

# Building the Control Plane for Data

Shirshanka Das  
Data Council  
Mar 28, 2023  
Austin, Texas


---

Hello!



## Shirshanka Das

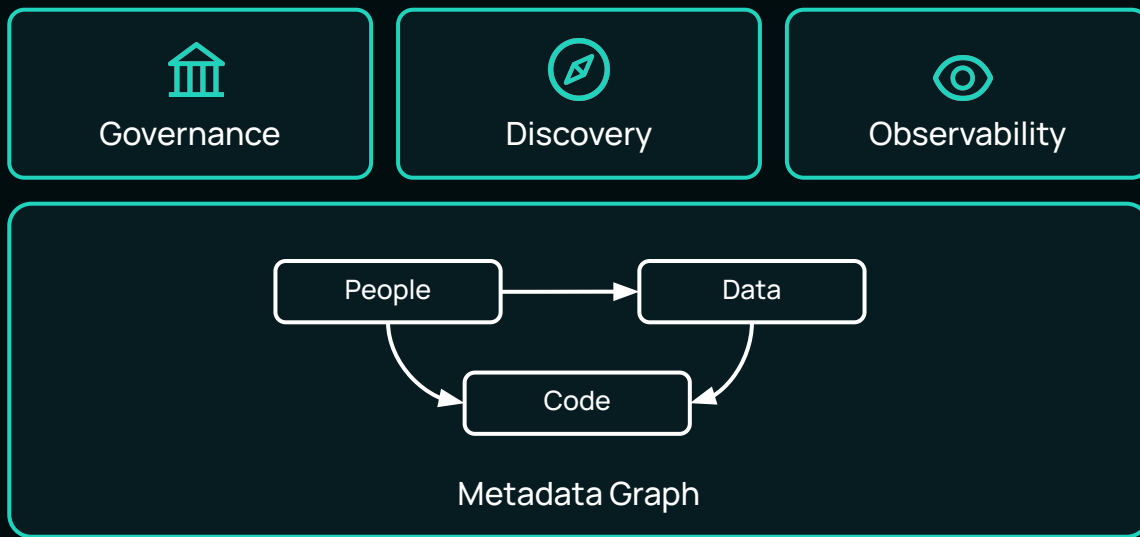
CEO and Co-Founder, Acryl Data  
Founder, DataHub Project, ex-LinkedIn

 @shirshanka

# What is DataHub Project?

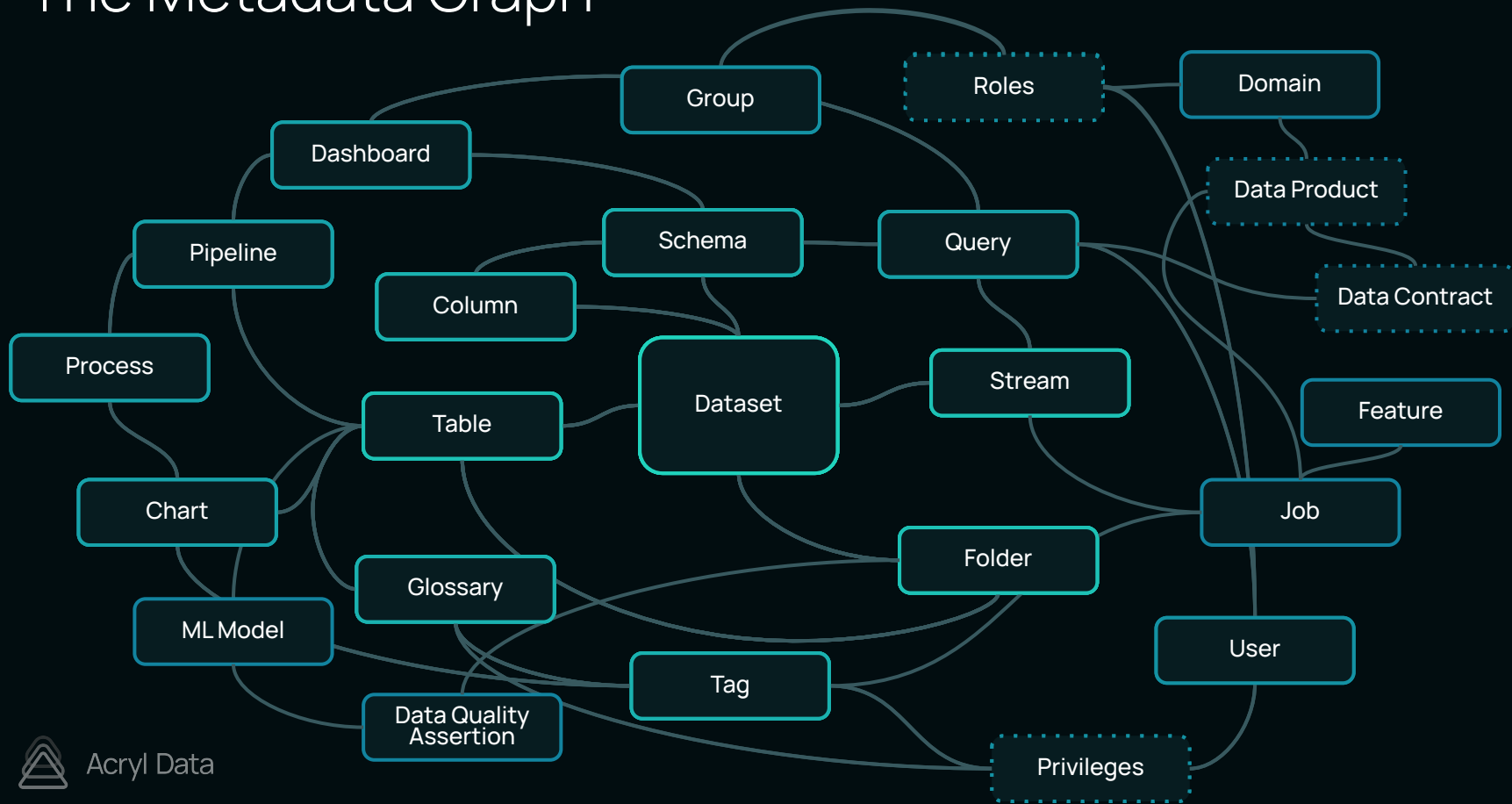
DataHub Project is an open source metadata platform that enables Data Discovery, Data Observability, and Federated Governance on top of a high-fidelity Metadata Graph.

**Acryl Data** is the company advancing the DataHub Project.



Learn more  
[datahubproject.io](https://datahubproject.io)

# The Metadata Graph





# The Application

The screenshot displays the Acryl DataHub application interface. The browser address bar shows the URL `longtailcompanions.acryl.io/search?page=1&query=p&unionType=0`. The search bar contains the query `pet`. The interface is divided into a left sidebar and a main content area.

**Filter Panel (Left Sidebar):**

- Type:** Datasets (14)
- Platform:** Snowflake (9), dbt (4), MongoDB (1)
- Tag:** pet (3), prod\_model (3), business\_critical (2), business\_critical (2), all (1)
- Glossary Term:** Tier 1 (14), Foster Rate (1), Return Rate (1), HighlyConfidential (1)
- Domain:** Marketing (1), E-Commerce (1)
- Owned By:** Shannon Lovett (3), Adoption (2)

**Main Content Area:**

Showing 1 - 10 of 14 results

**Table 1:** dbt & Snowflake > LONG\_TAIL\_COMPANIONS > ANALYTICS  
**PET\_DETAILS** View in Snowflake →  
test  
Marketing | Tier 1 | business critical  
43,517 rows, 16 columns, 2.98 MB, 71 queries last month- High, 1 unique users- High  
Changed 6 months ago  
Matches column pet\_fk  
pet\_details | PET\_DETAILS

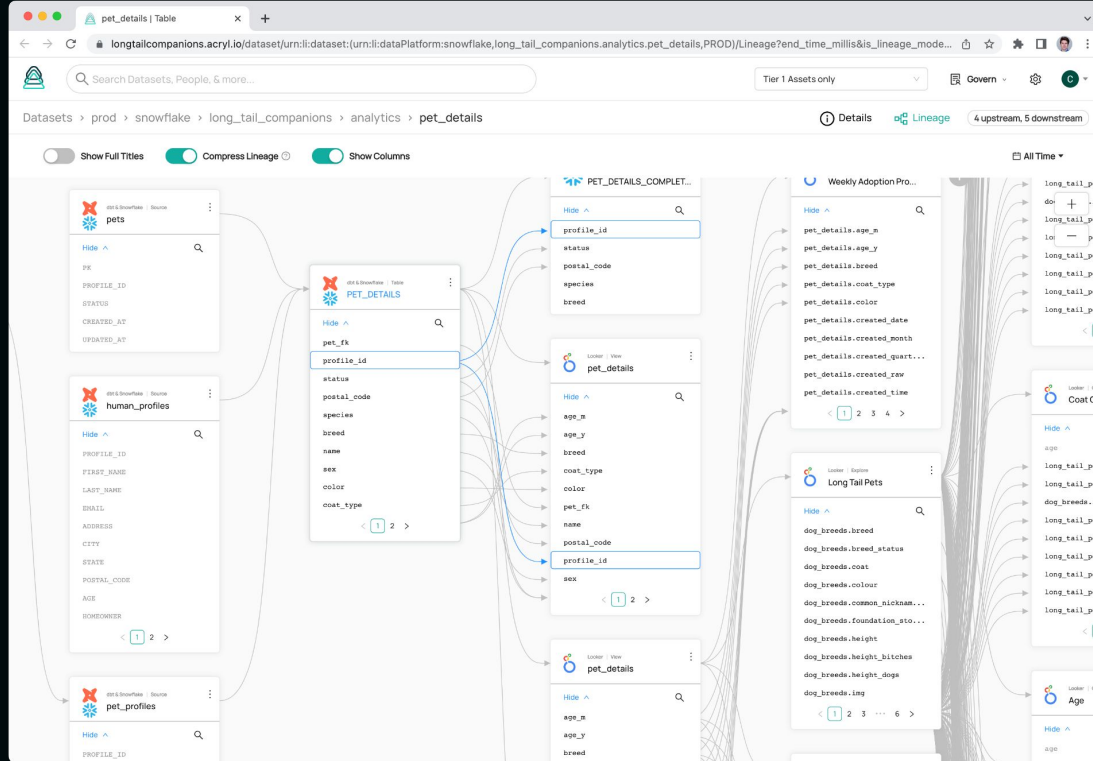
**Table 2:** Snowflake > LONG\_TAIL\_COMPANIONS > ANALYTICS  
**PET\_DETAILS\_COMPLETE** View in Snowflake →  
Tier 1  
45,517 rows, 5 columns, 870.4 KB, 48 queries last month- High, 1 unique users- Med  
Changed 4 months ago

**Table 3:** dbt & Snowflake > LONG\_TAIL\_COMPANIONS > ANALYTICS  
**PET\_STATUS\_HISTORY** View in Snowflake →  
Tier 1  
43,517 rows, 3 columns, 522.24 KB, 32 queries last month- High, 1 unique users- Med  
pet\_status\_history | PET\_STATUS\_HISTORY

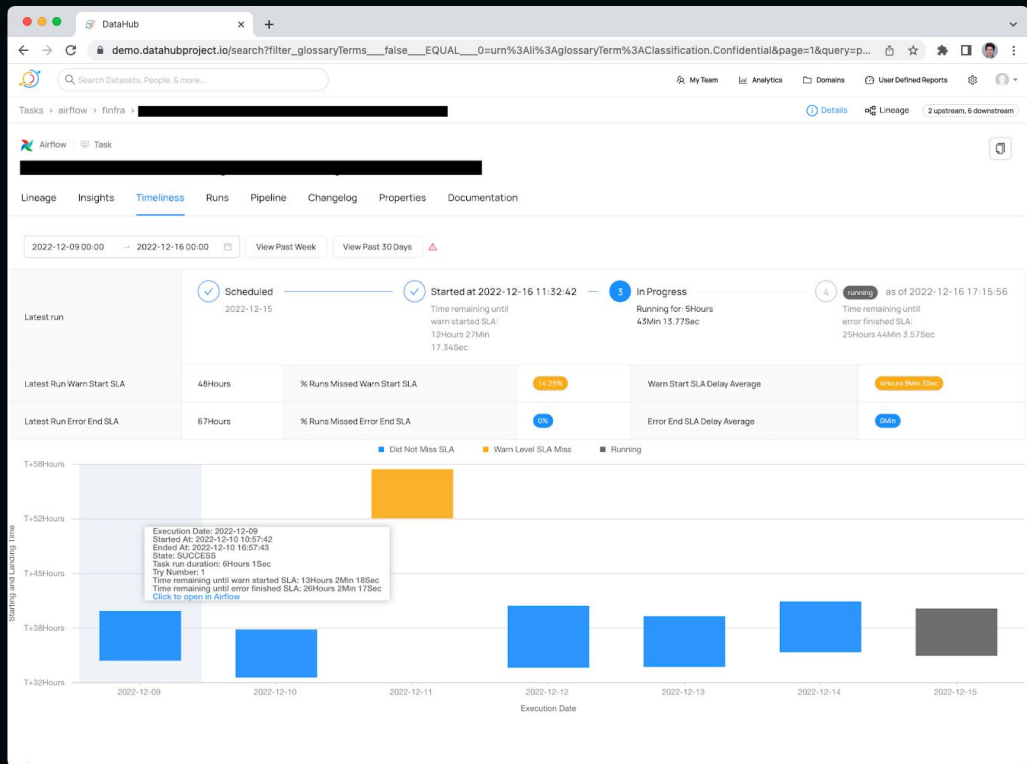
**Table 4:** dbt & Snowflake > LONG\_TAIL\_COMPANIONS > ADOPTION  
**PET\_PROFILES** View in Snowflake →  
My documentation

**Top Users:** swaroop

# The Application



# The Application



# DataHub is the #1 Open Source Metadata Platform



datahub-project / datahub

## Integrations

A grid of 28 logos representing various data integrations. The logos are arranged in five rows: the first four rows have seven logos each, and the fifth row has six logos followed by a '+ more...' text. The logos include Tableau, Snowflake, Databricks, Amazon Redshift, Google Cloud, Microsoft Azure, and many others.

## Adopters

A grid of logos for companies that have adopted DataHub. The logos are arranged in four rows: the first three rows have four logos each, and the fourth row has four logos followed by 'and more...'. The logos include Pinterest, Udemy, Peloton, LinkedIn, Zynga, Wikimedia, Expedia, ThoughtWorks, Klarna, Stripe, Optum, Adevinta, Viasat, Notion, MoLoco, and others.

What does this have to do  
with the control plane of  
data?

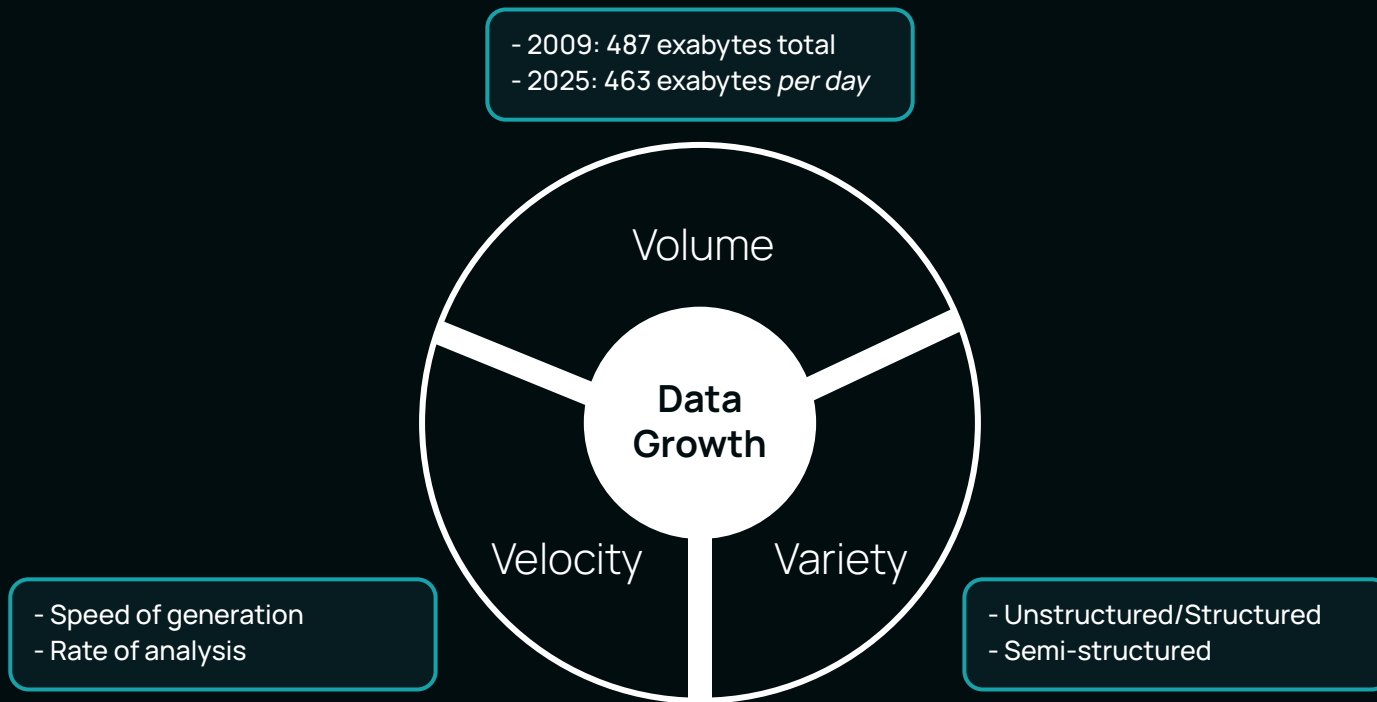
*a.k.a “Am I in the right room?”*



# How We Got to Now

A Brief History of Data

# Major shifts in data over the last decade



# Data Tools in 2012



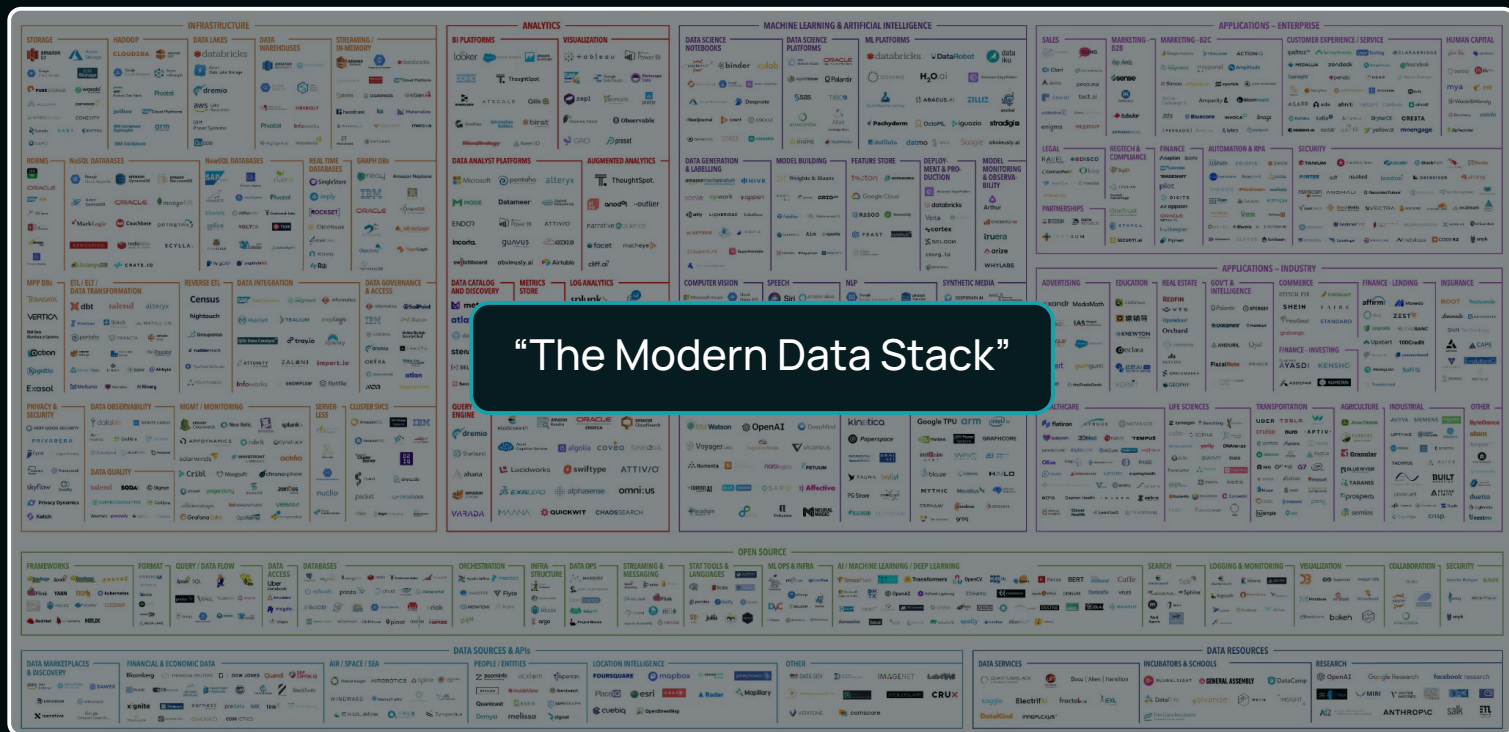


# Data Tools in 2021

The image is a large grid of data tool logos, organized into several main categories:

- INFRASTRUCTURE:** Includes STORAGE (e.g., Amazon S3, Microsoft Azure), HADOOP (e.g., Cloudera, Databricks), DATA LAKES, DATA WAREHOUSES, STREAMING/IN-MEMORY, NOSQL DATABASES, REAL TIME DATABASES, and GRAPH DBs.
- ANALYTICS:** Includes BI PLATFORMS (e.g., Tableau, Power BI), VISUALIZATION, DATA SCIENCE NOTEBOOKS, DATA SCIENCE PLATFORMS, ML PLATFORMS, DATA ANALYTICS PLATFORMS, AUGMENTED ANALYTICS, DATA CATALOG AND DISCOVERY, METRICS STORE, LOG ANALYTICS, QUERY ENGINE, SEARCH, and DATA SOURCES & APIs.
- MACHINE LEARNING & ARTIFICIAL INTELLIGENCE:** Includes DATA SCIENCE PLATFORMS, ML PLATFORMS, DATA GENERATION & LABELING, MODEL BUILDING, FEATURE STORE, DEPLOY, MONITORING & OBSERVABILITY, COMPUTER VISION, SPEECH, NLP, SYNTHETIC MEDIA, HORIZONTAL AI, GPU DBS & CLOUD, and AI HARDWARE.
- APPLICATIONS - ENTERPRISE:** Includes SALES, MARKETING B2B, MARKETING B2C, CUSTOMER EXPERIENCE / SERVICE, HUMAN CAPITAL, LEGAL, REGTech & COMPLIANCE, FINANCE, AUTOMATION & RPA, SECURITY, PARTNERSHIPS, ADVERTISING, EDUCATION, REAL ESTATE, COMMERCE, FINANCE-LENDING, INSURANCE, HEALTHCARE, LIFE SCIENCES, TRANSPORTATION, AGRICULTURE, INDUSTRIAL, and OTHER.
- OPEN SOURCE:** Includes FRAMEWORKS, FORMAT, QUERY / DATA FLOW, DATA ACCESS, DATABASES, ORCHESTRATION, META-STRUCTURE, DATA OPS & MTPA, STREAMING & MESSAGING, START TOOLS & LANGUAGES, AI / MACHINE LEARNING / DEEP LEARNING, SEARCH, LOGGING & MONITORING, VISUALIZATION, COLLABORATION, SECURITY, DATA MARKETPLACES & DISCOVERY, FINANCIAL & ECONOMIC DATA, AIR / SPACE / SEA, PEOPLE / ENTITIES, LOCATION INTELLIGENCE, OTHER, DATA SERVICES, INCUBATORS & SCHOOLS, and RESEARCH.

# Data Tools in 2021



---

# What got easier?



## Data Storage

blob stores, stream stores, nosql stores



## Data Movement

ELT, ETL, Reverse ETL



## Data Querying

disaggregated compute, federated querying, shared nothing systems



## Data Visualization

self-serve BI, notebooks

# What got easier?



## Orchestrators

DAGs, Tasks, Pipelines



## Transformation

SQL → SQL + friends



## Real Time

Streaming compute,  
correctness guarantees



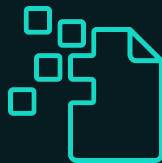
## AI Infra

Frameworks, Cloud-based  
AI platforms, Federated  
learning

# What got harder?



Data Discovery



Data Quality



Data Management

---

# What got harder?



Data Discovery



Data Quality



Data Management

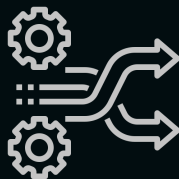
# What problems emerged in Data Discovery?



Physical Metadata is  
not intuitive to  
everyone



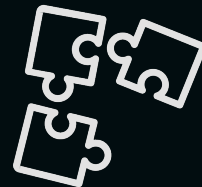
Crawl-Only  
Ingestion Leads to  
Stale Metadata



No approach to  
acting on changes  
in metadata



Manual enrichment  
of metadata leads to  
problems



Over-indexed on  
Data Warehouses

# Solving for Data Discovery

## Metadata 360



Combine *technical* and  
*business* metadata

## Shift Left



Declare & collect metadata  
at the source

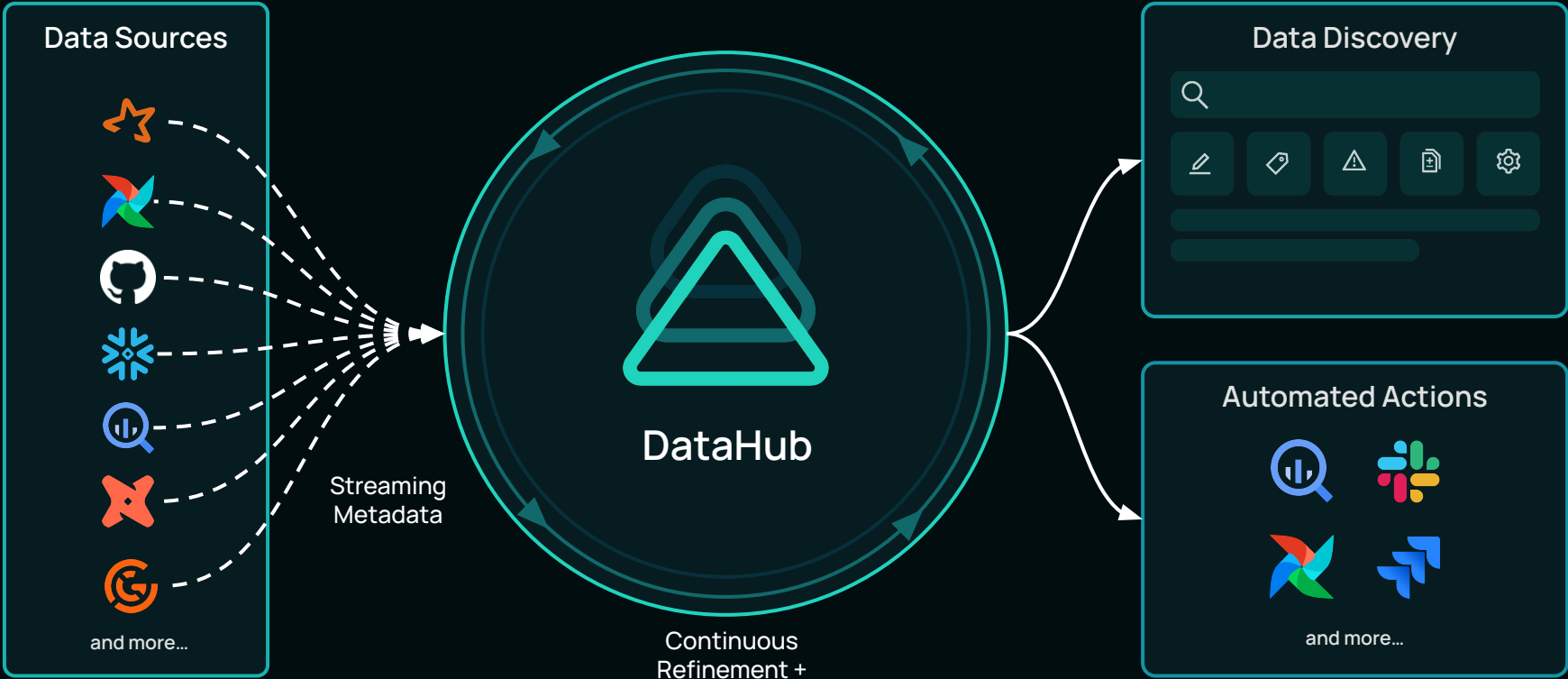
## Active Metadata



Inject metadata into the  
operational plane



# DataHub Architecture

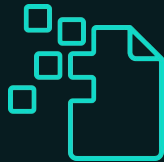


---

# What got harder?



Data Discovery

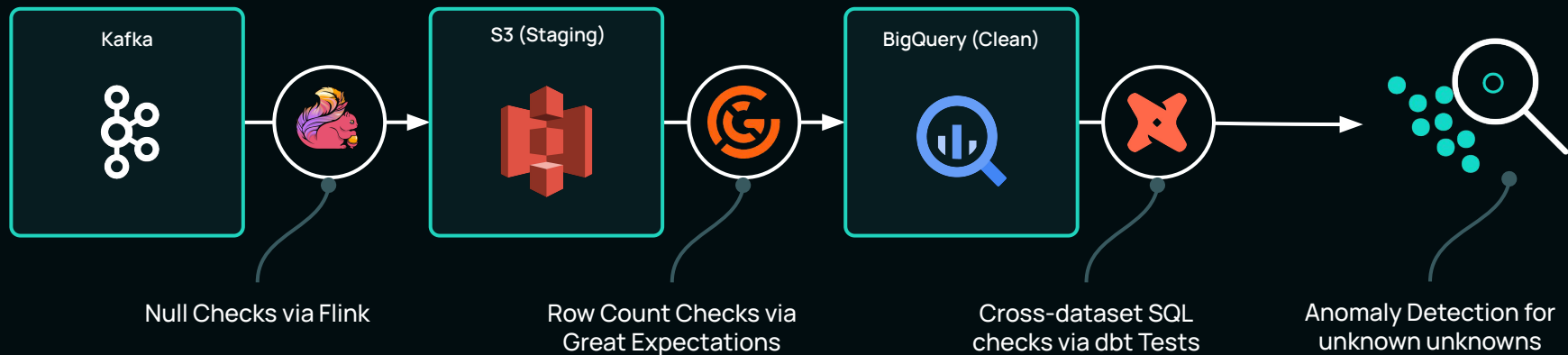


Data Quality



Data Management

# How do we manage data quality today?



---

# Data Quality - Broken Glass

- Data quality is checked inconsistently across different tools
- Not sharing outputs of checks in a uniform way
- Pipelines cannot operate on data that meets the quality bar with confidence

# What got harder?



Data Discovery



Data Quality



Data Management

# What does it take to set up access management *once*?

```
"Effect": "Allow",
"Action": [
  "rds:DescribeDBInstances",
  "rds:DescribeDBClusters",
  "rds:DescribeGlobalClusters"
],
"Resource": "*"
```

Define Actions and Tags for attribute-based access control

```
JSON
{
  "rds:DescribeDBInstances",
  "rds:DescribeDBClusters",
  "rds:DescribeGlobalClusters"
},
"Resource": "*"
}
```

Create and review policy in AWS IAM console

Attach Policy

Select one or more policies to attach. Each group can have u

Filter: Policy Type - C-RDS

Create group and attach policy in AWS IAM console

Tag key

Environment

Add another Tag

Specify tags in database in AWS RDS console

Databases

Filter databases

DB identifier

Locate or create database in AWS RDS console

Select AWS access type

Select how these users will access AWS. Access keys and autoger

Access type\*  Programmatic access  
Enables an access key for other development too

AWS Management C  
Enables a password

Create user and associate with group

Creation time 2020-11-04

Permissions Groups (1) Tags (1) Se

IAM tags are key-value pairs you can add to your u  
access for this user. Learn more

Specify tags for user back in AWS IAM console

⊗ User: arn:aws:iam::08588124  
resource: arn:aws:rds:us-east

Test and review permissions and access

Repeat bespoke steps in admin console for each tool

---

# Data Management - Broken Glass

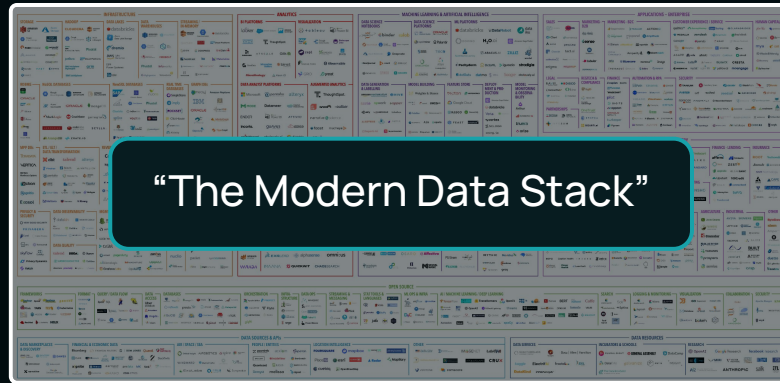
- Access Management
- Masking of sensitive information
- Table-based retention
- Key-based retention
- Data Locality

# How do we solve for this?

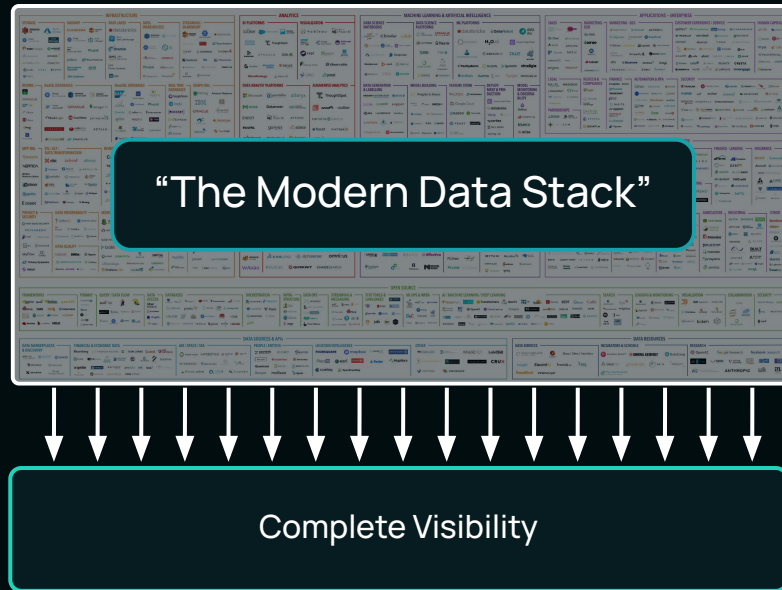
a.k.a *“Is this still about the control plane?”*



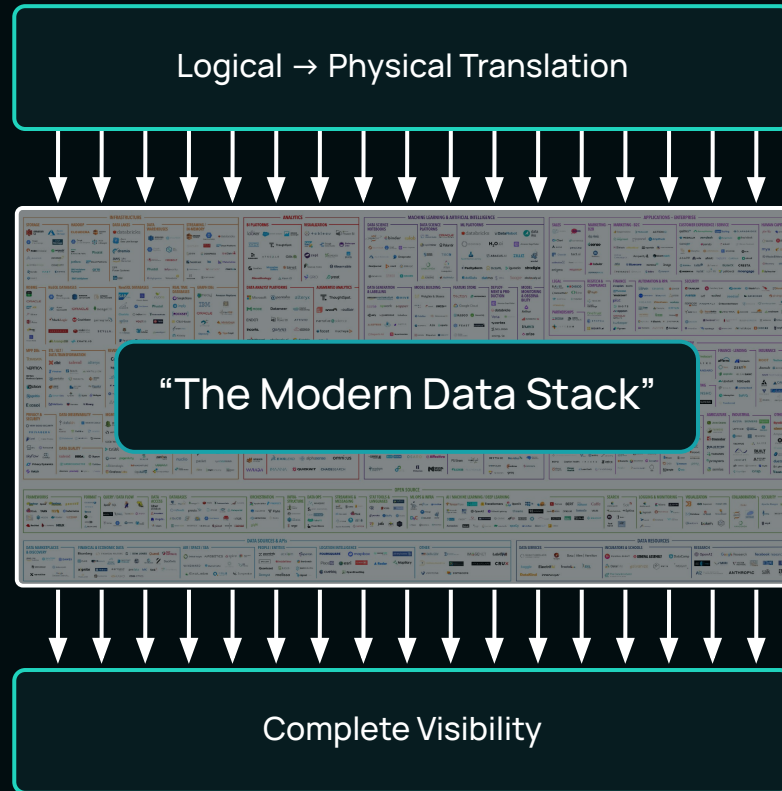
“All problems in Computer Science can be solved by another layer of indirection”



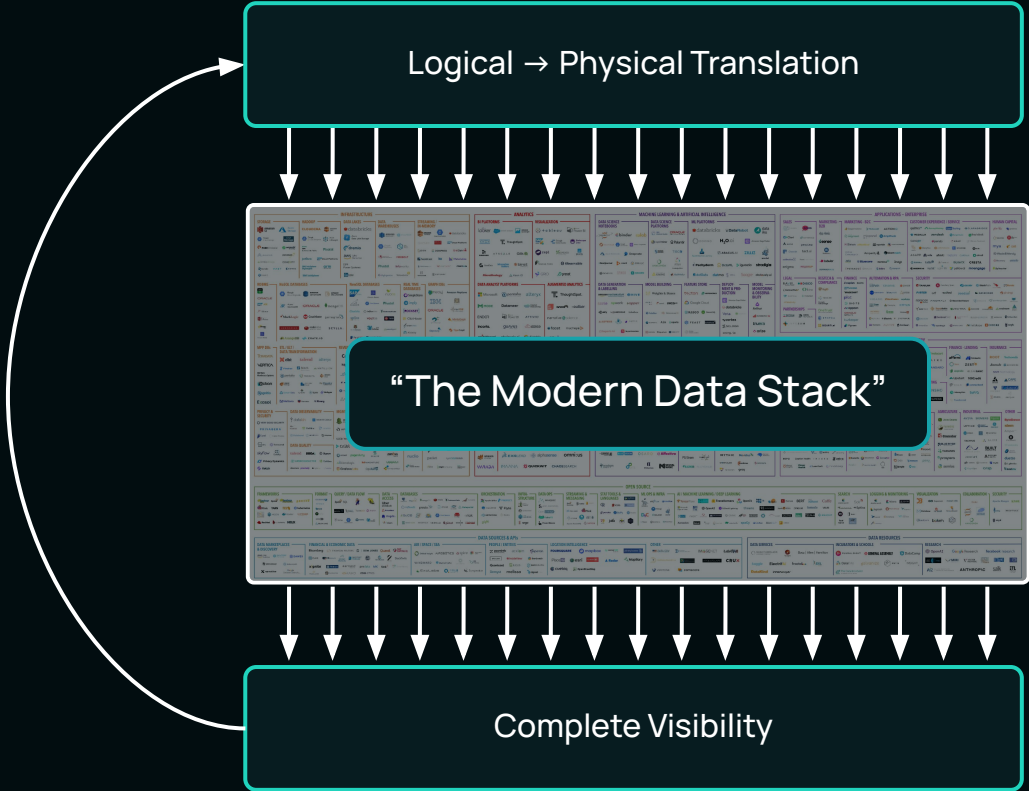
# Data needs a Control Plane



# Data needs a Control Plane



# Data needs a Control Plane



---

# A Control Plane

The control plane resides above the data plane, as a separate entity, and enforces rules for the data plane.

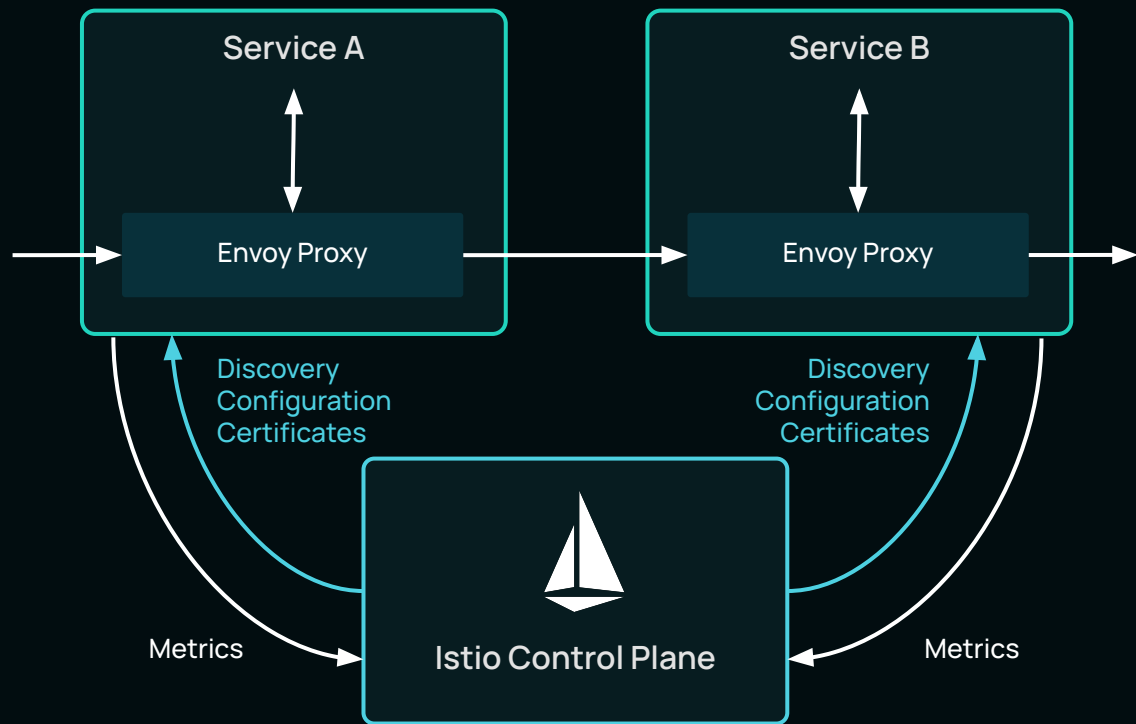
Where have we seen this before?

What have we done?

What did we learn?



# Exhibit 1: The Control Plane for the Service Mesh



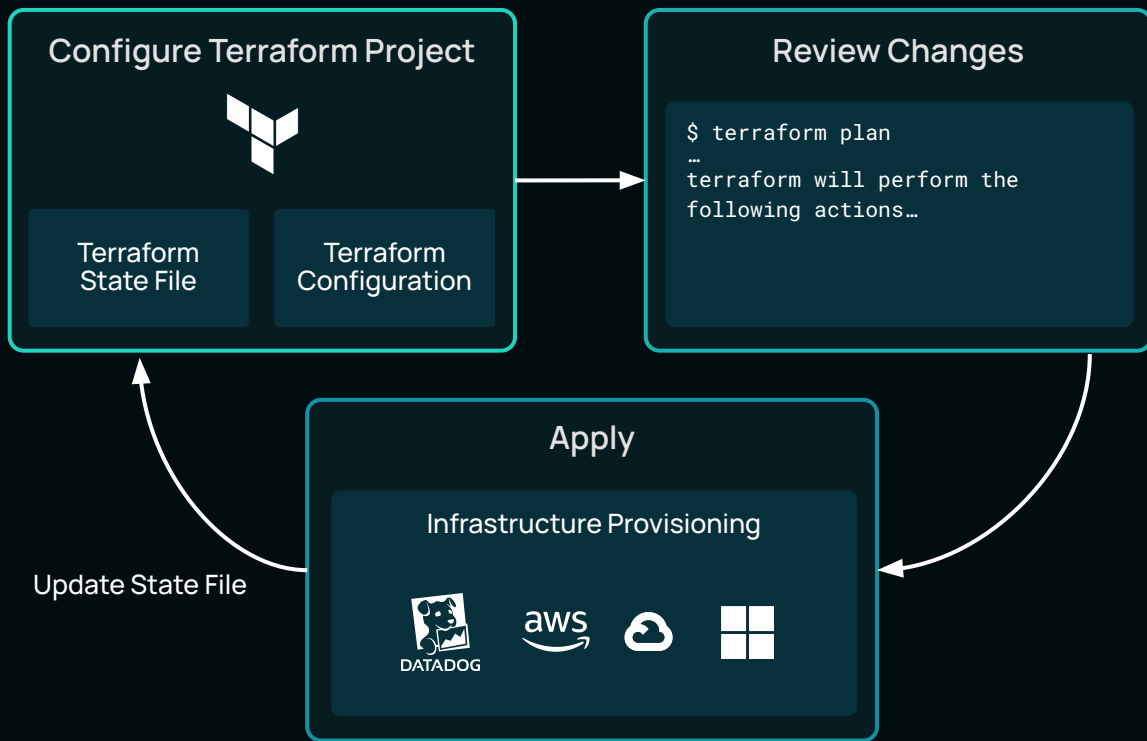
## Istio

Centralizes definition and application of service policies, discovery

Removes manual specification of physical service configuration

Enables fine-grain routing, access control in the service data plane

# Exhibit 2: Infrastructure Provisioning for any Cloud



## Terraform

Centralizes definition and application of infrastructure policies

Removes manual specification of physical instance configuration

Enables fine-grain provisioning in the compute plane



---

# What have we learnt from these systems?

State what you know, let the system figure out the full plan a.k.a. **what not how**

Decentralization should be paired with **policy as code / configuration as code** a.k.a **shift left**

You can get a lot done without requiring standardization upfront a.k.a **adapter pattern works**



# Core Capabilities of a Successful Control Plane

*adapted for Data*



---

# Core Capabilities



## Breadth

Integrate with  
everything

Represent everything

Embrace  
decentralization

---

# Core Capabilities



## Breadth

Integrate with  
everything

Represent everything

Embrace  
decentralization



## Low Latency

Operational use-cases  
demand this

Freshness and  
responsiveness

# Core Capabilities



## Breadth

Integrate with everything

Represent everything

Embrace decentralization



## Low Latency

Operational use-cases demand this

Freshness and responsiveness



## Scale

Drivers: Breadth, Temporality, Versioning

# Core Capabilities



## Breadth

Integrate with everything

Represent everything

Embrace decentralization



## Low Latency

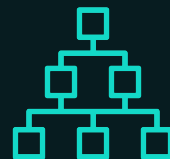
Operational use-cases demand this

Freshness and responsiveness



## Scale

Drivers: Breadth, Temporality, Versioning



## Source of Truth

System of record for logical specifications

# Core Capabilities



## Breadth

Integrate with everything

Represent everything

Embrace decentralization



## Low Latency

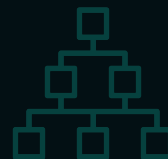
Operational use-cases demand this

Freshness and responsiveness



## Scale

Drivers: Breadth, Temporality, Versioning



## Source of Truth

System of record for logical specifications



## Auditable

Why is the system in its current state?

Who performed what action?

# Core Capabilities



## Breadth

Integrate with everything

Represent everything

Embrace decentralization



## Low Latency

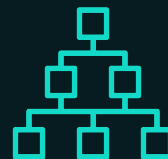
Operational use-cases demand this

Freshness and responsiveness



## Scale

Drivers: Breadth, Temporality, Versioning



## Source of Truth

System of record for logical specifications



## Auditable

Why is the system in its current state?

Who performed what action?

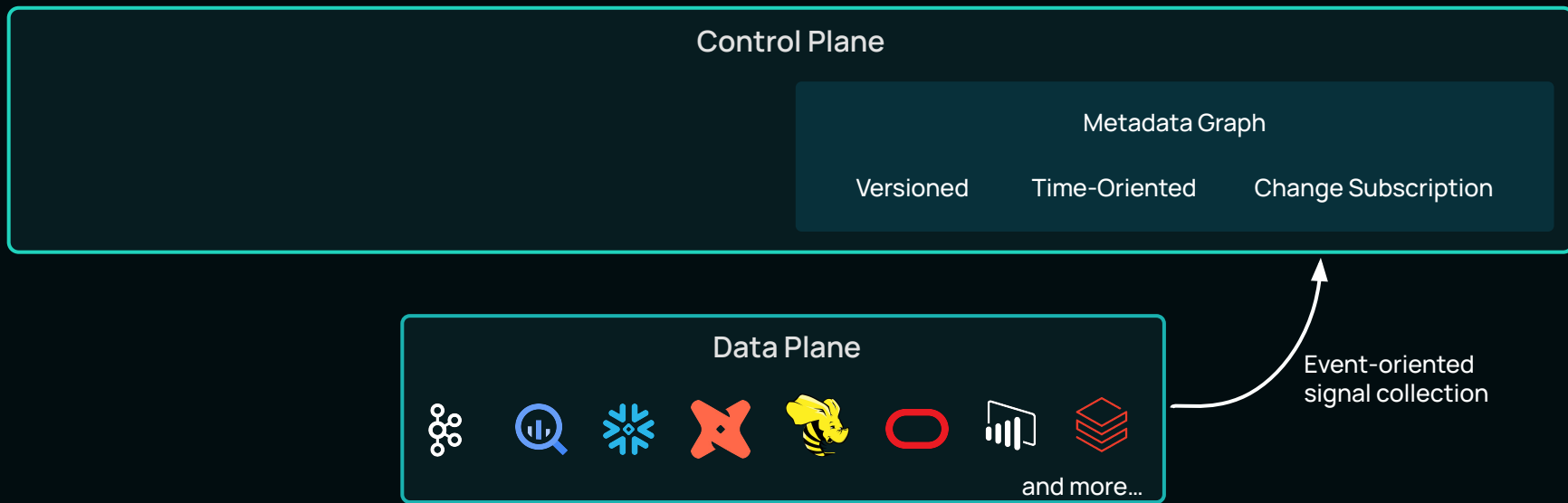




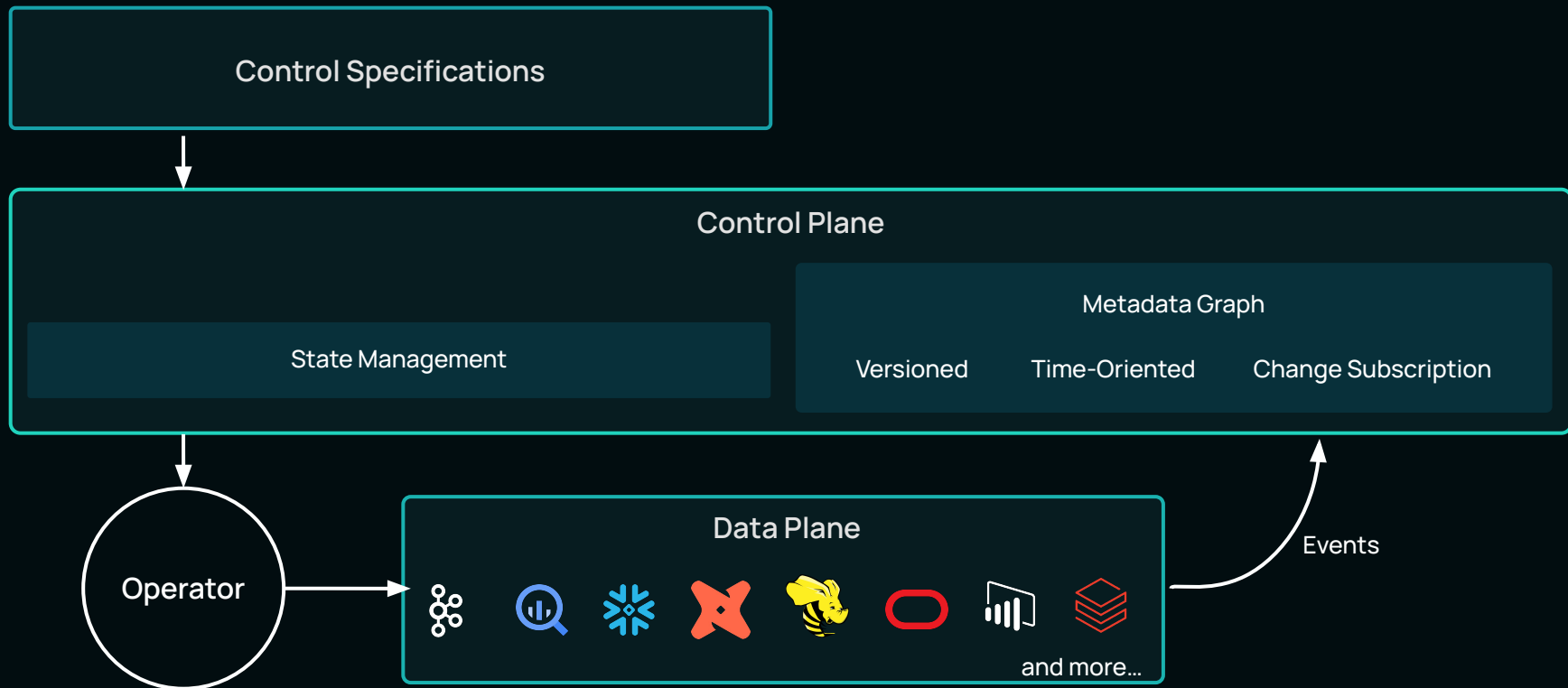
# Architecture Blueprint

*for the control plane for data*

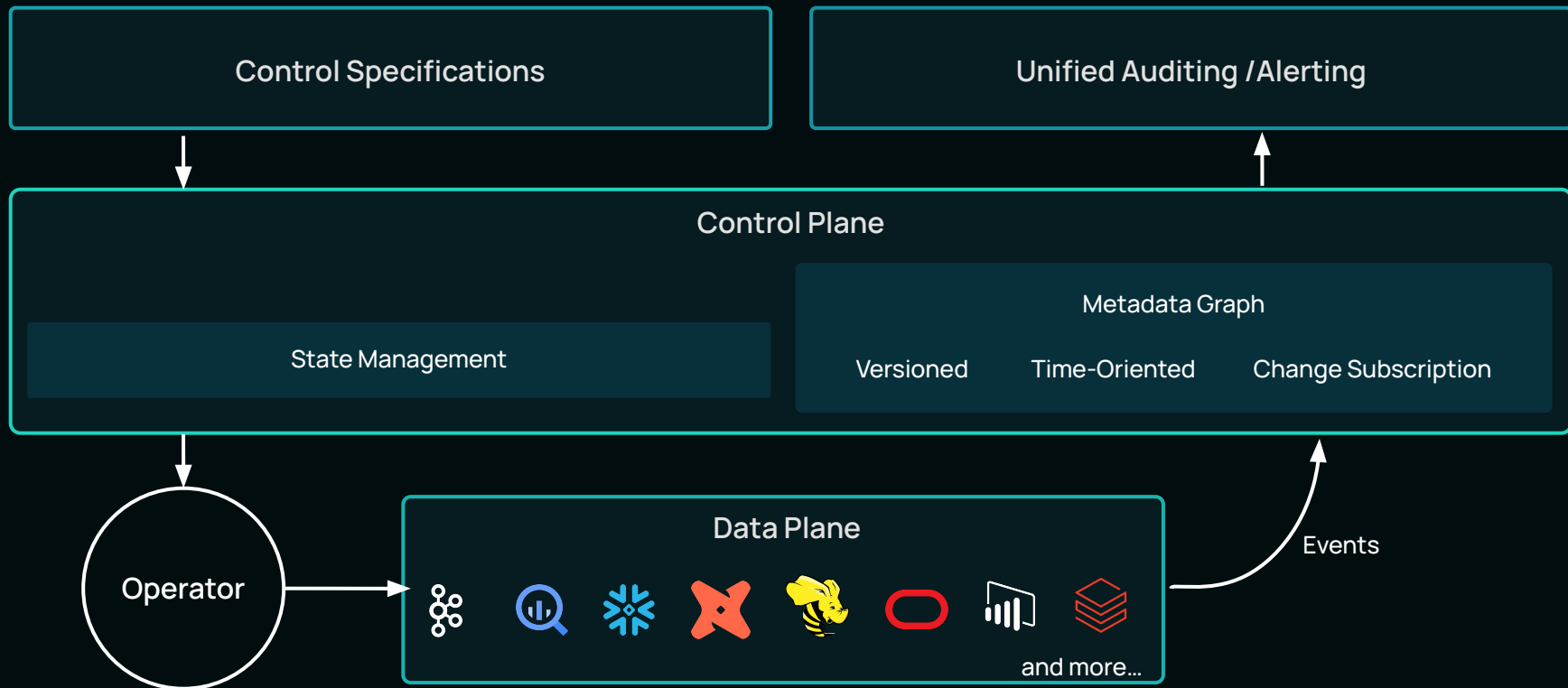
# Solving for real-time visibility and breadth



# Solving for logical $\rightarrow$ physical translation

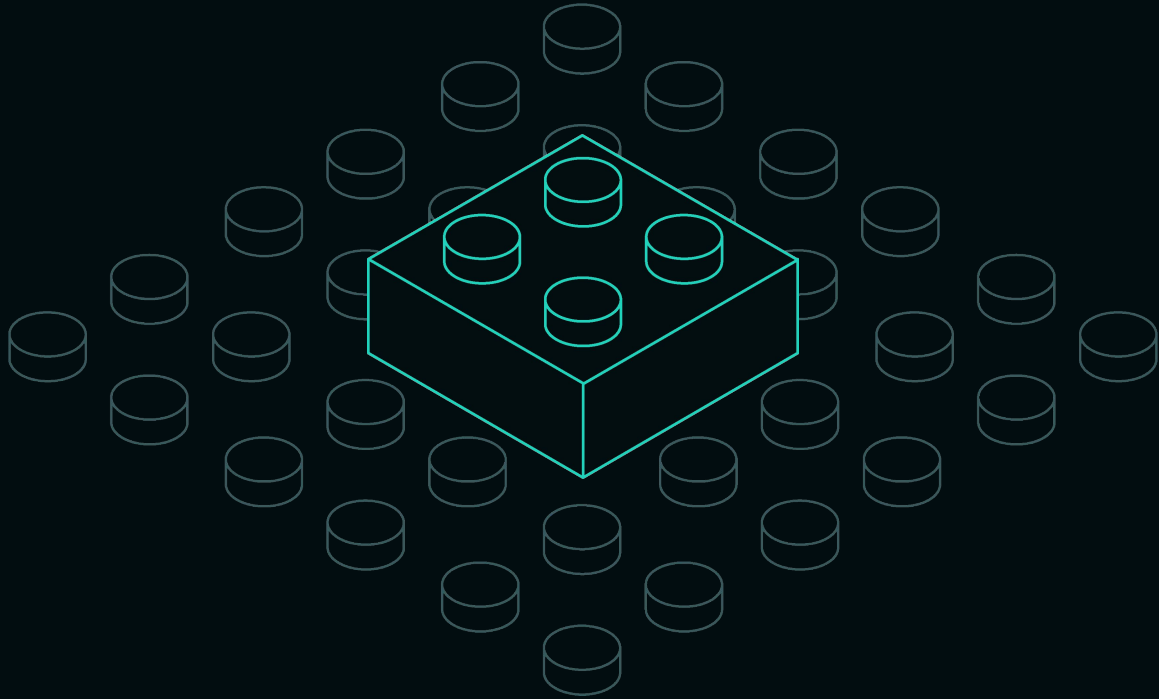


# Putting it all together



---

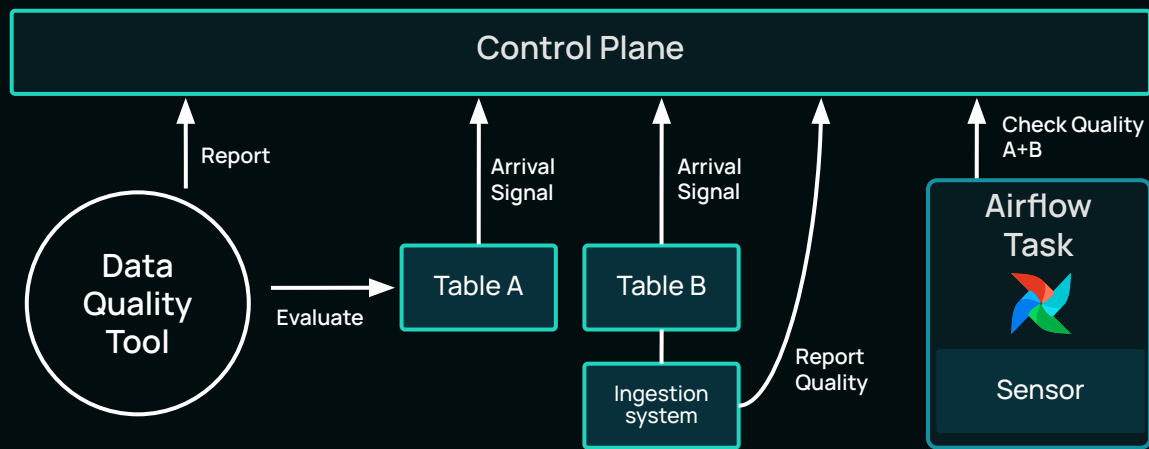
# Use-Cases



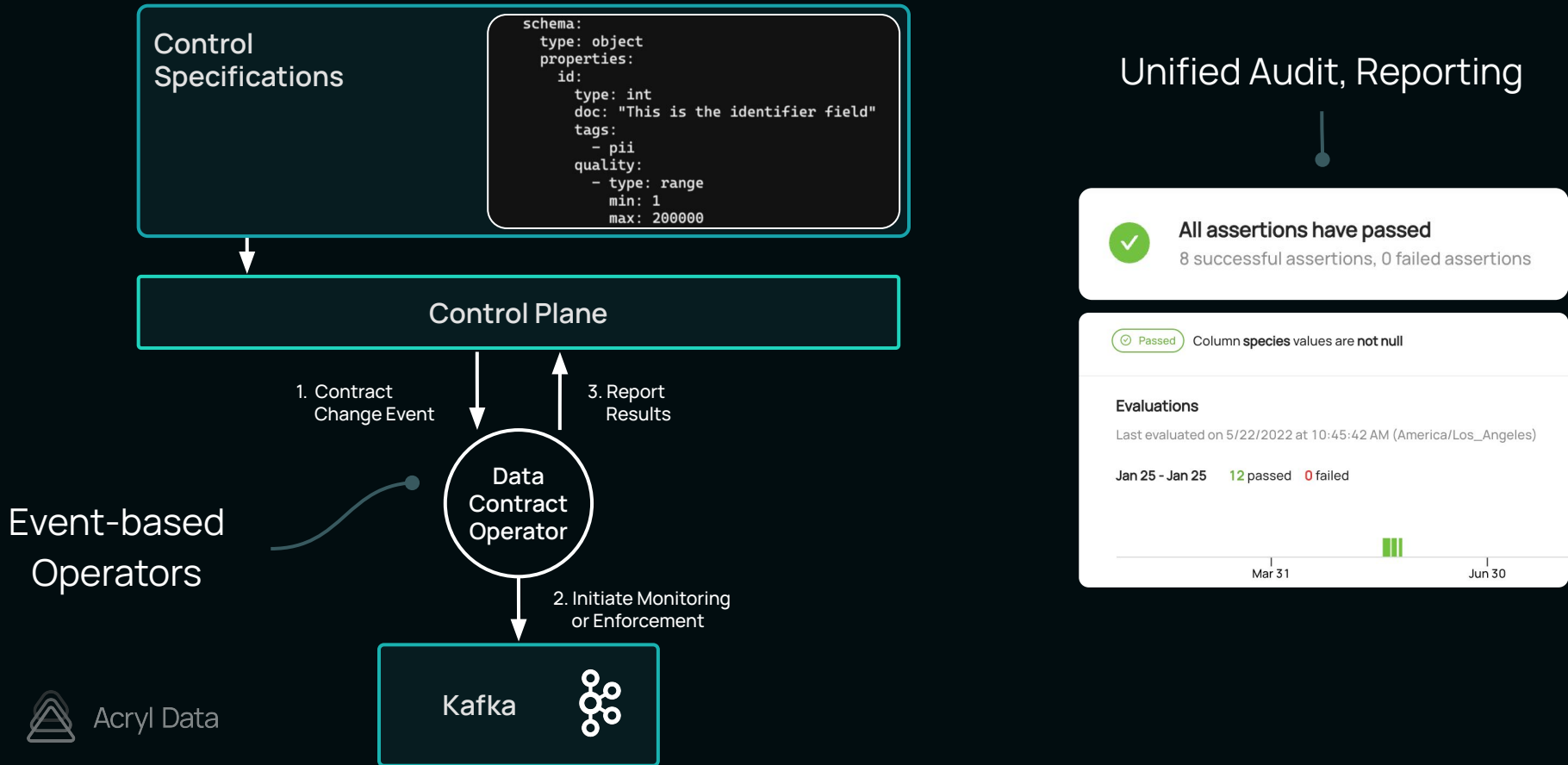
# Use Case 1: Reliable Data Pipelines

Integrate Quality signals across multiple tools

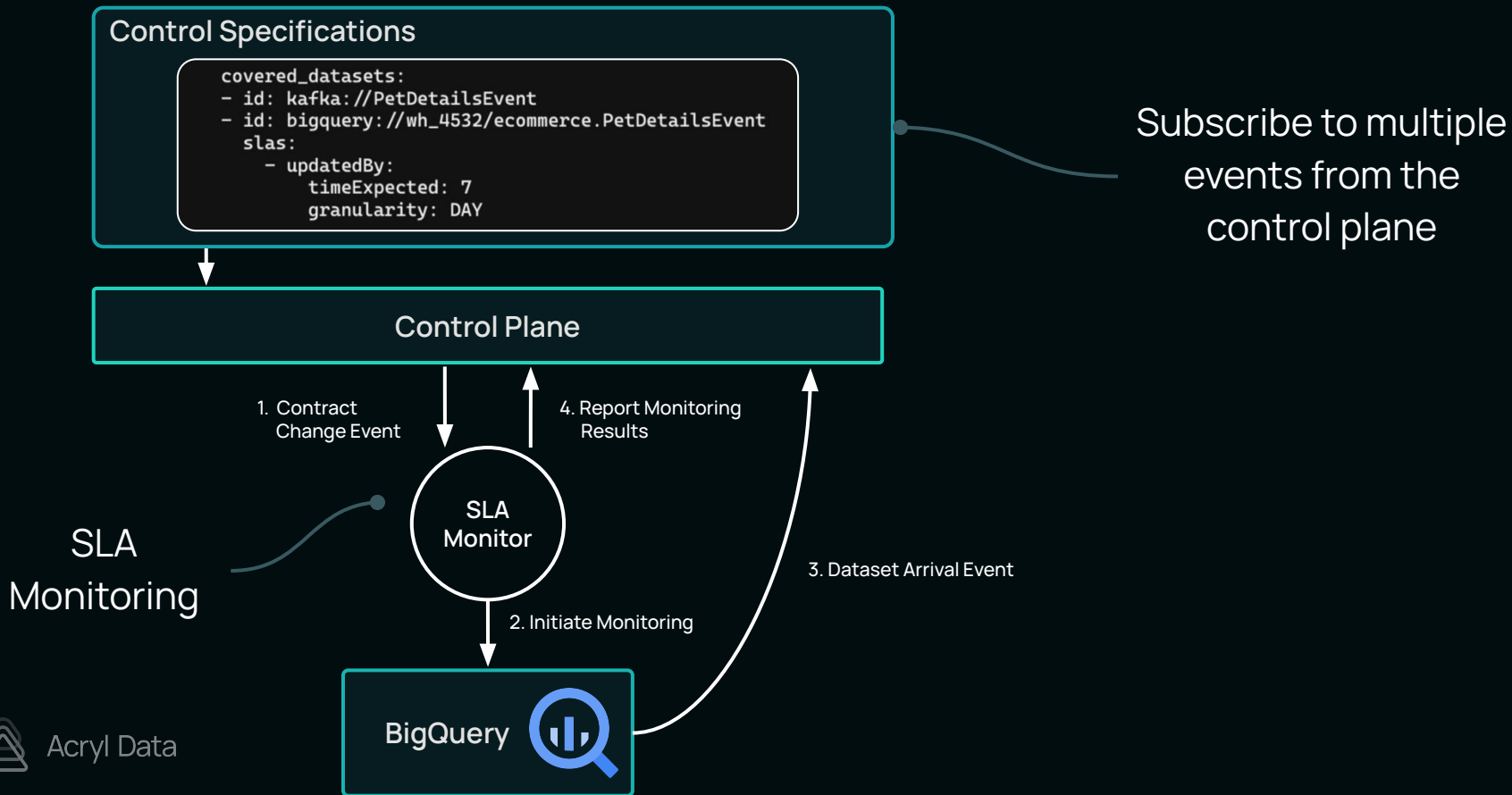
Intelligent computation, triggering based on arrival



# Use Case 2: Data Contract - Quality Enforcement



# Use Case 3: Data Contract - SLA Monitoring

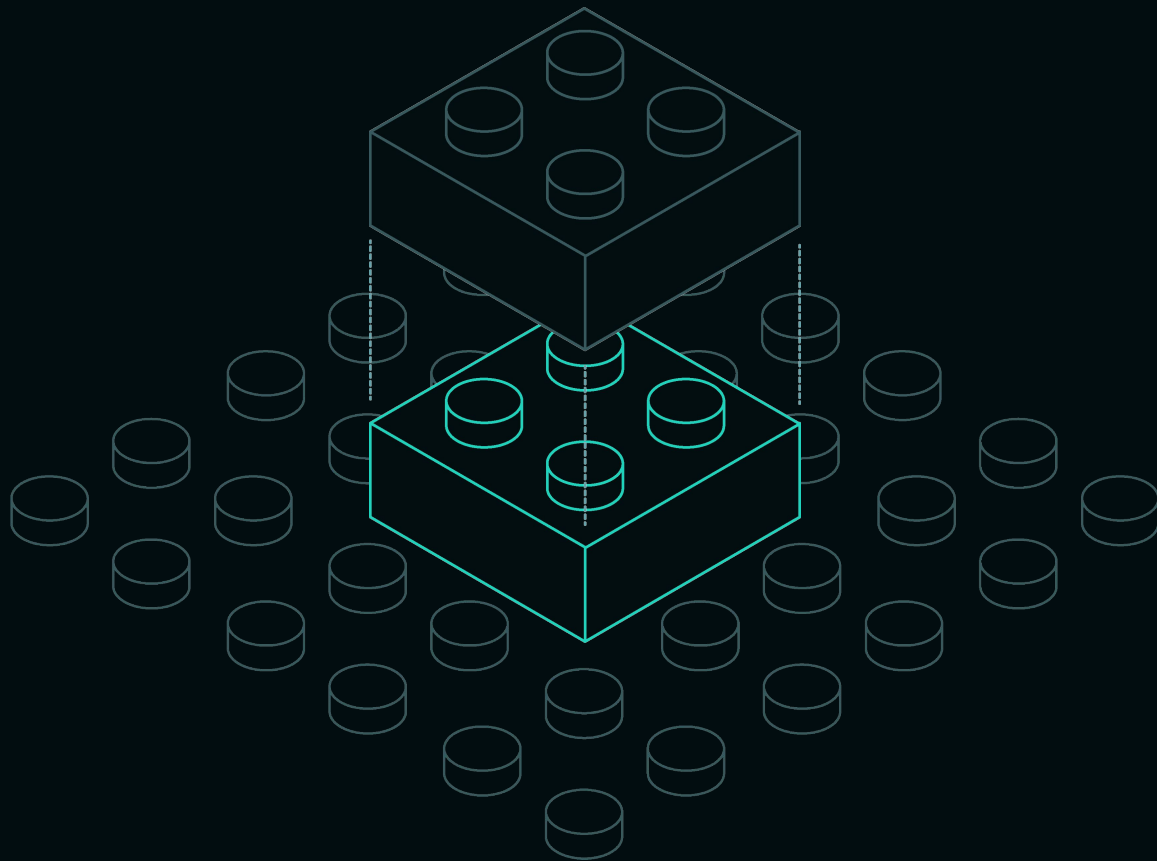






---

What other use-cases could we tackle with this arch?



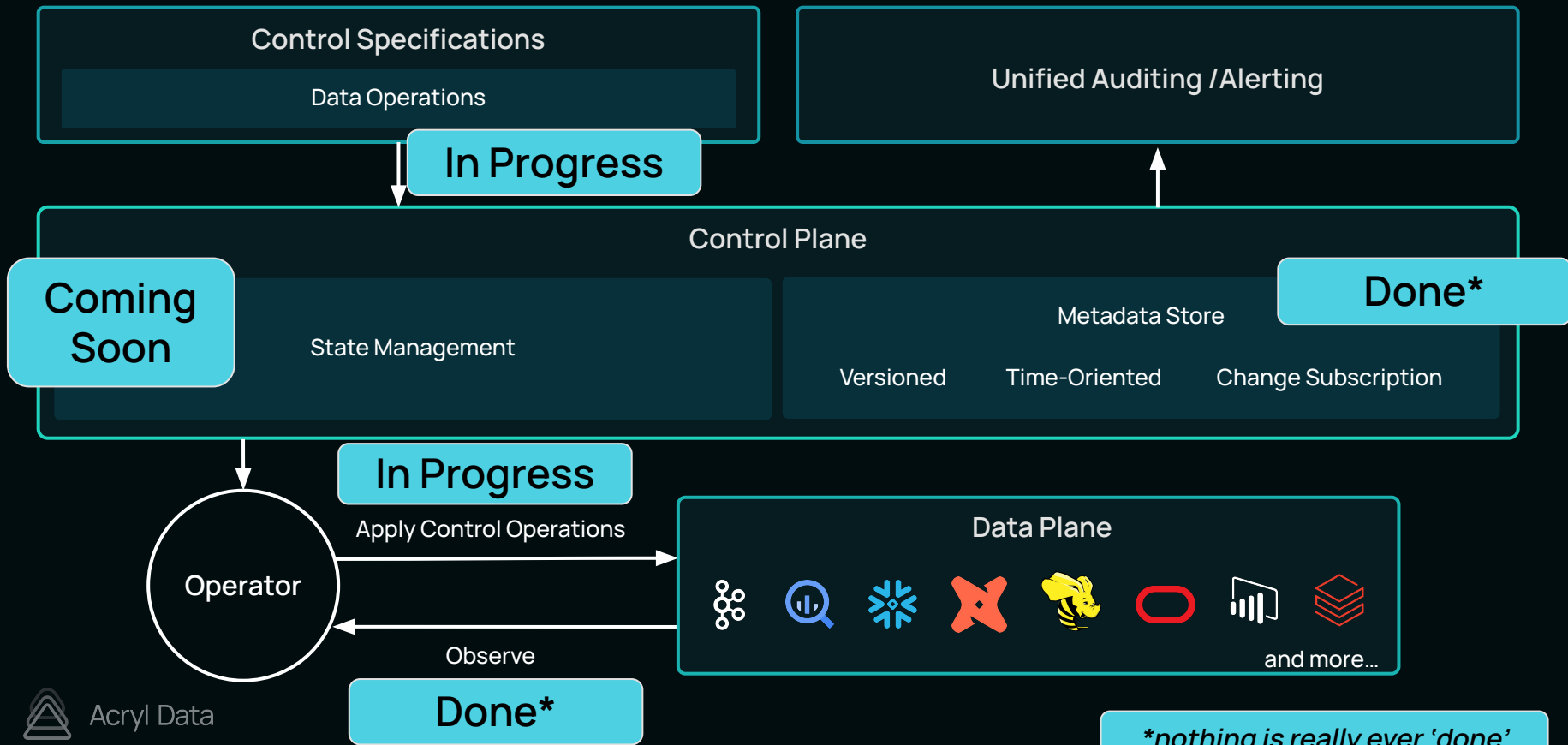
---

# Quite a few!

- Masking of sensitive information
- Table-based retention
- Key-based retention
- Data Locality
- Auto ETL

Are we there yet?

# The Control Plane for Data is Almost Here



*\*nothing is really ever 'done'*

---

# Key Takeaways

- The Data Stack continues to be highly fragmented
- A metadata-driven control plane holds the key to bringing order to the chaos
- Data Quality, Data Governance, Data Management can be harmonized
- We are building towards this vision collaboratively with our partners:  
**Help Acryl build the Control Plane for Data!**

# Join the movement!



[slack.datahubproject.io](https://slack.datahubproject.io)



[acryldata.io/sign-up](https://acryldata.io/sign-up)  
[acryldata.io/careers](https://acryldata.io/careers)



Acryl Data