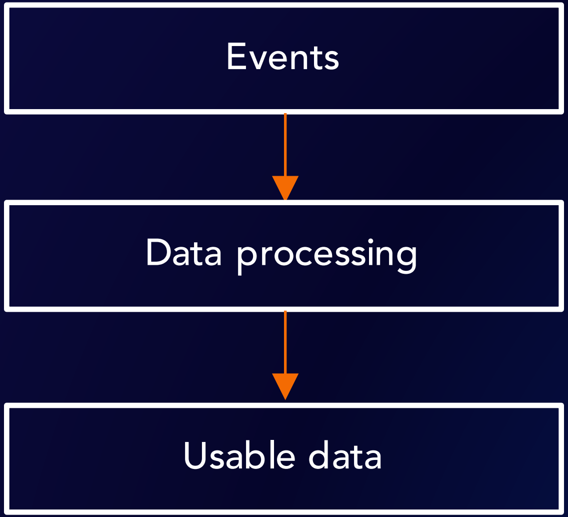
Mouseover

Video

More and more and more

# THE EVOLUTION

# SO WHAT'S A DATA PIPELINE?

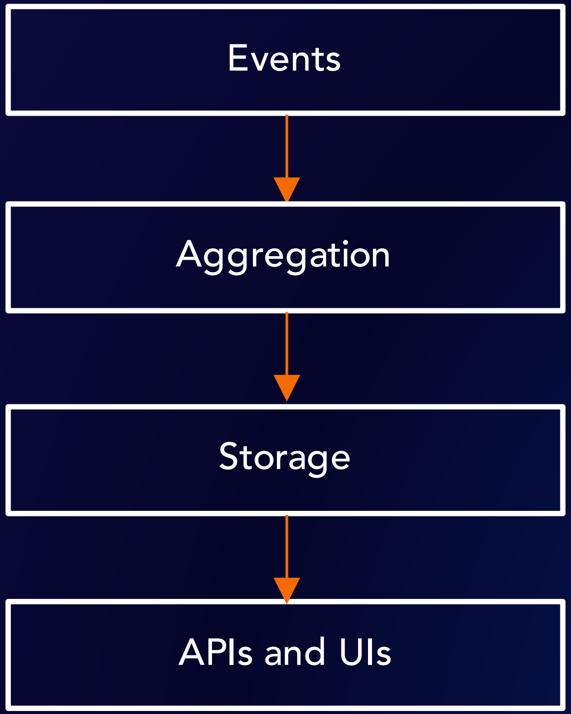


Events

Data processing

Usable data

# SO WHAT'S A DATA PIPELINE?



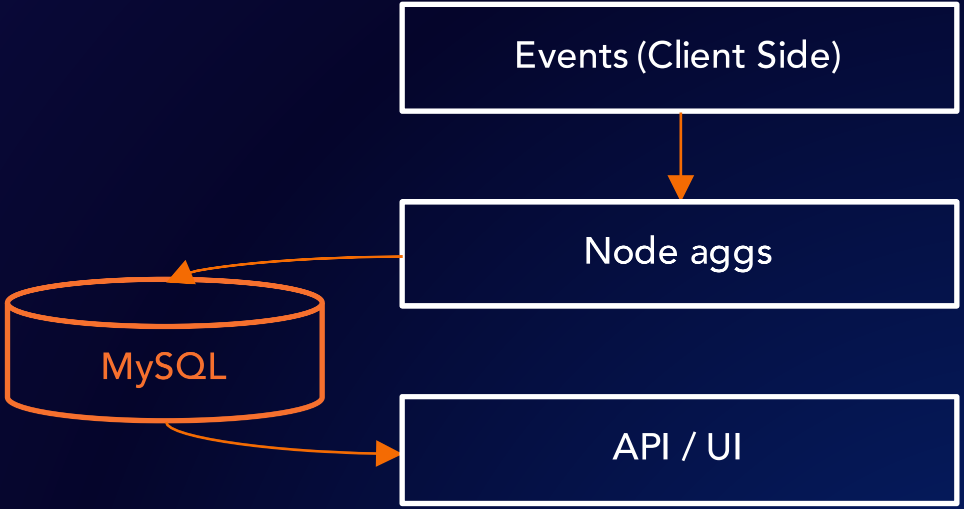
# v1: Sad but true

Implementation highlights

- Variable sample rate
- Keep a running sum in memory and write to MySQL every few minutes

Challenges

- Constant open connection to DB
- Tables became large and unwieldy
- Difficult to slice and dice sampled data
- Easy to lose data



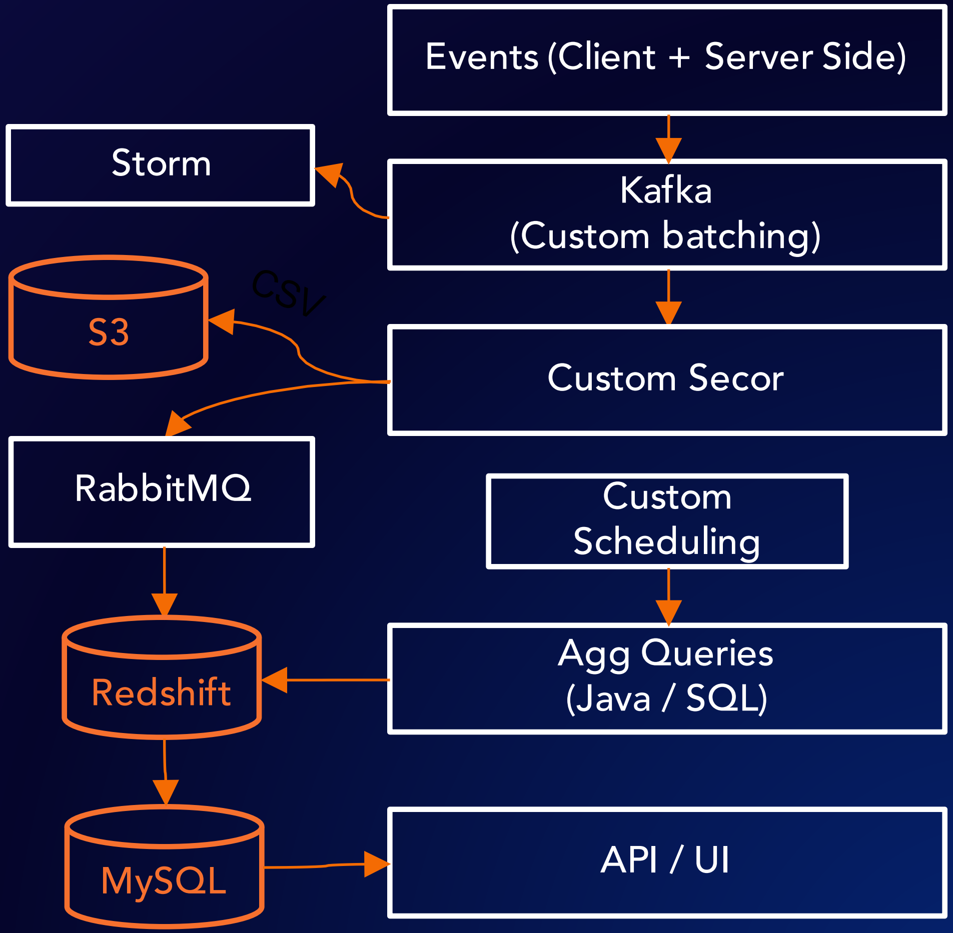
# v2: Kafka, Secor, and Redshift

Implementation highlights

- Collect every event in Kafka
- Upload to S3 and load into Redshift
- All jobs done through Redshift queries
- Storm to handle real time pacing

Challenges

- Dependencies tough to manage
- Couldn't do everything via SQL queries
- Redshift became expensive



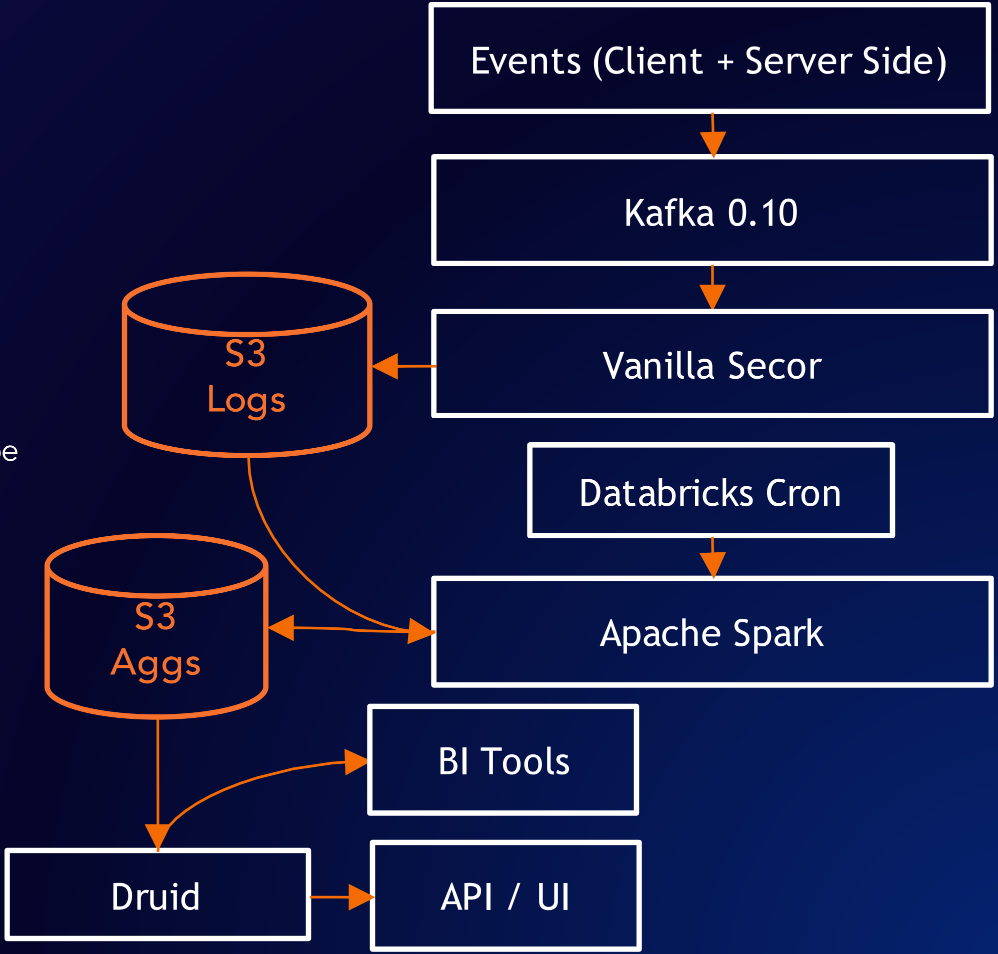
# v3: Hello Spark; Hello Druid

Implementation highlights

- Kafka 0.10
- Failed attempt at Spark Streaming
- Spark was a big improvement
  - Cheaper & more scalable than Redshift
  - More advanced query logic
- Druid also helped
  - Trivial to scale to 100s of metrics and dimensions
  - Replaced a dozen tables with a single cube
  - Improved query times

Challenges

- More tech to maintain
- Scheduling still a challenge
- More complex development process



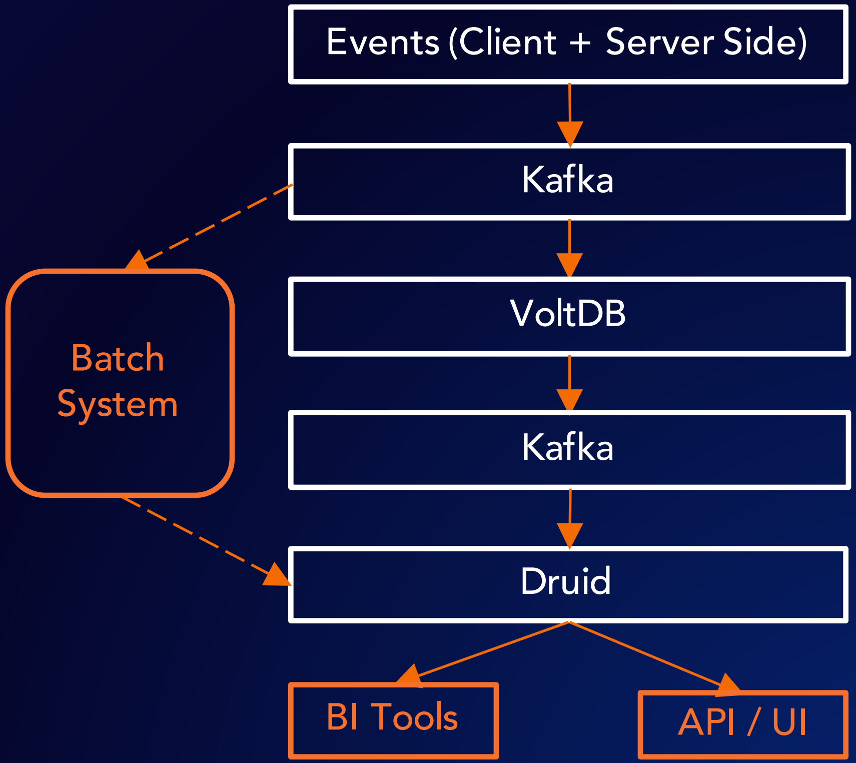
# v4: Lambda, the ultimate?

Implementation highlights

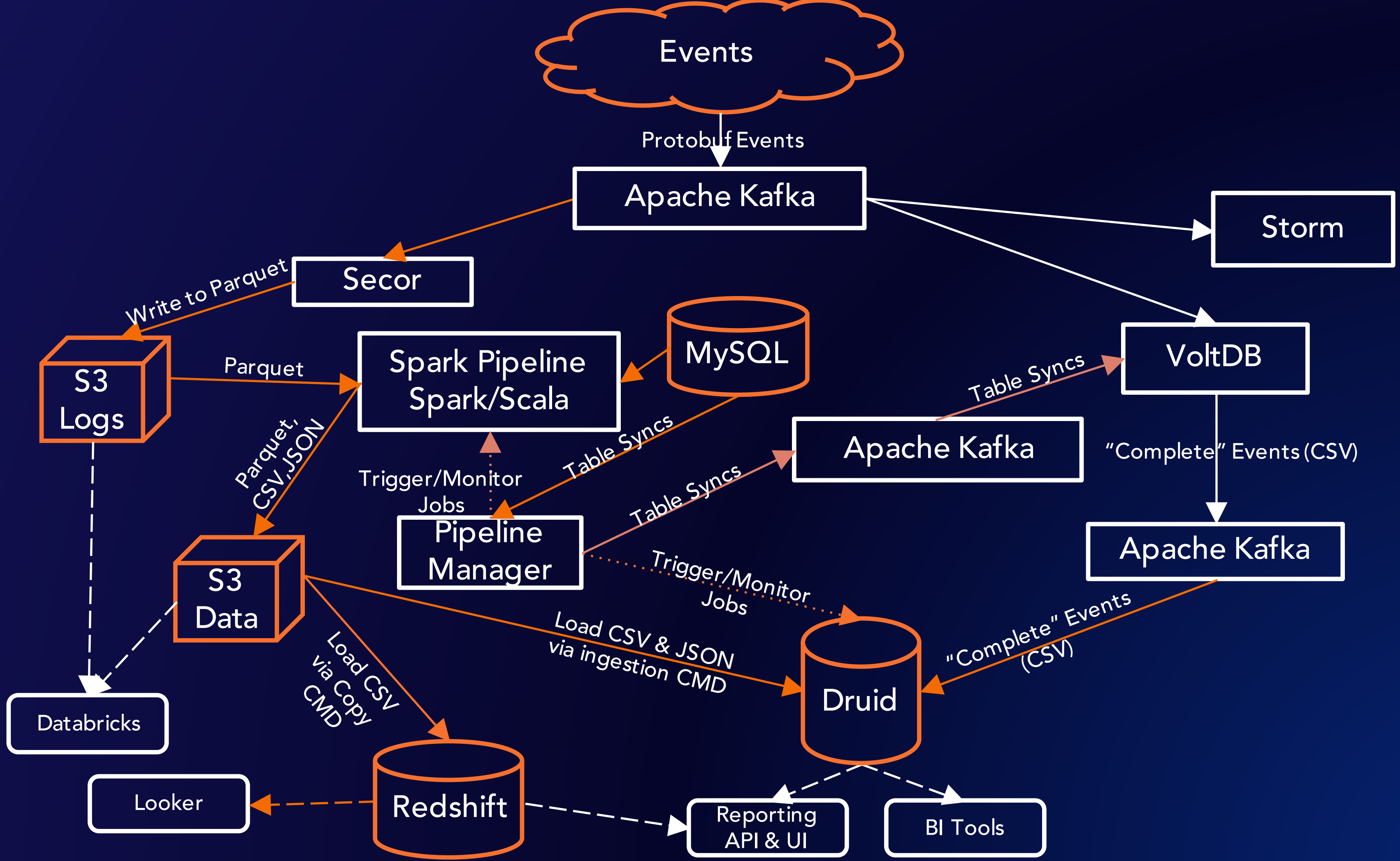
- Introduced VoltDB
- Feeds back into our Druid cluster
- Delays in batch jobs masked

Challenges

- Even more tech to maintain
- Real time can get real expensive



# CURRENT STATE


Events
Protobuf Events
Apache Kafka
Storm
Secor
Write to Parquet
S3 Logs
Parquet
Spark Pipeline Spark/Scala
MySQL
VoltDB
Table Syncs
Parquet, CSV,JSON
Trigger/Monitor Jobs
Table Syncs
Apache Kafka
"Complete" Events (CSV)
Table Syncs
Pipeline Manager
Apache Kafka
S3 Data
Trigger/Monitor Jobs
Load CSV & JSON via ingestion CMD
Load CSV via Copy CMD
Druid
"Complete" Events (CSV)
Databricks
Looker
Redshift
Reporting API & UI
BI Tools

# LESSONS LEARNED

- The seemingly simple stuff is difficult
  - Dependencies
  - Scheduling

- Stop hacking open source libraries: Vanilla is an uninspired yet classic and delicious flavor
  - Secor
  - Kafka

- SQL really is everywhere

- Changing code is much easier than changing data

- The big data ecosystem is huge with tons of tools

# Q & A

# THANK YOU