Swimming in the data river

Or, when “streaming analytics” isn’t

Gian Merlino
gian@imply.io

Who am I?

Gian Merlino

Committer & PMC member on druid

Cofounder at imply

10 years working on scalable systems

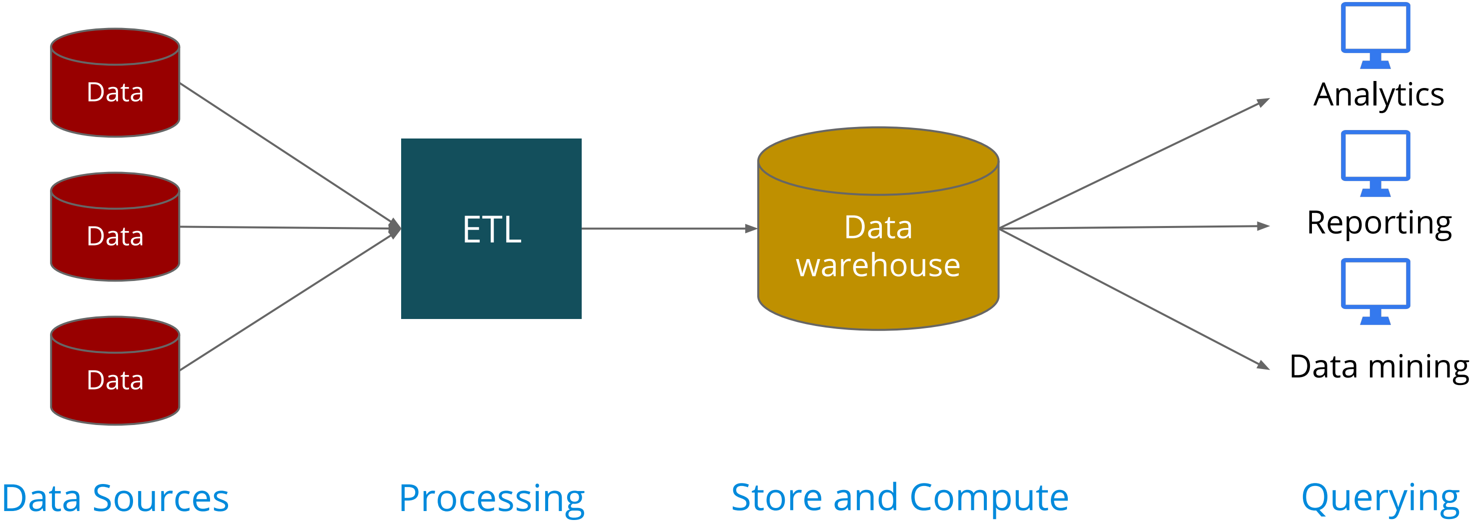
Agenda

- From warehouses to rivers
- What can we do with streaming data?
- Streaming analytics
- Enter the Druid
- Do try this at home!

Rolling down the river

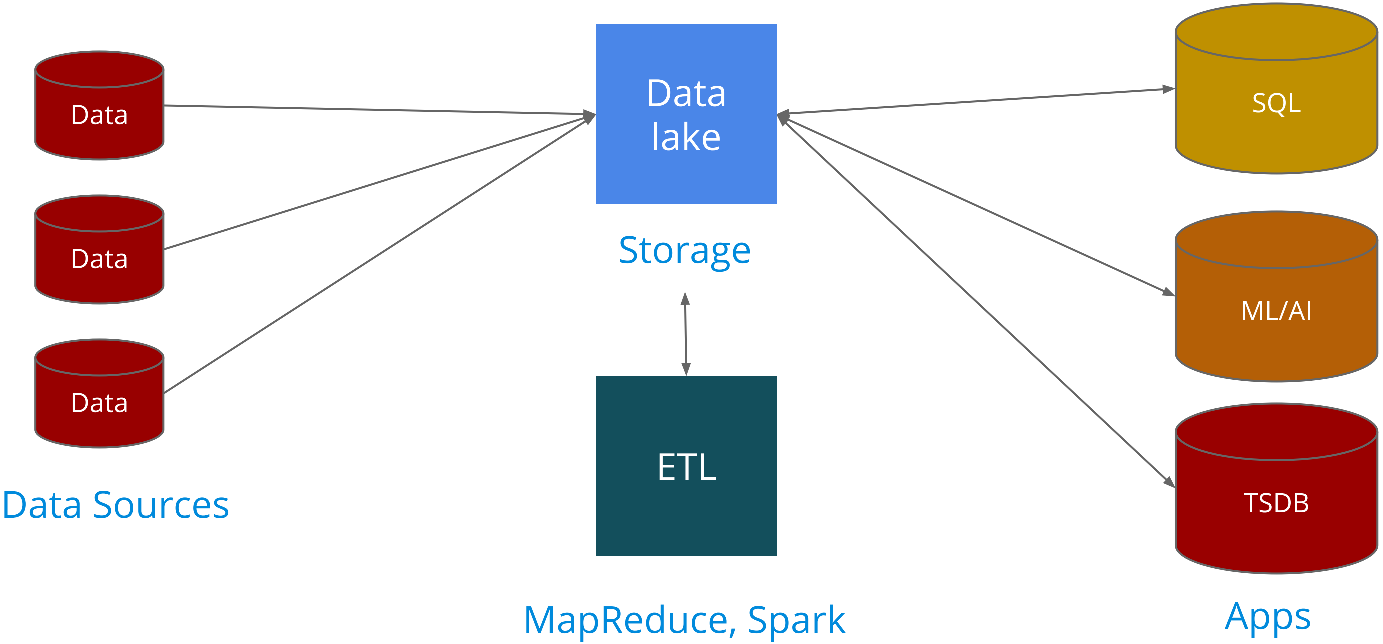
Data warehouses

Tightly coupled architecture with limited flexibility.



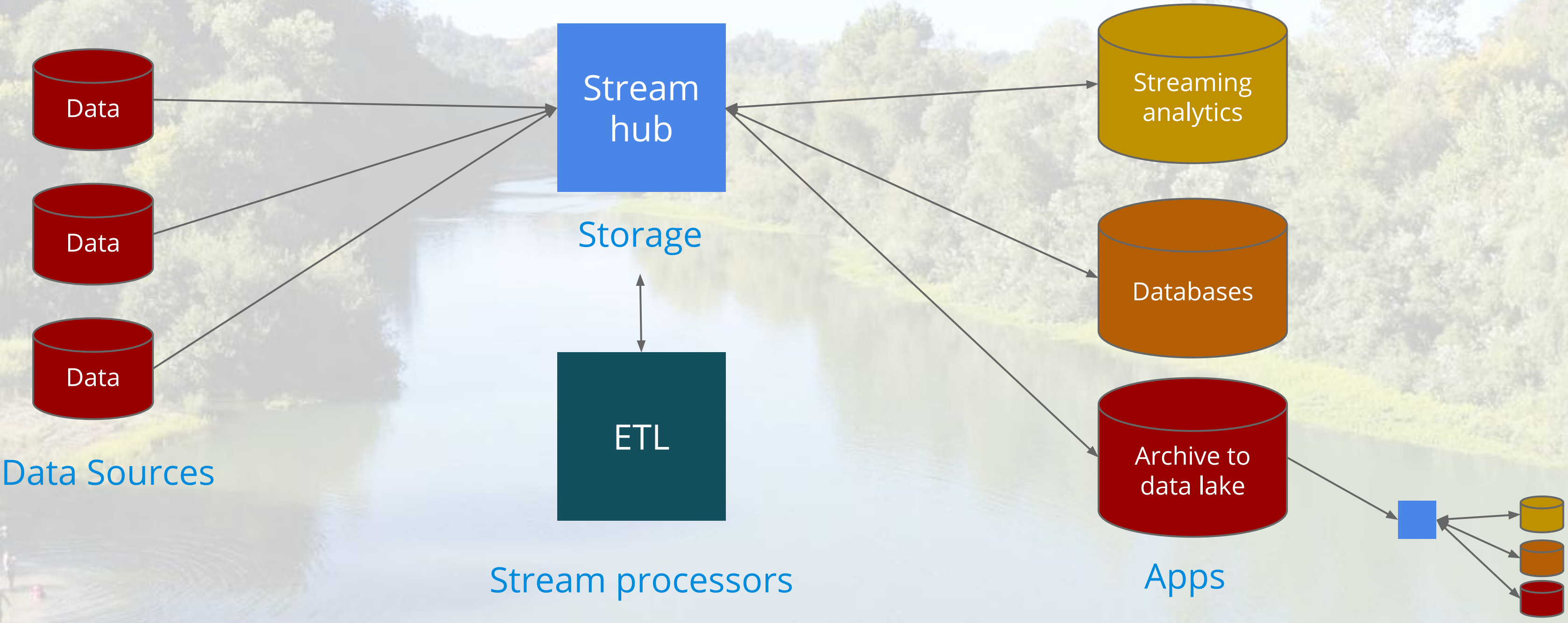
Data lakes

Modern data architectures are more application-centric.

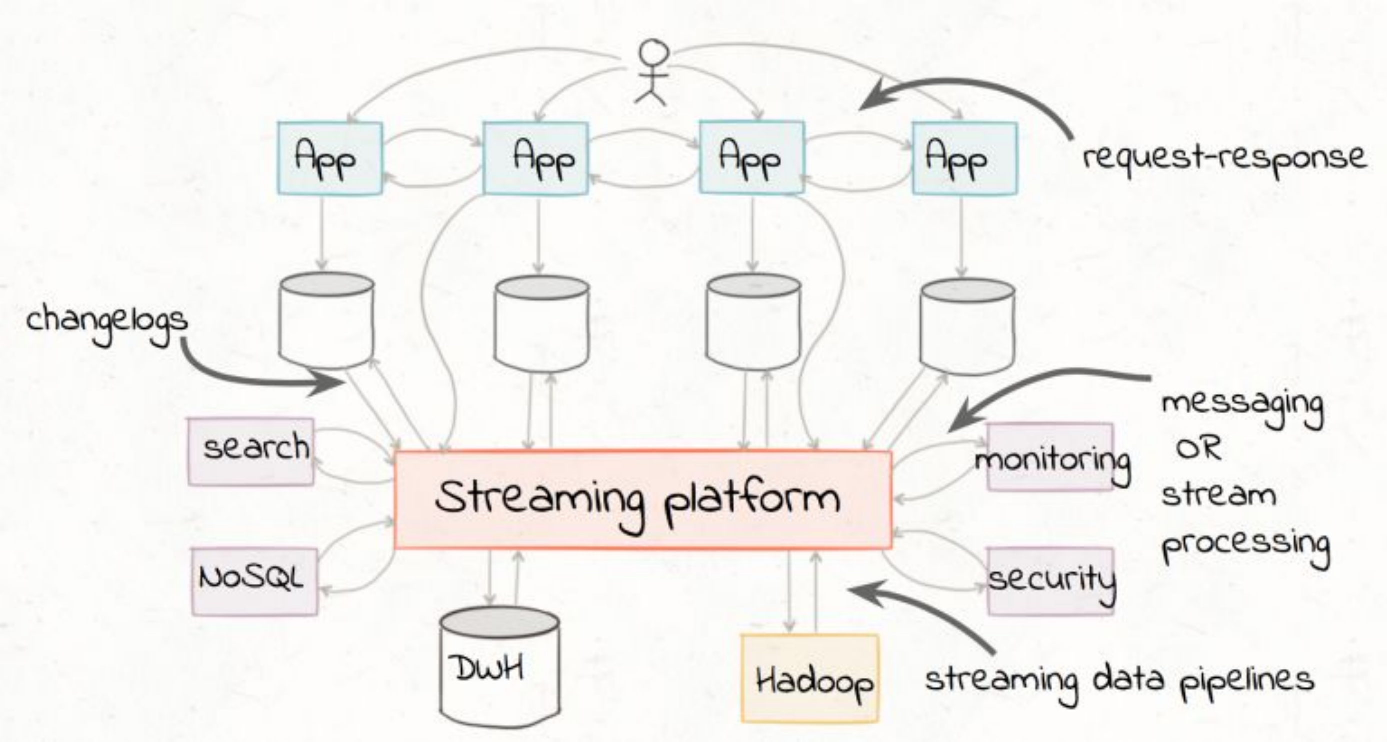


Data rivers

Streaming architectures are true-to-life and enable faster decision cycles.

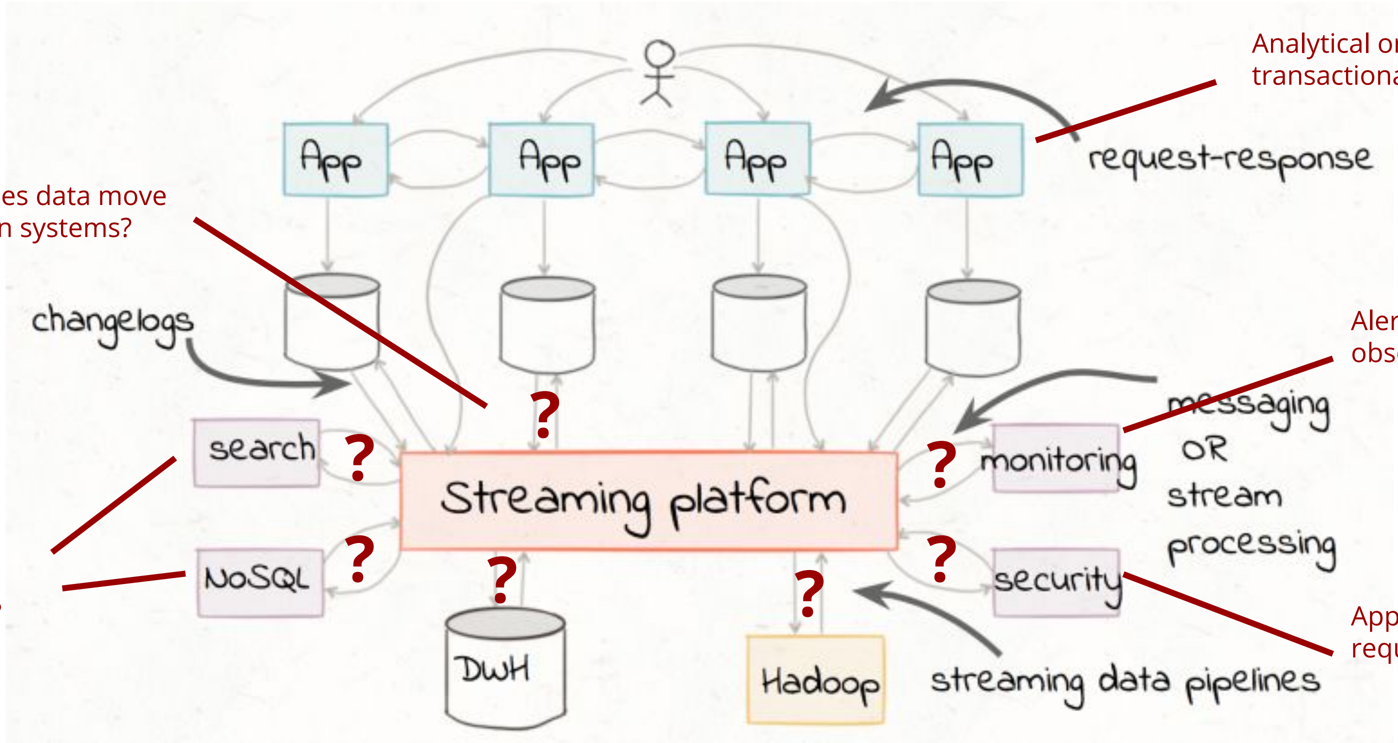


Streaming data



Source: https://www.confluent.io/blog/building-real-time-streaming-etl-pipeline-20-minutes/

Streaming data



Analytical or transactional?

How does data move between systems?

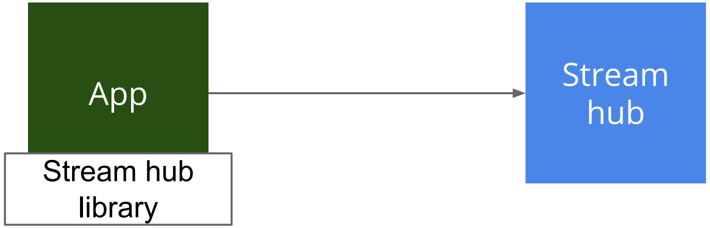
Alerting, charting, observability, APM?

Do search and NoSQL need apps?

App or set of requirements?

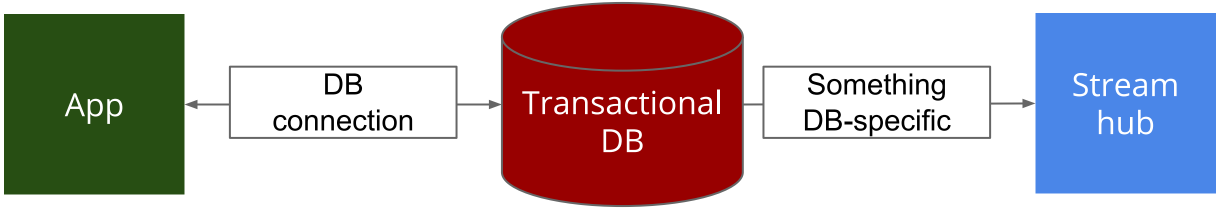
Source: https://www.confluent.io/blog/building-real-time-streaming-etl-pipeline-20-minutes/ (plus question marks)

Streaming data



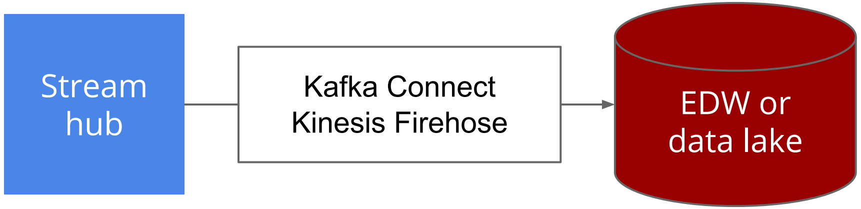
Direct production

Streaming data



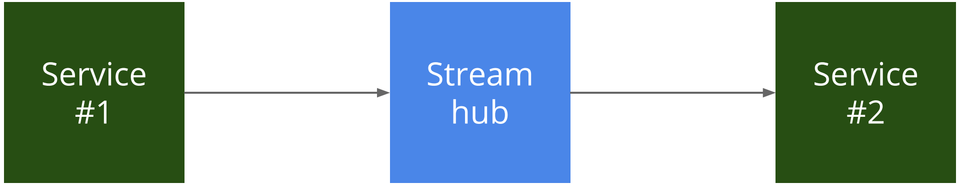
Change data capture

Streaming data



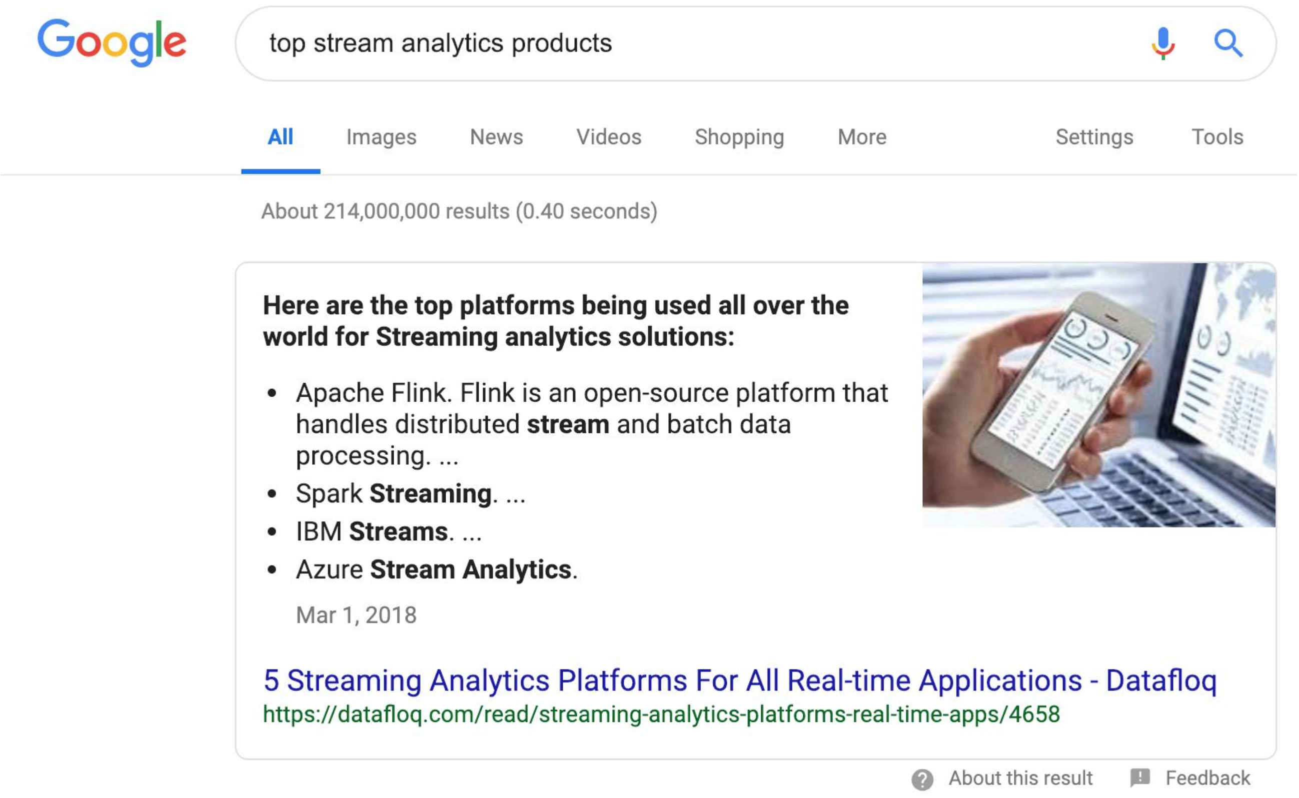
Streaming data pipeline

Streaming data



Microservice communication

Streaming data



top stream analytics products

About 214,000,000 results (0.40 seconds)

Here are the top platforms being used all over the world for Streaming analytics solutions:

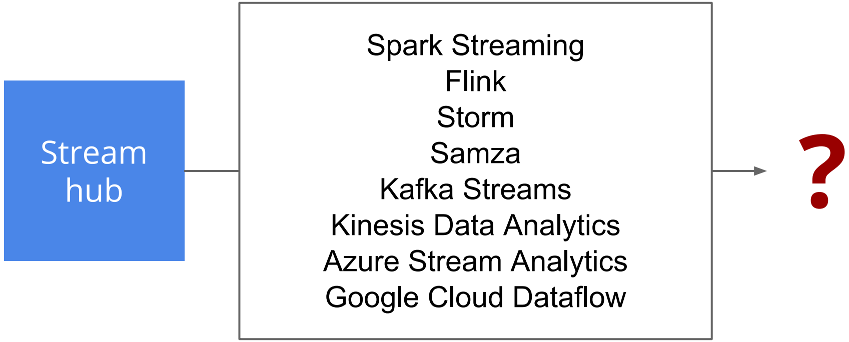
- Apache Flink. Flink is an open-source platform that handles distributed **stream** and batch data processing. ...
- Spark **Streaming**. ...
- IBM **Streams**. ...
- Azure **Stream Analytics**.

Mar 1, 2018

5 Streaming Analytics Platforms For All Real-time Applications - Datafloq

https://datafloq.com/read/streaming-analytics-platforms-real-time-apps/4658

Streaming data

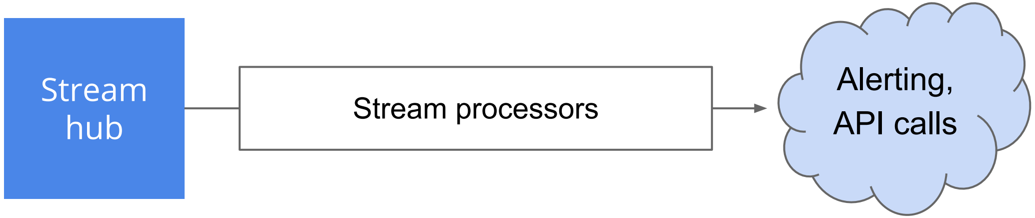


Stream hub

Spark Streaming
Flink
Storm
Samza
Kafka Streams
Kinesis Data Analytics
Azure Stream Analytics
Google Cloud Dataflow

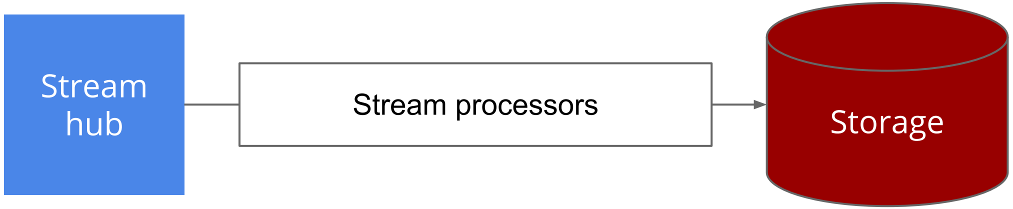
?

Streaming data



Real-time actions

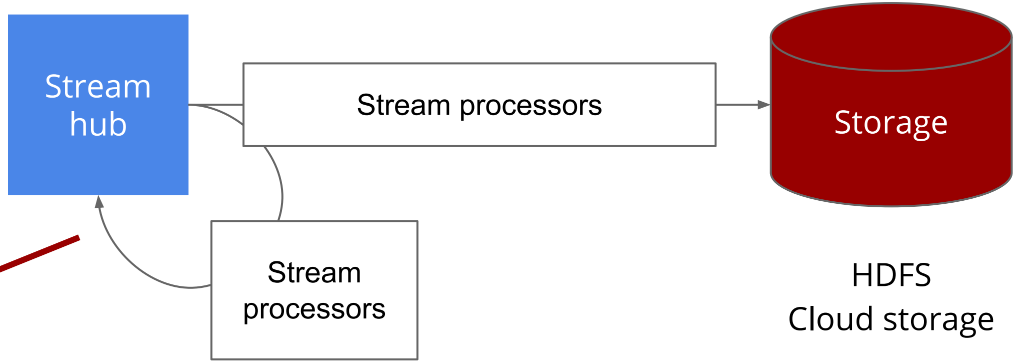
Streaming data



HDFS
Cloud storage

Data movement

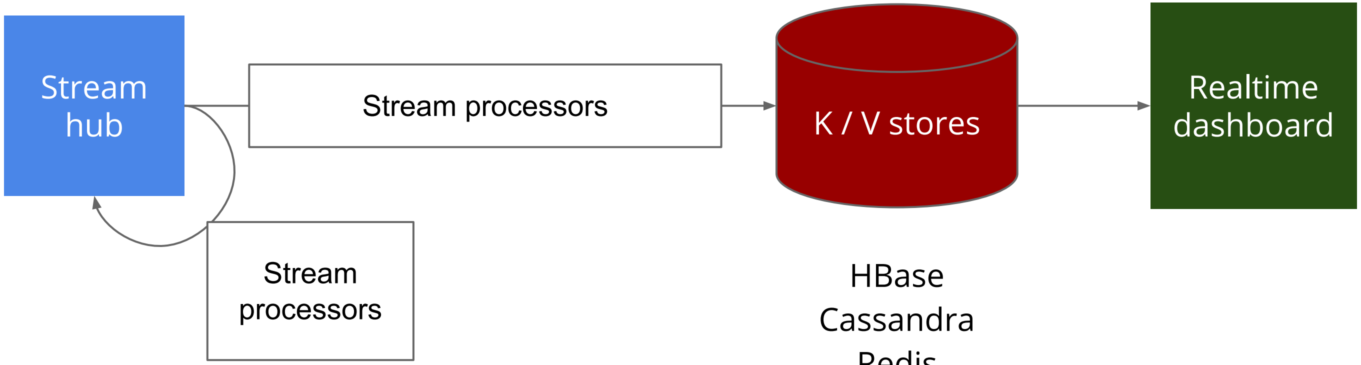
Streaming data



Continuous query

Data movement + enrichment

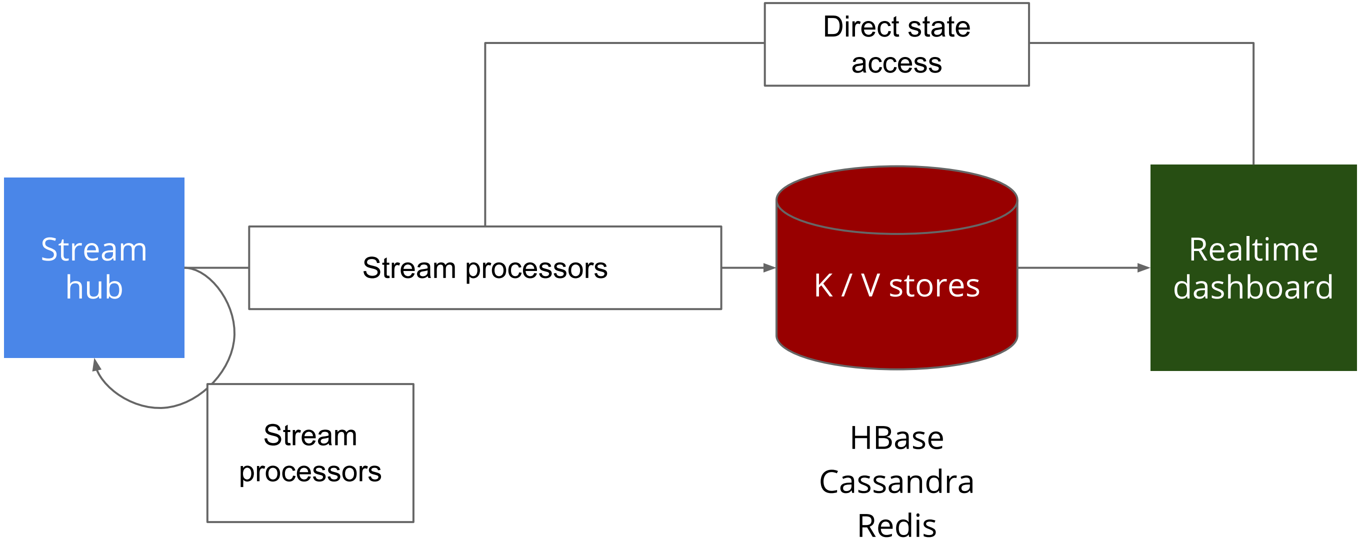
Streaming data



HBase
Cassandra
Redis

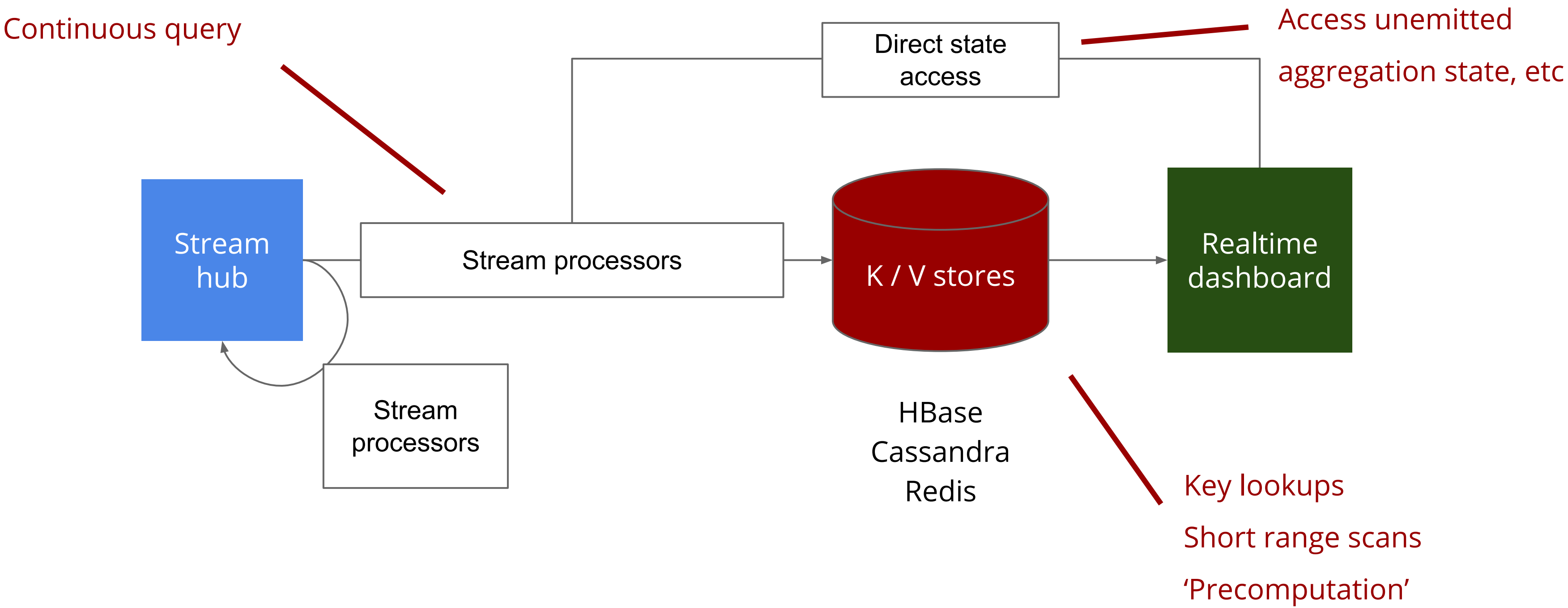
Continuous query + write to serving layer

Streaming data



Continuous query + write to serving layer + unemitted state serving

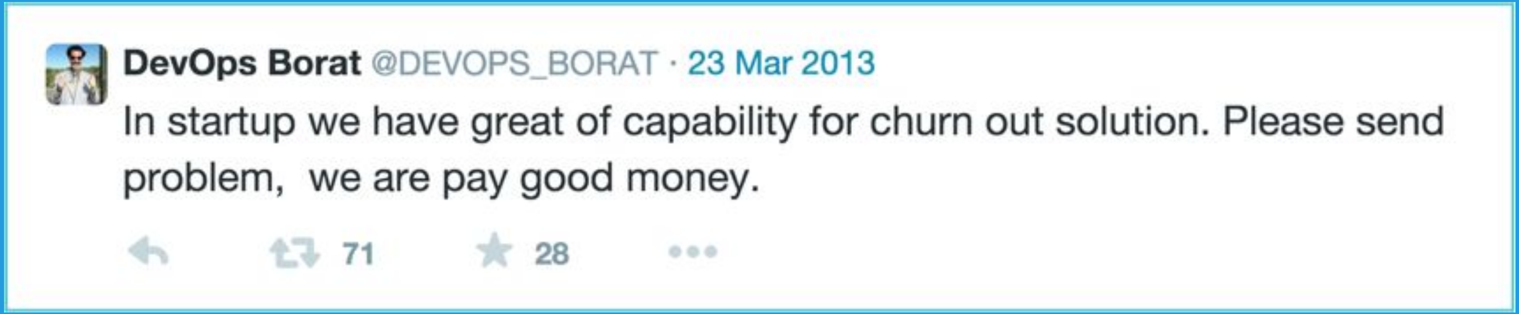
Streaming data



Continuous query + write to serving layer + unemitted state serving

Spark Streaming
百科事典
Is this analytics?

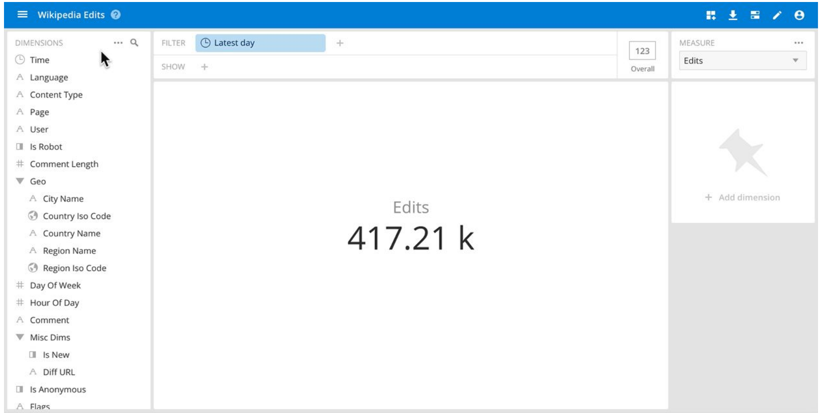
The problem



DevOps Borat @DEVOPS_BORAT · 23 Mar 2013

In startup we have great of capability for churn out solution. Please send problem, we are pay good money.

The problem

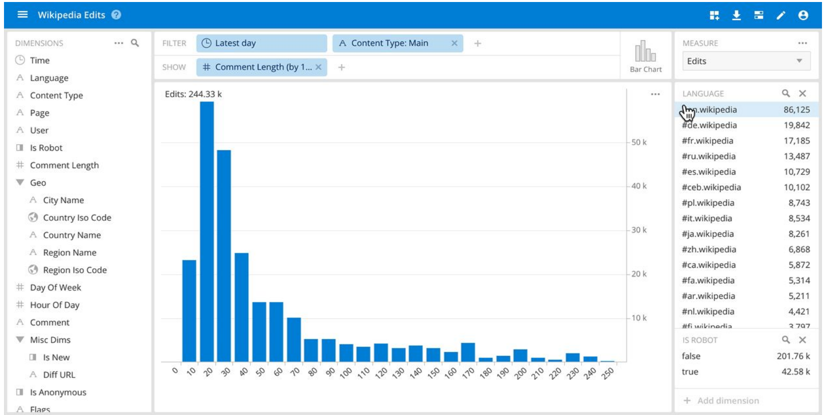


FILTER: Latest day

SHOW

MEASURE: Edits

Edits
417.21 k



FILTER: Latest day | Content Type: Main

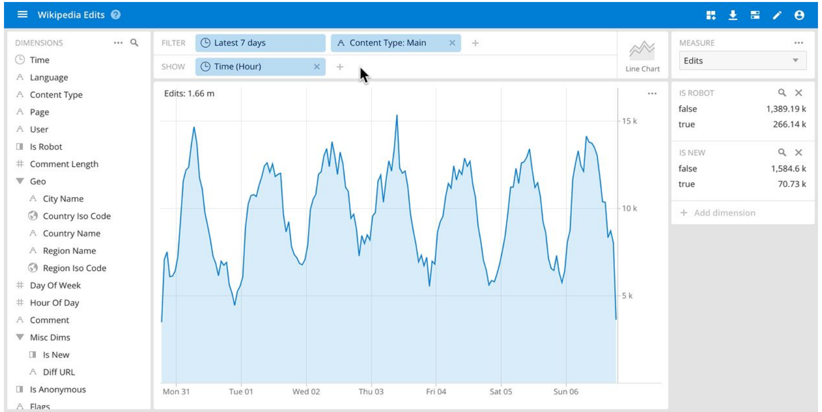
SHOW: Comment Length (by 1...

MEASURE: Edits

Edits: 244.33 k

LANGUAGE	
n.wikipedia	86,125
#de.wikipedia	19,842
#fr.wikipedia	17,185
#ru.wikipedia	13,487
#es.wikipedia	10,729
#ceb.wikipedia	10,102
#pl.wikipedia	8,743
#it.wikipedia	8,534
#ja.wikipedia	8,261
#zh.wikipedia	6,868
#ca.wikipedia	5,872
#fa.wikipedia	5,314
#ar.wikipedia	5,211
#nl.wikipedia	4,421
#fi.wikipedia	3,797

IS ROBOT	
false	201.76 k
true	42.58 k



FILTER: Latest 7 days | Content Type: Main

SHOW: Time (Hour)

MEASURE: Edits

Edits: 1.66 m

IS ROBOT	
false	1,389.19 k
true	266.14 k

IS NEW	
false	1,584.6 k
true	70.73 k

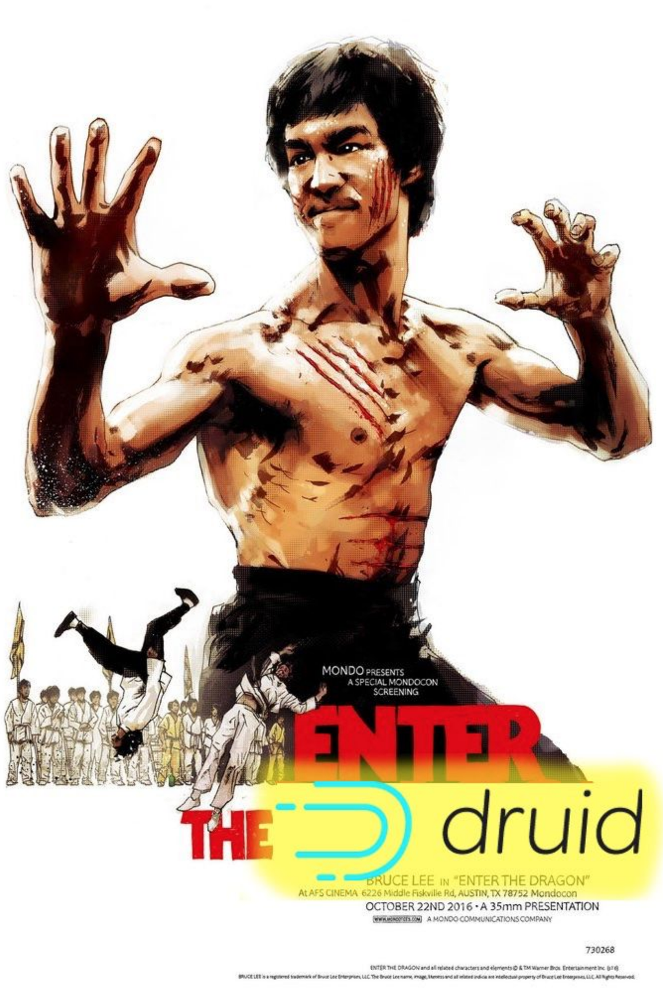
The problem

- Slice-and-dice for big data streams
- Interactive exploration
- Look under the hood of reports and dashboards
- And we want our data fresh, too

Challenges

- Scale: when data is large, we need a lot of servers
- Speed: aiming for sub-second response time
- Complexity: too much fine grain to precompute
- High dimensionality: 10s or 100s of dimensions
- Concurrency: many users and tenants
- Freshness: load from streams

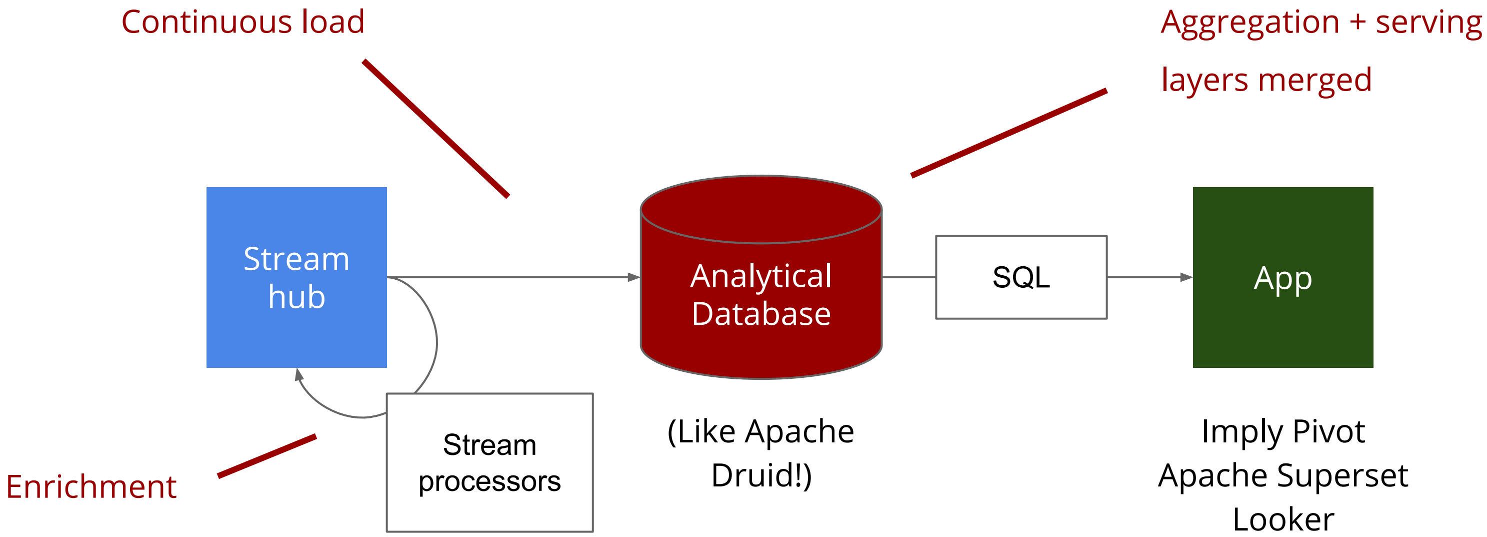
high performance
analytics data store for
event-driven data


MONDO PRESENTS
A SPECIAL MONDOCON
SCREENING
ENTER
THE
druid
BRUCE LEE IN "ENTER THE DRAGON"
At AFS CINEMA 6226 Middle Fiskville Rd, AUSTIN, TX 78752 Mondocon
OCTOBER 22ND 2016 • A 35mm PRESENTATION
A MONDO COMMUNICATIONS COMPANY
730268
ENTER THE DRAGON and all related characters and elements © & TM Warner Bros. Entertainment Inc.

What is Druid?

- **“high performance”:** low query latency, high ingest rates
- **“analytics”:** counting, ranking, groupBy, time trend
- **“data store”:** the cluster stores a copy of your data
- **“event-driven data”:** fact data like clickstream, network flows, user behavior, digital marketing, server metrics, IoT

Streaming data



(Like Apache Druid!)

Imply Pivot
Apache Superset
Looker
Your App Here™

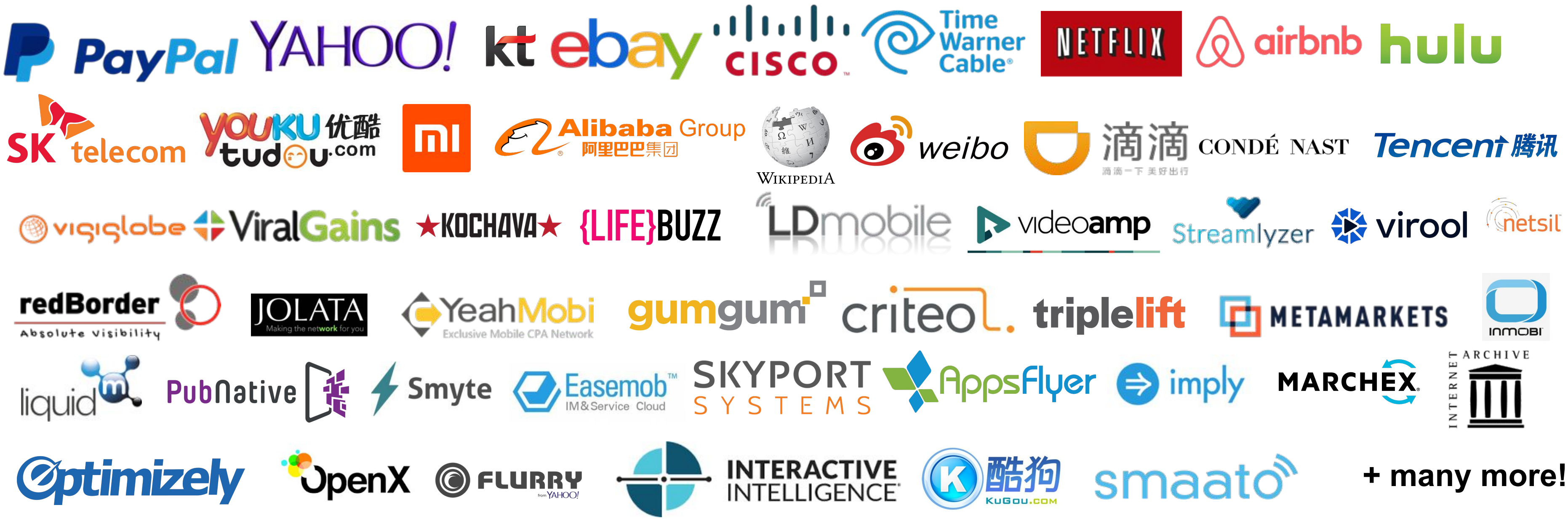
Key features

- Column oriented
- High concurrency
- Scalable to 100s of servers, millions of messages/sec
- Continuous, real-time ingest
- Indexes on all dimensions by default
- Query through SQL
- Target query latency sub-second to a few seconds

Use cases

- Clickstreams, user behavior
- Digital advertising
- Application performance management
- Network flows
- IoT

Powered by Apache Druid



+ many more!

Source: http://druid.io/druid-powered.html

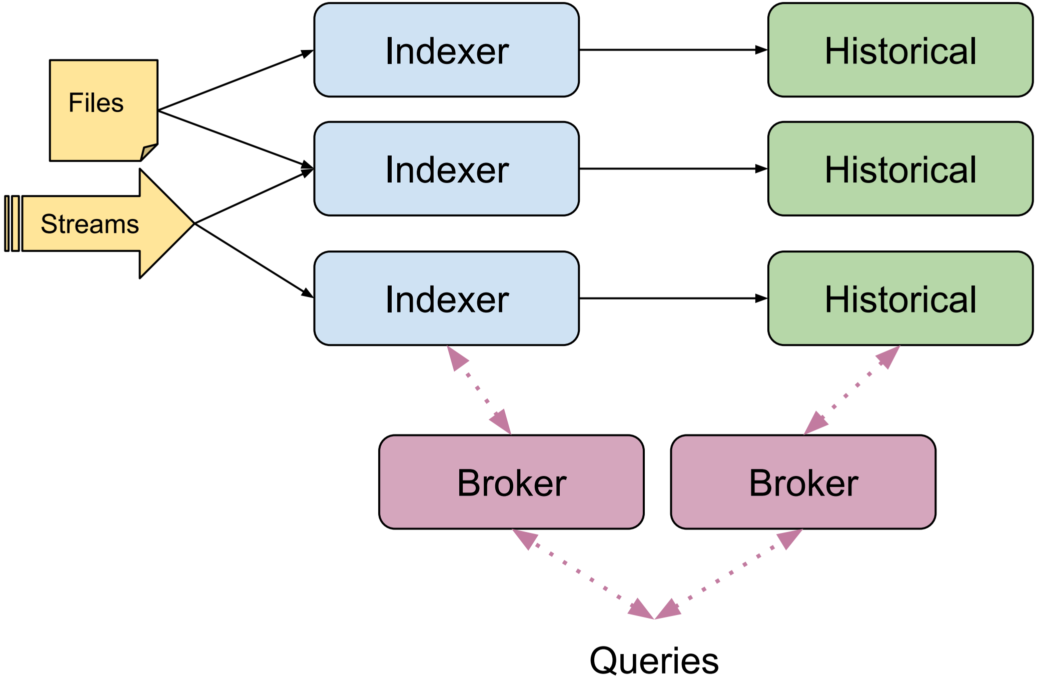
Powered by Apache Druid

From Yahoo:

“The performance is great ... some of the tables that we have internally in Druid have **billions and billions of events** in them, and we’re scanning them in **under a second**.”

Source: https://www.infoworld.com/article/2949168/hadoop/yahoo-struts-its-hadoop-stuff.html

Architecture

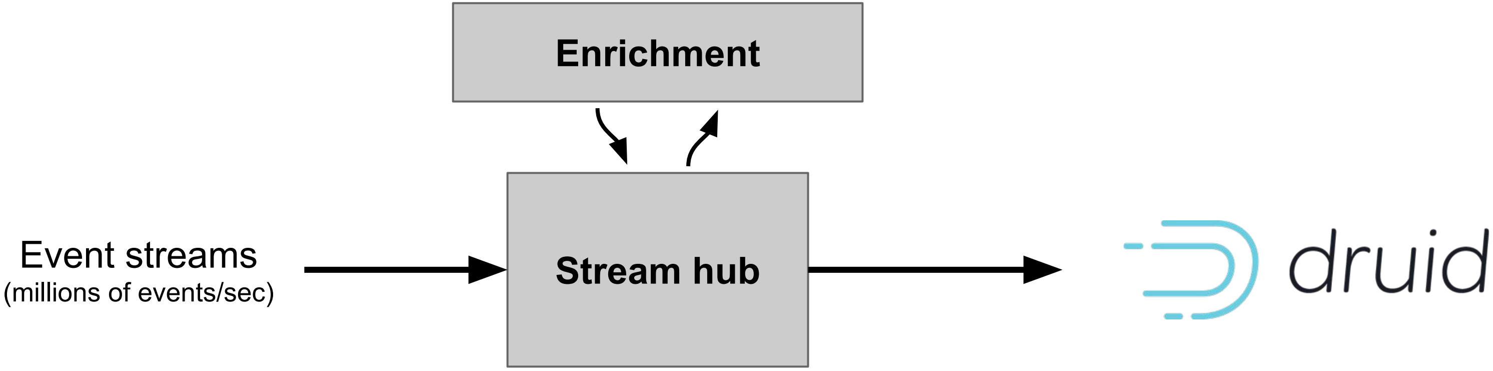


Why this works

- Computers are fast these days
- Indexes help save work and cost
- But don’t be afraid to scan tables — it can be done efficiently

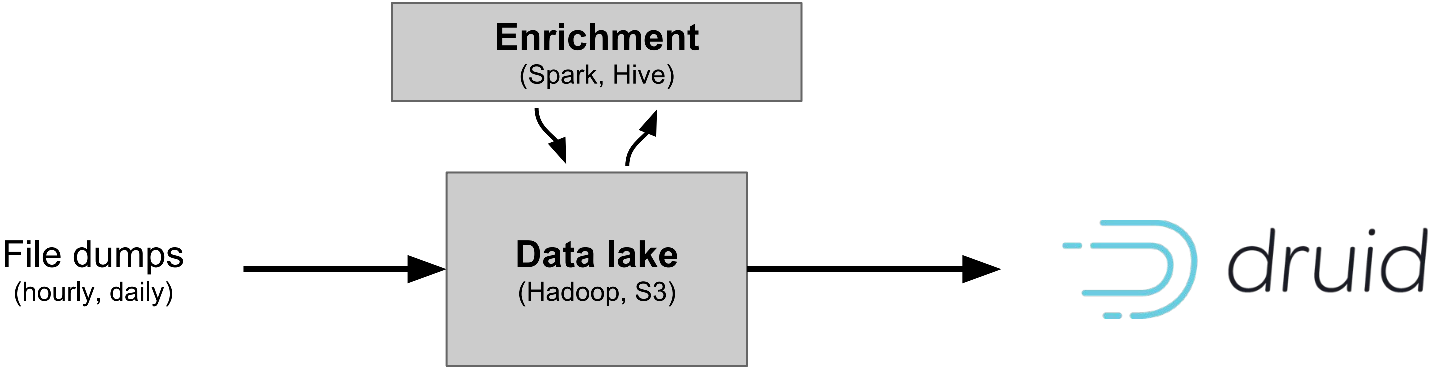
Integration patterns

Deployment patterns

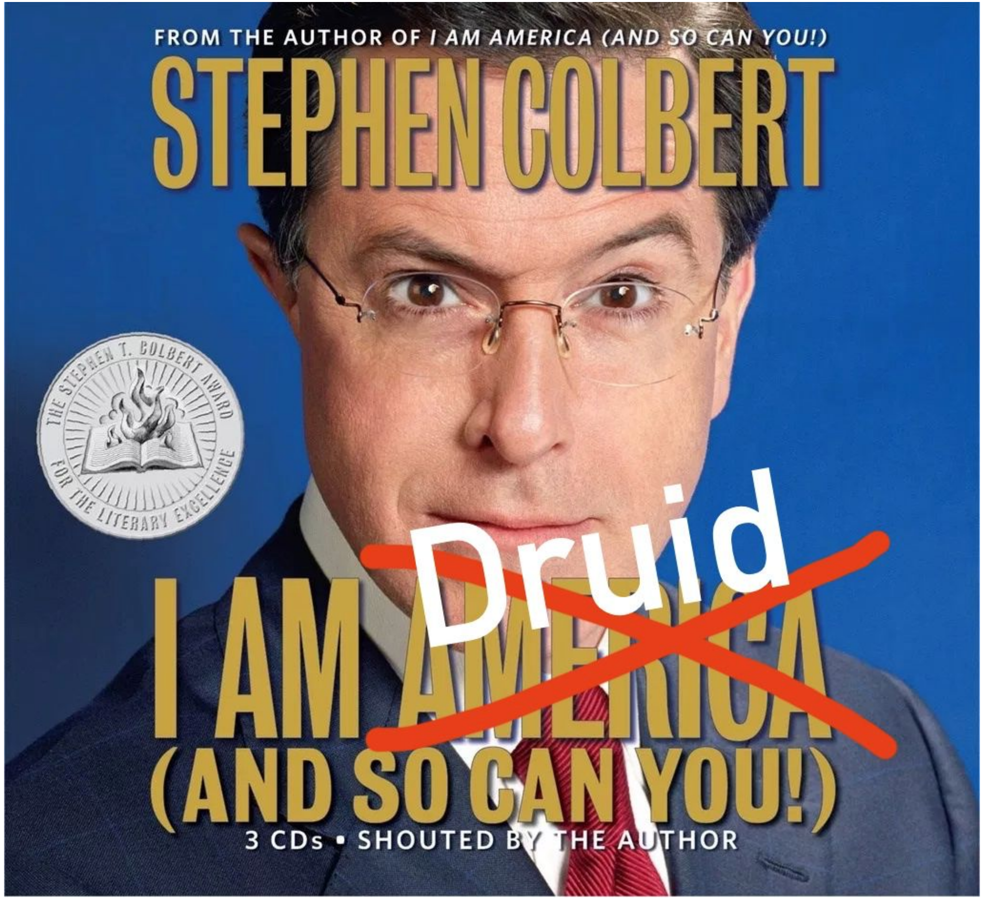


- Modern data architecture
- Centered around stream hub

Deployment patterns



- (Slightly less) modern data architecture
- Centered around data lake


FROM THE AUTHOR OF I AM AMERICA (AND SO CAN YOU!)
STEPHEN COLBERT
THE STEPHEN T. COLBERT AWARD FOR THE LITERARY EXCELLENCE
Druid
I AM ~~AMERICA~~
(AND SO CAN YOU!)
3 CDs • SHOUTED BY THE AUTHOR

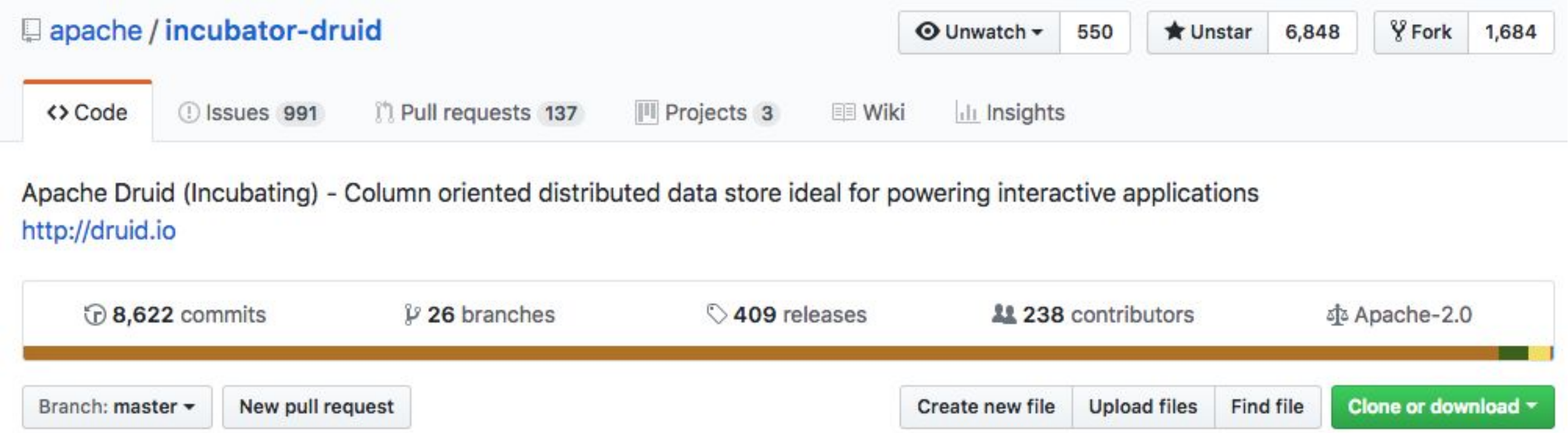
Download

Apache Druid community site (new): https://druid.apache.org/

Apache Druid community site (legacy): http://druid.io/

Imply distribution: https://imply.io/get-started

Contribute



Apache Druid (Incubating) - Column oriented distributed data store ideal for powering interactive applications http://druid.io

https://github.com/apache/druid

Stay in touch

Follow the Druid project on Twitter!

@druidio

Join the community!

http://druid.apache.org/