Open Source & the Cloud

Open Source development adopting cloud at Twitter

ApacheCon 2022

Daniel Templeton @templedf
Lohit VijayRenu @lohitvijayarenu

Lohit VijayaRenu

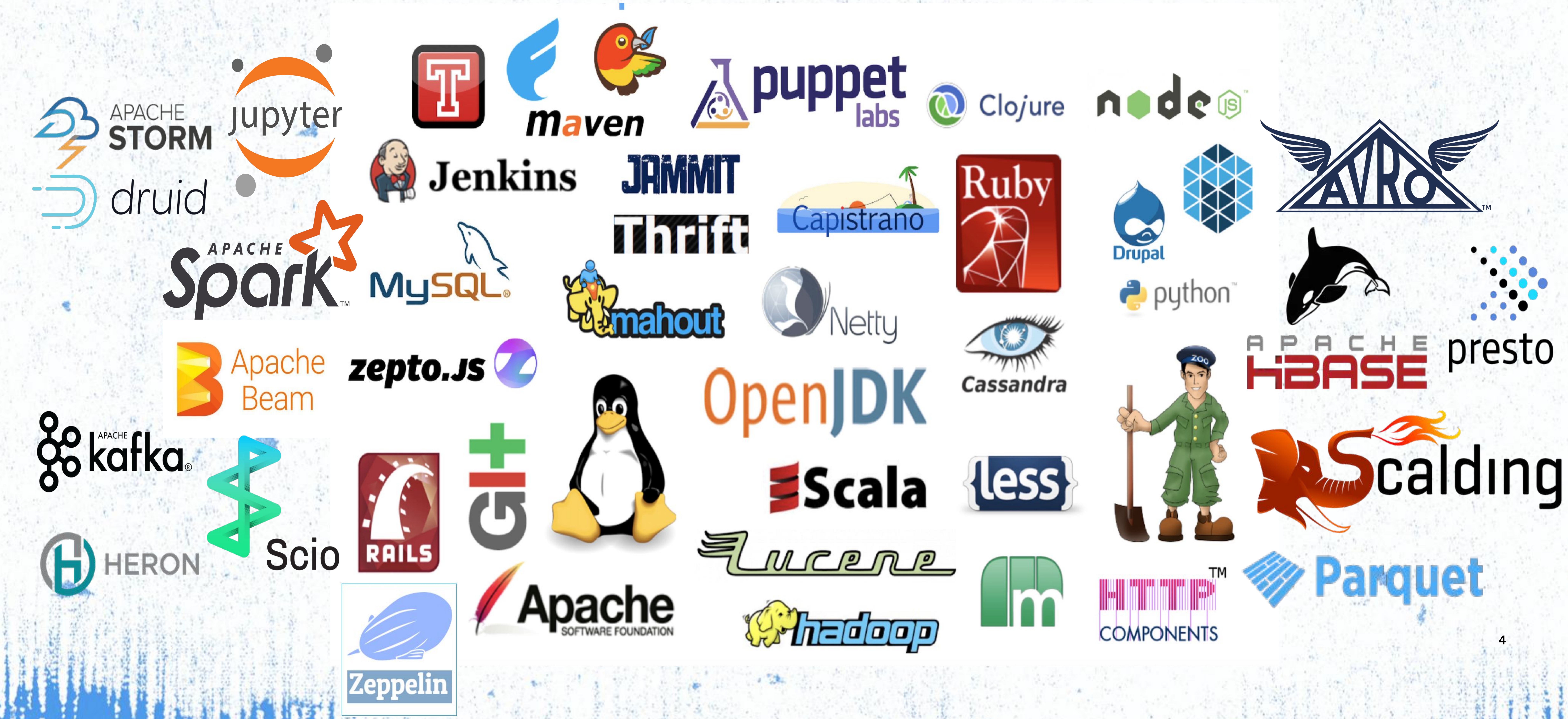
He/Him
Principal Software Eng
Apache Hadoop committer
@lohitvijayarenu

Daniel Templeton

He/Him
Sr EM Data Lifecycle
Apache Hadoop PMC
@templedf

Open Source adoption at Twitter and how it is evolving with extending infrastructure support for Cloud

Open Source & Twitter



Data Platform use cases

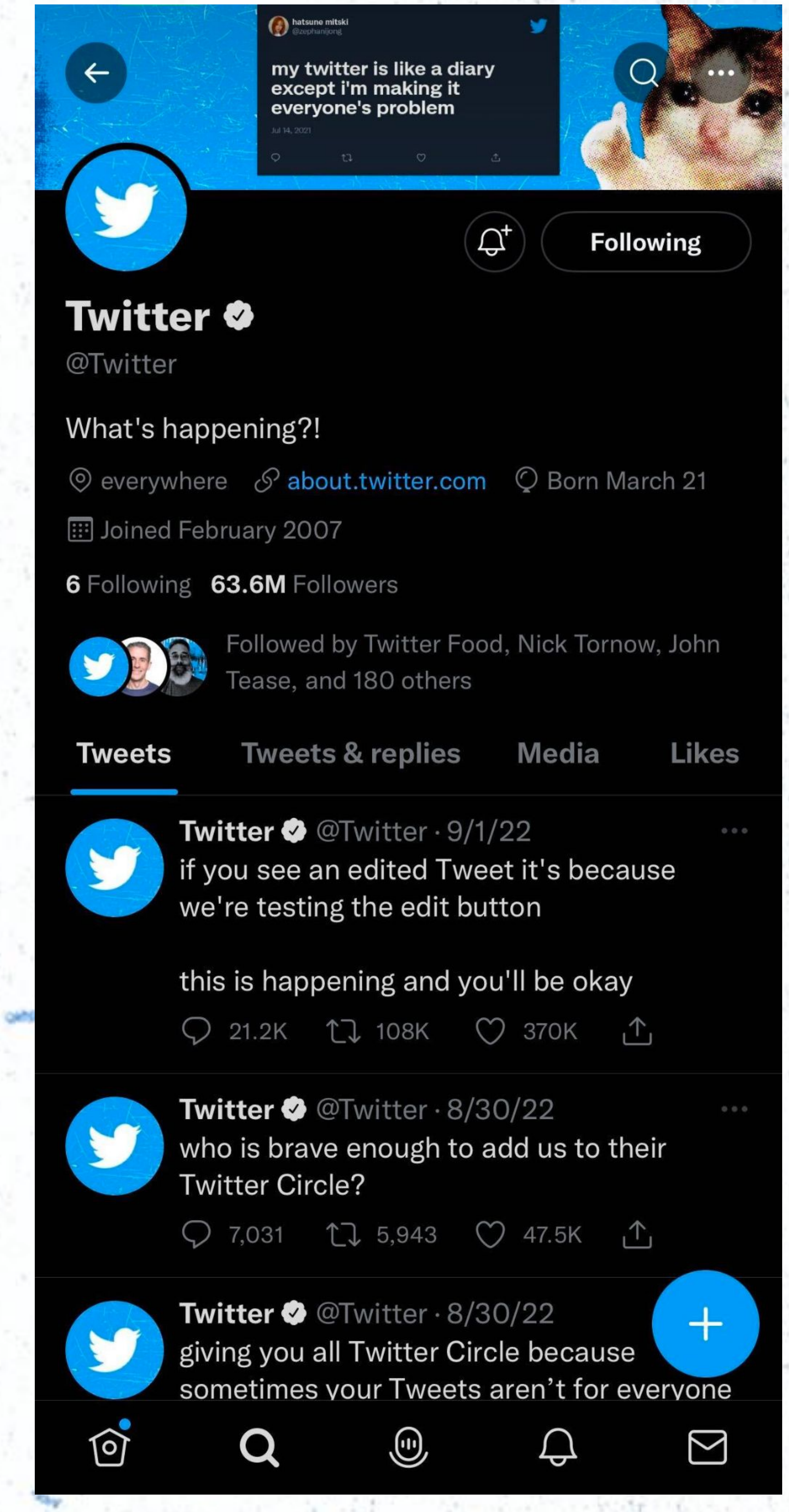
Centralized data processing infrastructure

Data management - storage, metadata, log ingestion, replication, retention

Data processing - batch, streaming, aggregations

Data analytics - SQL, reporting, quality

Data products - curated datasets



hatsune mitski
@zephanijong

my twitter is like a diary except i'm making it everyone's problem

Jul 14, 2021

Twitter
@Twitter

What's happening?!

everywhere about.twitter.com Born March 21

Joined February 2007

6 Following 63.6M Followers

Followed by Twitter Food, Nick Tornow, John Tease, and 180 others

Twitter @Twitter · 9/1/22
if you see an edited Tweet it's because we're testing the edit button

this is happening and you'll be okay

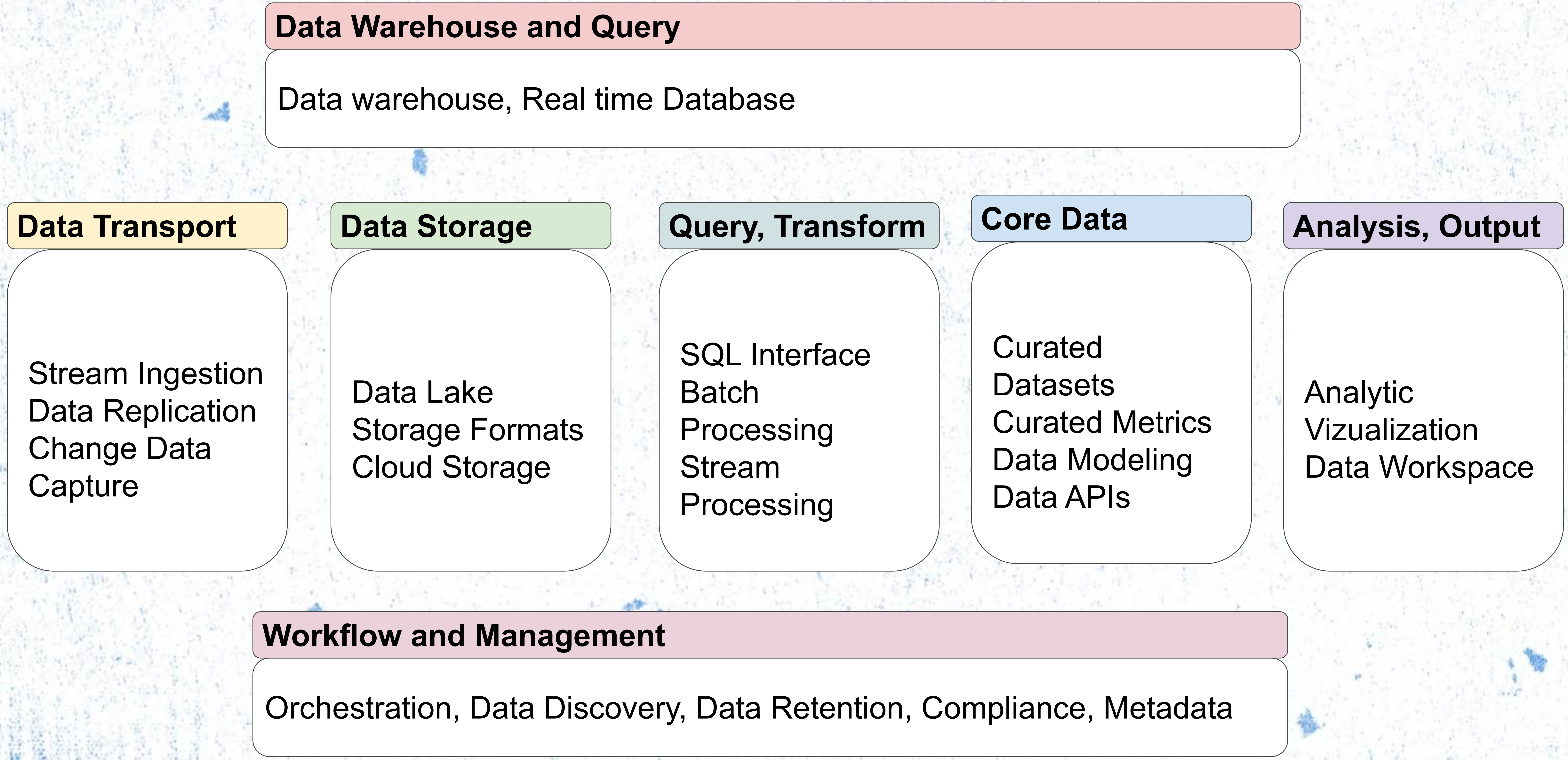
21.2K 108K 370K

Twitter @Twitter · 8/30/22
who is brave enough to add us to their Twitter Circle?

7,031 5,943 47.5K

Twitter @Twitter · 8/30/22
giving you all Twitter Circle because sometimes your Tweets aren't for everyone

Data Platform



Data Warehouse and Query

Data warehouse, Real time Database

Data Transport

Stream Ingestion
Data Replication
Change Data Capture

Data Storage

Data Lake
Storage Formats
Cloud Storage

Query, Transform

SQL Interface
Batch Processing
Stream Processing

Core Data

Curated Datasets
Curated Metrics
Data Modeling
Data APIs

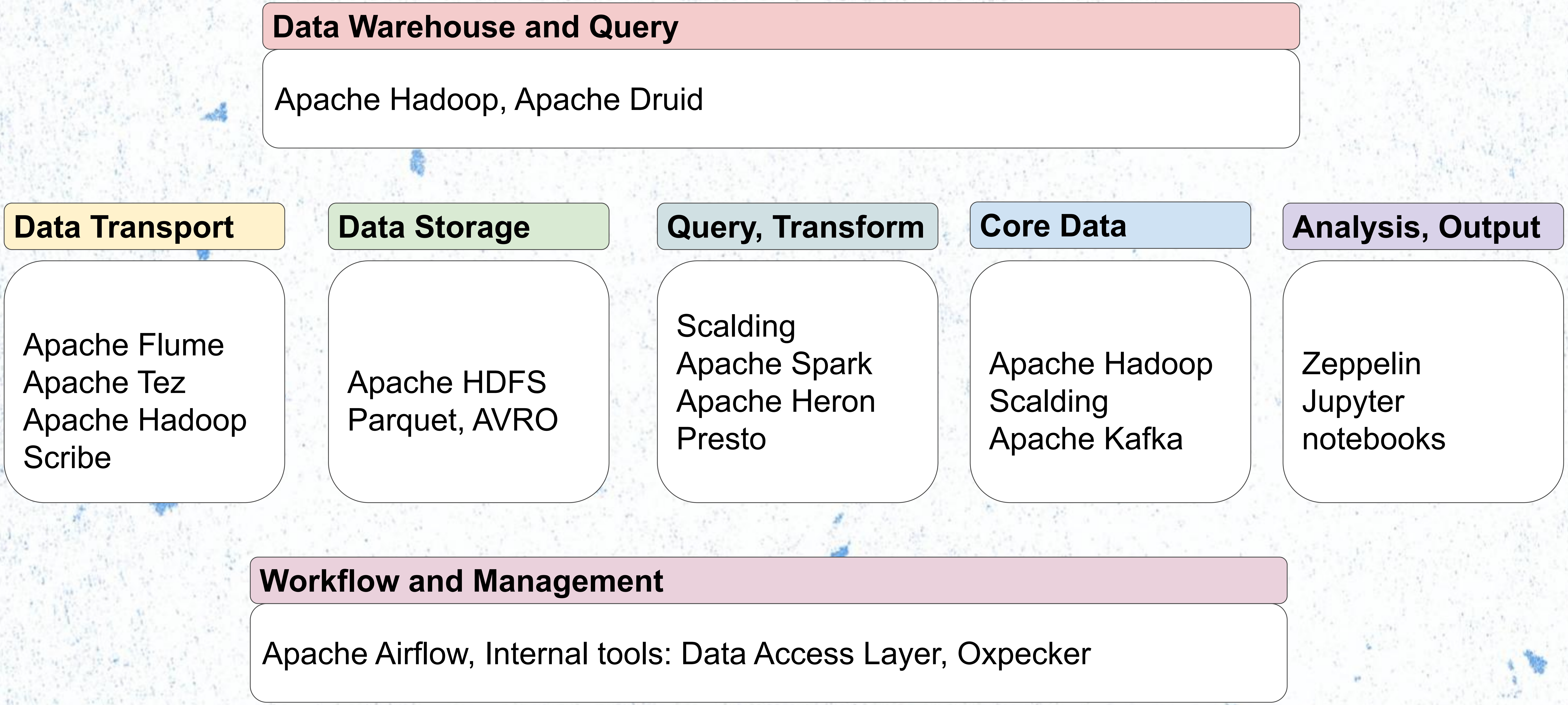
Analysis, Output

Analytic
Vizualization
Data Workspace

Workflow and Management

Orchestration, Data Discovery, Data Retention, Compliance, Metadata

On-prem and Open Source



Data Warehouse and Query

Apache Hadoop, Apache Druid

Data Transport

Apache Flume
Apache Tez
Apache Hadoop
Scribe

Data Storage

Apache HDFS
Parquet, AVRO

Query, Transform

Scalding
Apache Spark
Apache Heron
Presto

Core Data

Apache Hadoop
Scalding
Apache Kafka

Analysis, Output

Zeppelin
Jupyter
notebooks

Workflow and Management

Apache Airflow, Internal tools: Data Access Layer, Oxpecker

Complexities and contributions

- Adoption of open source ecosystem **components at scale**
  - Solving Data Processing and Data Analytic use cases
- Contribution to open source with features around **scalability and reliability**
  - Unique problems seen at scale and solutions for those
- Active engagement with **vendor** and **open source contributors**
- Discuss requirements and opportunities to solve complex problems
- Building strong relationships with **community**
  - Building strong technical teams to scale Twitter business
  - Graduating **developers to committers** and beyond

Why Cloud?

- Realize story around **Unified Data** and **Machine Learning**
- Rapidly **grow / shrink**
- A broader geographical footprint for locality and **business continuity**
- Solve complicated problems for **max ROI**
  - Capacity management
  - New features and technologies
  - Ecosystem integration
- Access to **other Google offerings** such as BigQuery, CloudML, Cloud DataFlow, VertexAI etc

Beginning of the Cloud Journey

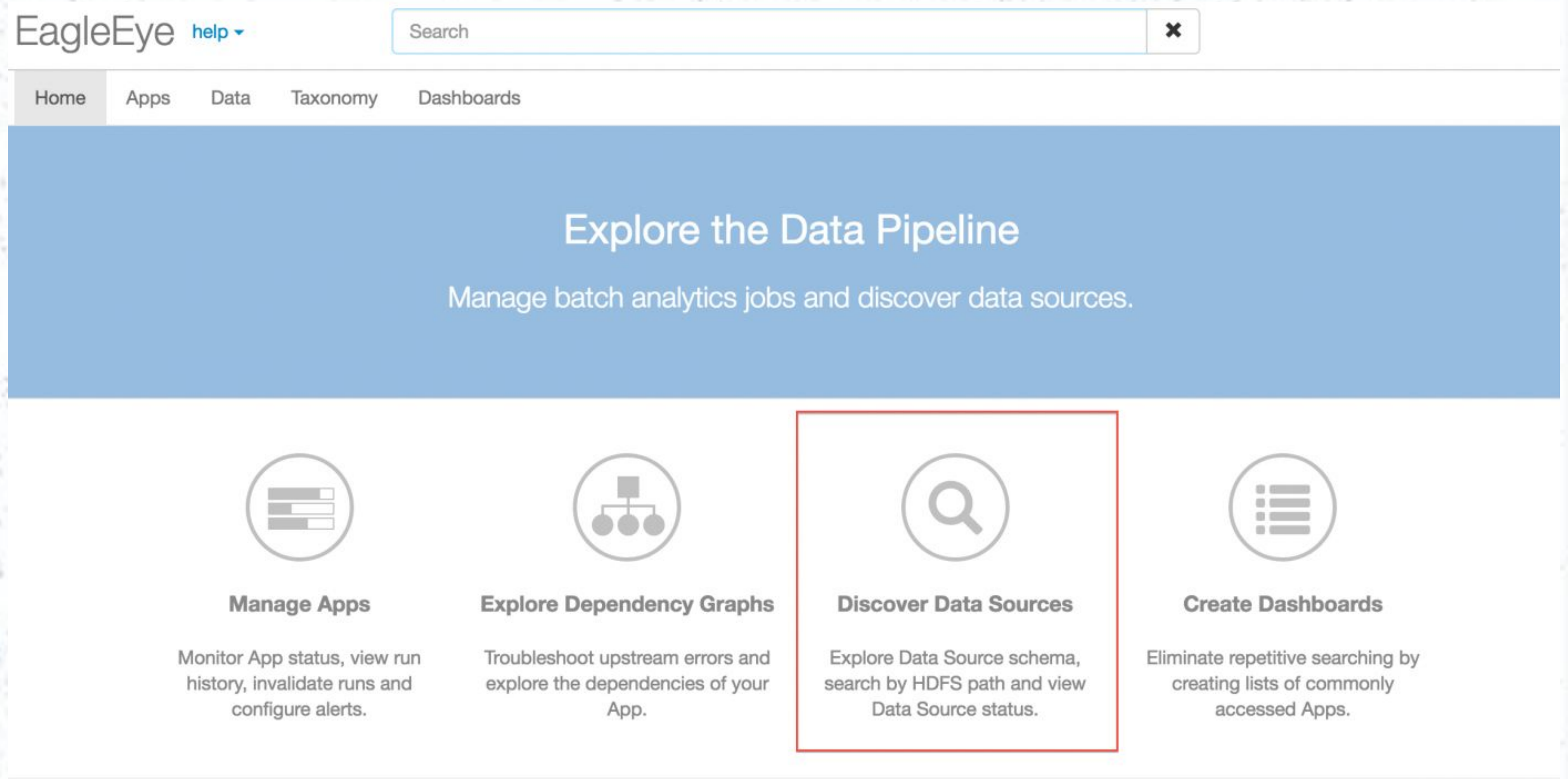
- Evaluate new **features and capabilities** provided by cloud vendors
    - Particularly to support ML/AI use cases
    - Cost v/s capability
- **Compare and contrast** with on-prem infrastructure
    - Benchmarking, stress testing and evaluation
    - Identify scalable and extensible components
- Justify the need for **adoption**
    - Pick right use cases
    - Utilize new features and capabilities
- Targeted rollout for **specific use cases**
    - Learn, rinse, repeat

Projects for enablement of cloud

- Integration with existing Twitter infra
  - **Metadata** integration
- Cloud **Resource organization**
  - Projects, buckets, tables and more
- **Security**
  - Identity management, extending Twitter security controls
- Data **Replication Service**
  - Batch, Streaming, and CDC
- **Networking**
- Evaluation of **new services**
  - Dataflow, PubSub, BigQuery, etc.

Metadata Integration

- Started with in-house **Data Access Layer**
  - Data Replication
  - Data Retention
  - Data management (Permission, schema, ownership…)
- Integrate cloud services with DAL
- Ensure **compliance and security** for cloud storage services:
  - Data annotation
  - Compliance enforcement
- Considering open source alternatives
  - DataHub, Open Metadata, Open Lineage
  - No end-to-end out-of-the-box solution



Explore the Data Pipeline

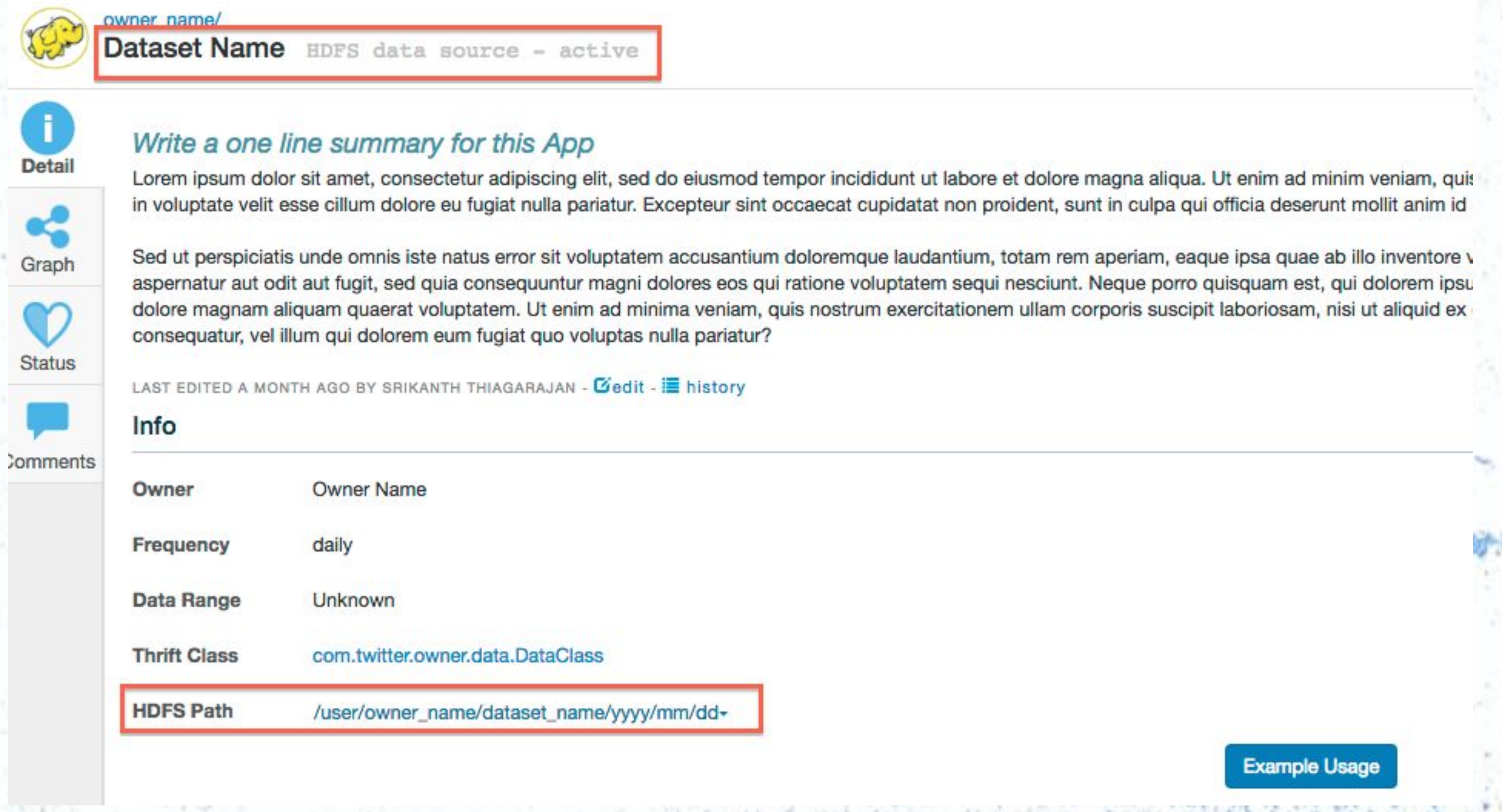
Manage batch analytics jobs and discover data sources.

Manage Apps
Monitor App status, view run history, invalidate runs and configure alerts.

Explore Dependency Graphs
Troubleshoot upstream errors and explore the dependencies of your App.

Discover Data Sources
Explore Data Source schema, search by HDFS path and view Data Source status.

Create Dashboards
Eliminate repetitive searching by creating lists of commonly accessed Apps.



owner_name/

Dataset Name HDFS data source - active

Write a one line summary for this App

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsu dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

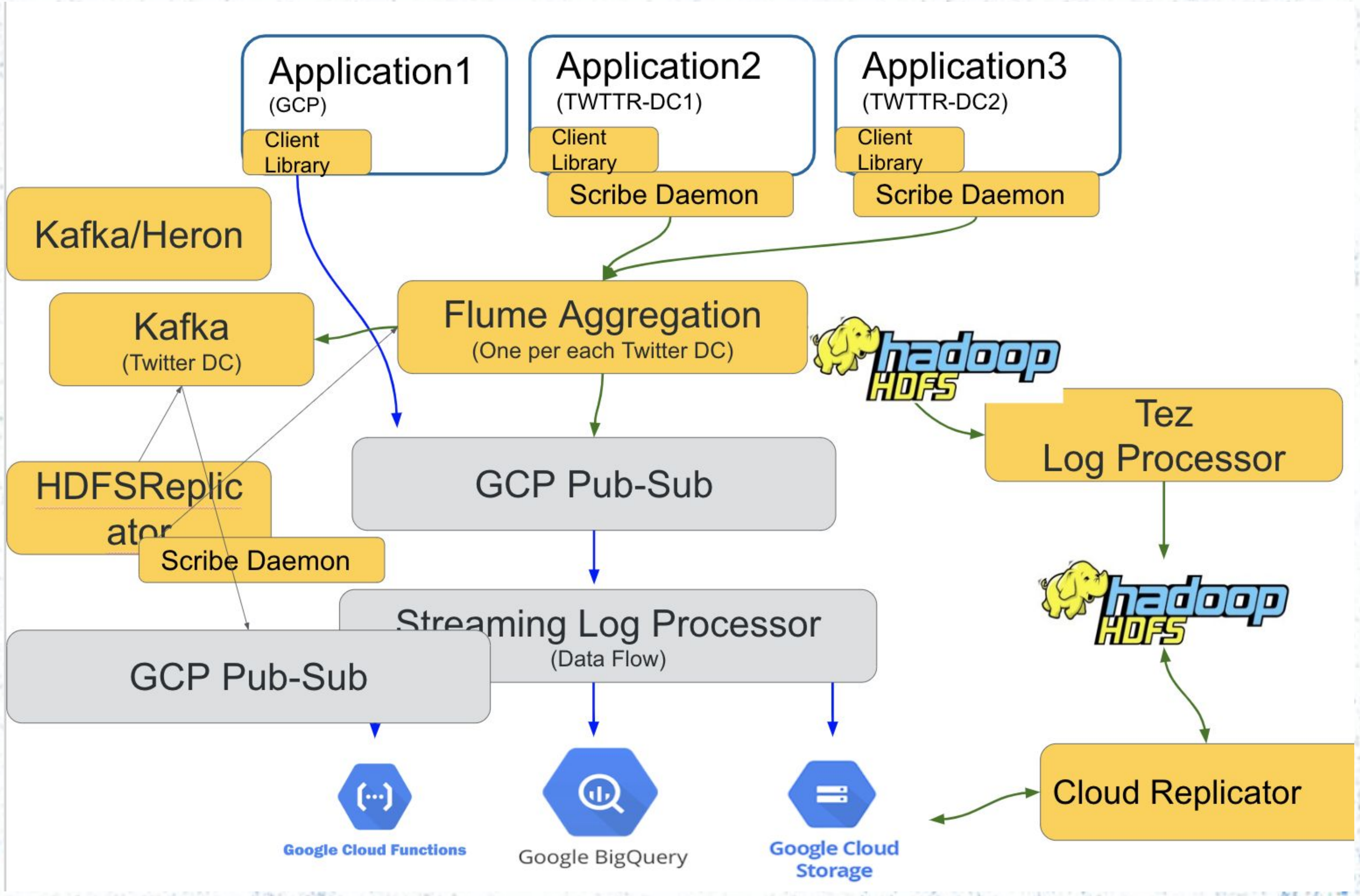
LAST EDITED A MONTH AGO BY SRIKANTH THIAGARAJAN - edit - history

Info

Owner	Owner Name
Frequency	daily
Data Range	Unknown
Thrift Class	com.twitter.owner.data.DataClass
HDFS Path	/user/owner_name/dataset_name/yyyy/mm/dd▾

Data Replication at Scale (PBs)

- Ability to **scale replication** to and from cloud
  - High throughput data transfer
  - Streaming
  - Change data streams
- Heavy **network requirements**
- Built using **open source**
  - Apache Hadoop
  - Apache Flume
  - Apache Kafka
  - Apache Beam
- No complete open source solution
  - Cloud services
  - Lots of in-house code



Application1 (GCP)

Client Library

Application2 (TWTTR-DC1)

Client Library

Scribe Daemon

Application3 (TWTTR-DC2)

Client Library

Scribe Daemon

Kafka/Heron

Kafka (Twitter DC)

Flume Aggregation (One per each Twitter DC)

hadoop HDFS

Tez Log Processor

HDFSReplic ator

Scribe Daemon

GCP Pub-Sub

Streaming Log Processor (Data Flow)

GCP Pub-Sub

hadoop HDFS

Google Cloud Functions

Google BigQuery

Google Cloud Storage

Cloud Replicator

Easy Onboarding

- Tooling to manage **security perimeters**
  - VPC SC setup
  - Identity and access management mapping
  - Extending onprem roles to cloud
- **Onboarding** Cloud services
  - Terraform setup
  - Cloud agnostic APIs for resource provisioning
- **Chargeback** of cloud services
  - Cloud resource utilization monitoring and alerting
  - Integrate with other systems (including on prem)
- Tooling for easier **cloud adoption**
  - Migration tooling, Onboarding guides and templates, Provisioning, Auditing…

Support Impact

- Increased **support load**
    - Different systems on prem and in the cloud
    - Tooling to support new systems
    - Permissions, setup, and resource provisioning
- **Guidance** for customers
    - Choosing the right technology for the use case
    - Communicating cost implications
- Ongoing **support**
    - Migration, Monitoring, Alerting, Capacity management, auditing …

Focus on standards

- Uniformity across on-prem and the cloud
- Focus on **standards** and widely adopted **interfaces**
  - Kubernetes
  - Apache Beam
  - Open Lineage
  - Open Metadata
  - SQL
- Adoption of **open source projects**
  - Supporting complex data intensive applications
  - Machine Learning pipelines

Multi Cloud

- To support continued growth
  - Explored **AWS for serving** and **GCP for offline use cases**
- Challenges of using cloud agnostic vs cloud native solutions
  - For storage, compute and larger use cases
- Data **Movement challenges**
  - Across cloud and onprem
  - Realtime and cost efficient data movement solutions
- **Compliance and Security**
  - Providing uniform constructs for users
- Serving and Processing **near cloud**
  - How to satisfy either needs on different clouds

OnPrem

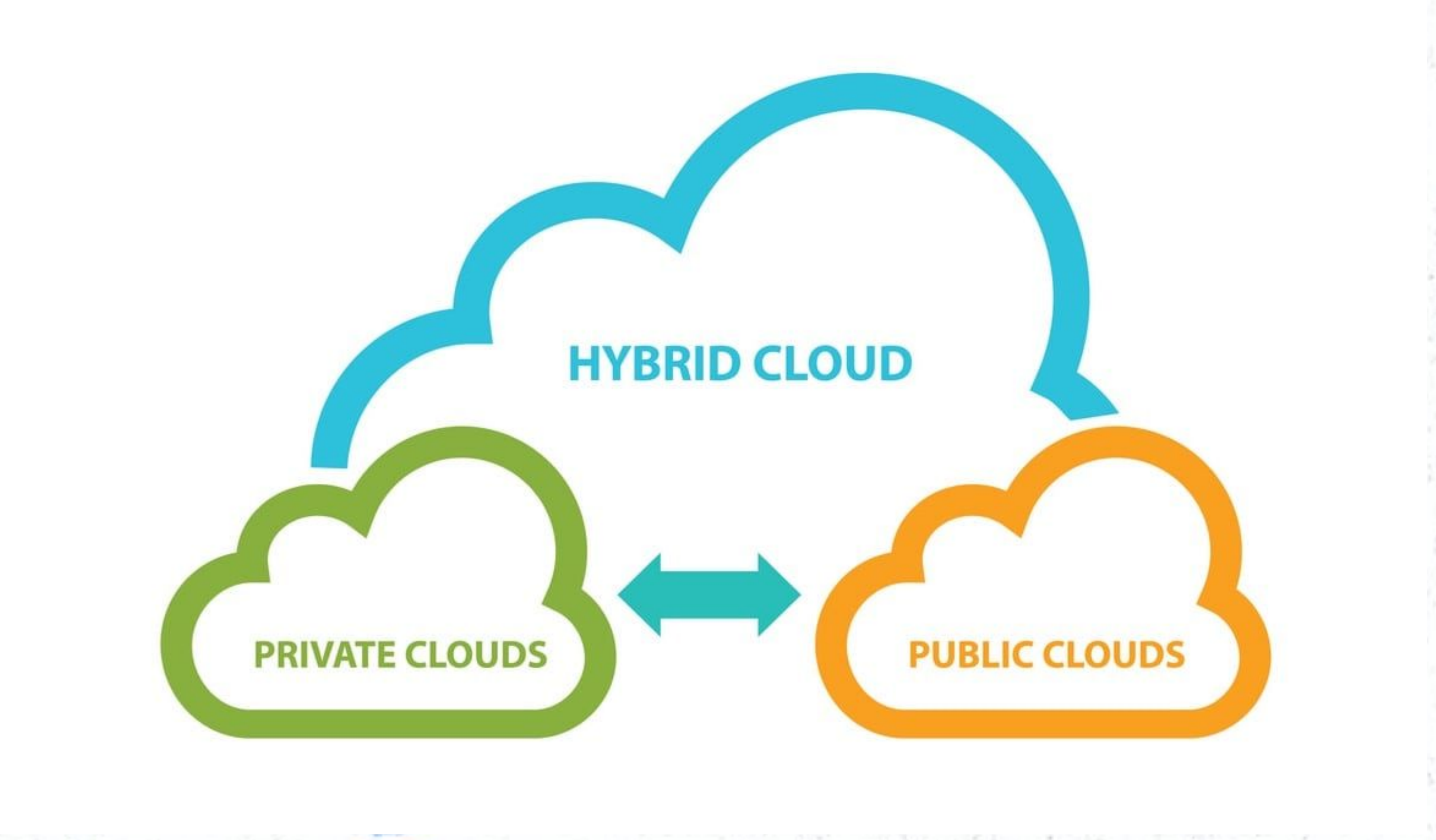
- Build or adopt projects for all parts of infrastructure
- Focus on scaling infrastructure deployments
- Operate and evolve onprem stack with small teams
- Utilize open source where possible

Cloud

- Take advantage of infrastructure in Cloud
- Utilize elasticity, scability and reliability of Cloud
- Increase velocity by adopting new solutions
- Focus on integration projects
- Concentrate on adopting interface and standards

Cloud or Hybrid

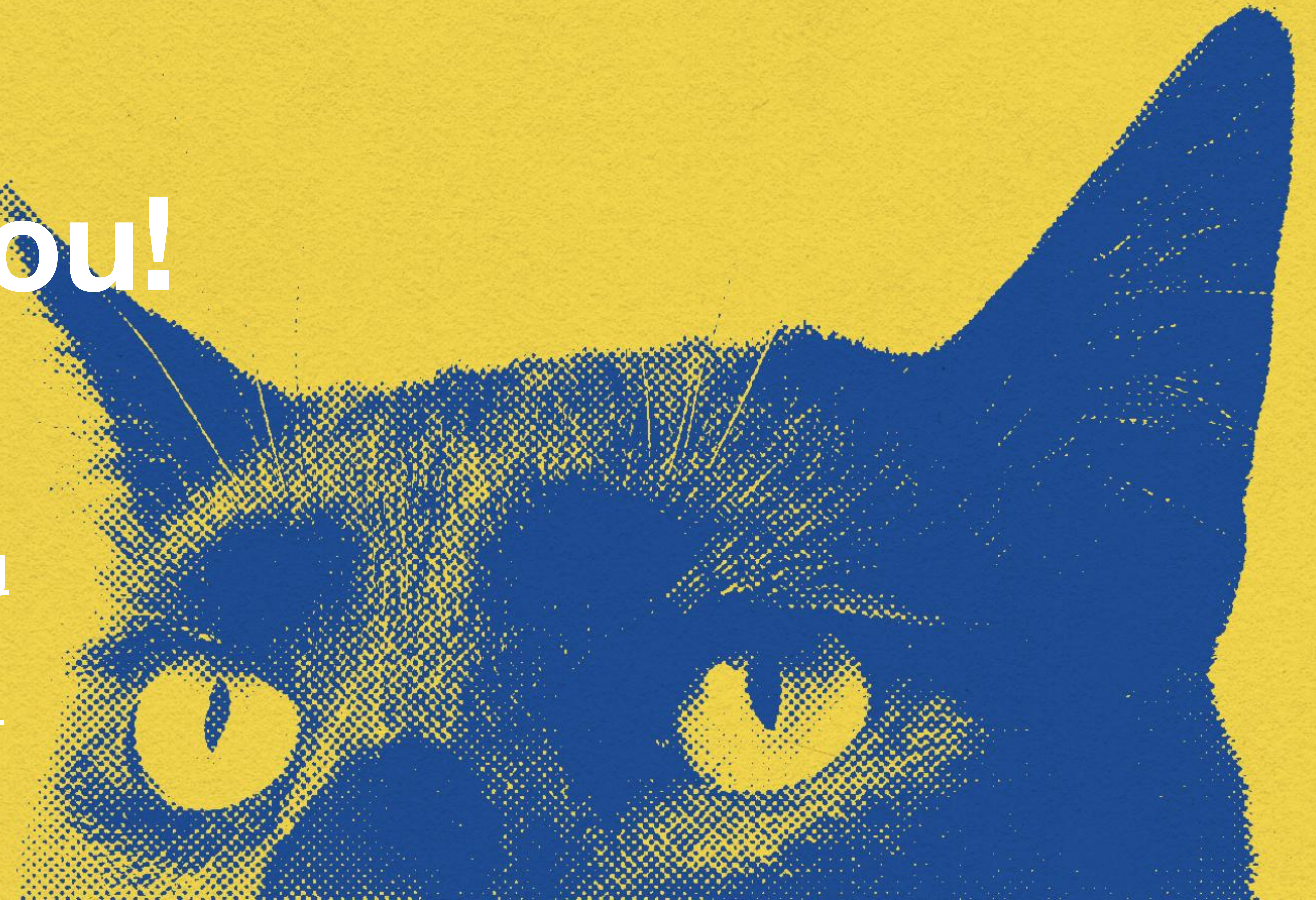
- Choice of all **Cloud** or **Hybrid**
- What is the **future**?
  - Adoption cost
  - Migration cost
  - Maintenance cost
- Where do we **innovate and invest**
  - Build solutions on-prem or hybrid
  - Buy solutions which are cloud specific



Take away

- Use cloud to **solve a specific problem**
  - Understand its limitations
- Migration is the **majority of the work**
  - Projects should take that into account
- **Lots of room for development** efforts and projects
- Adopt **interfaces, standards, and open source**
- More **hybrid cloud awareness** in OSS projects

Thank You!



Follow @TwitterEng
@TwitterCareers
careers.twitter.com