Metadata, Data Provenance, Metadata, and looking to the future with Data Mesh

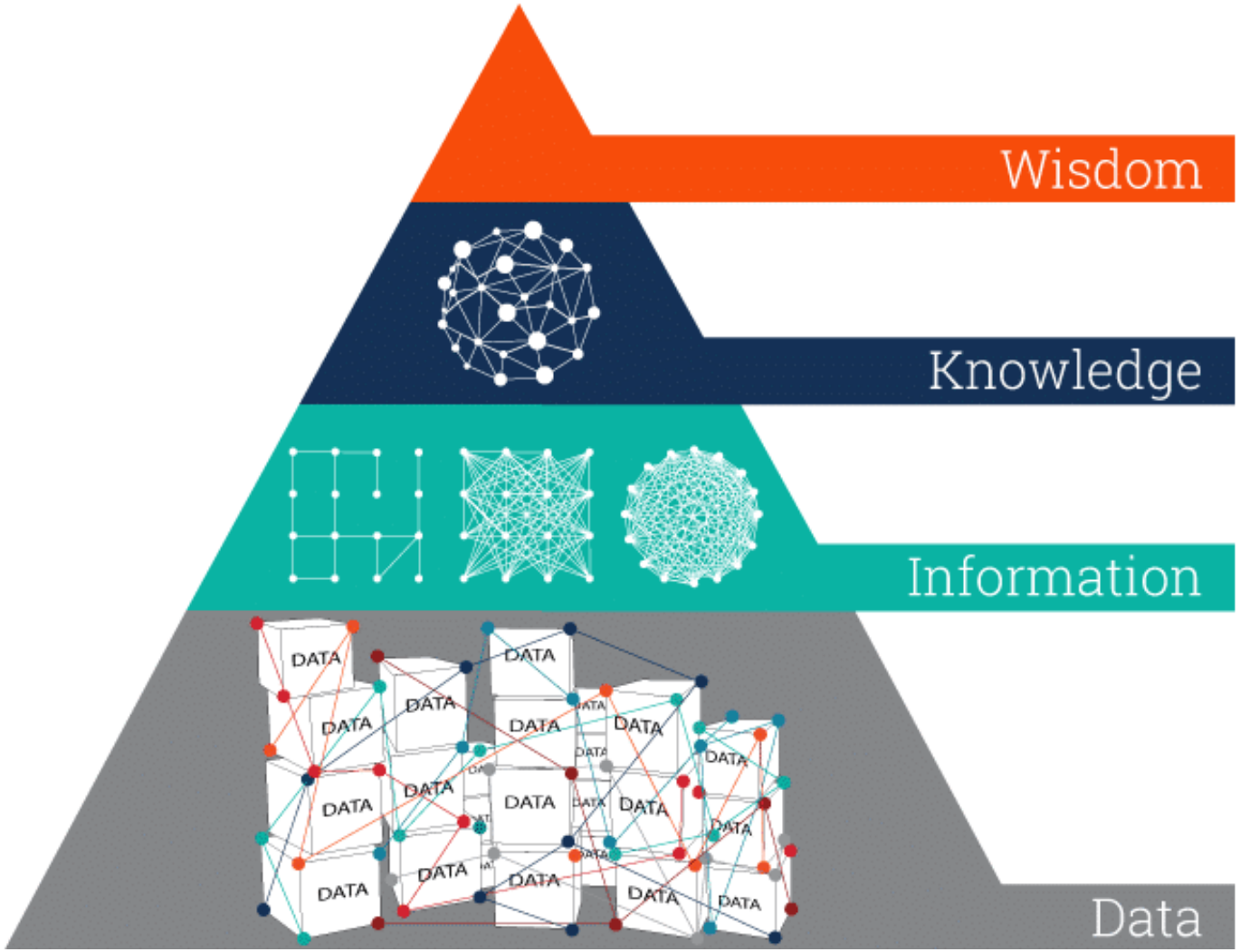
SDU

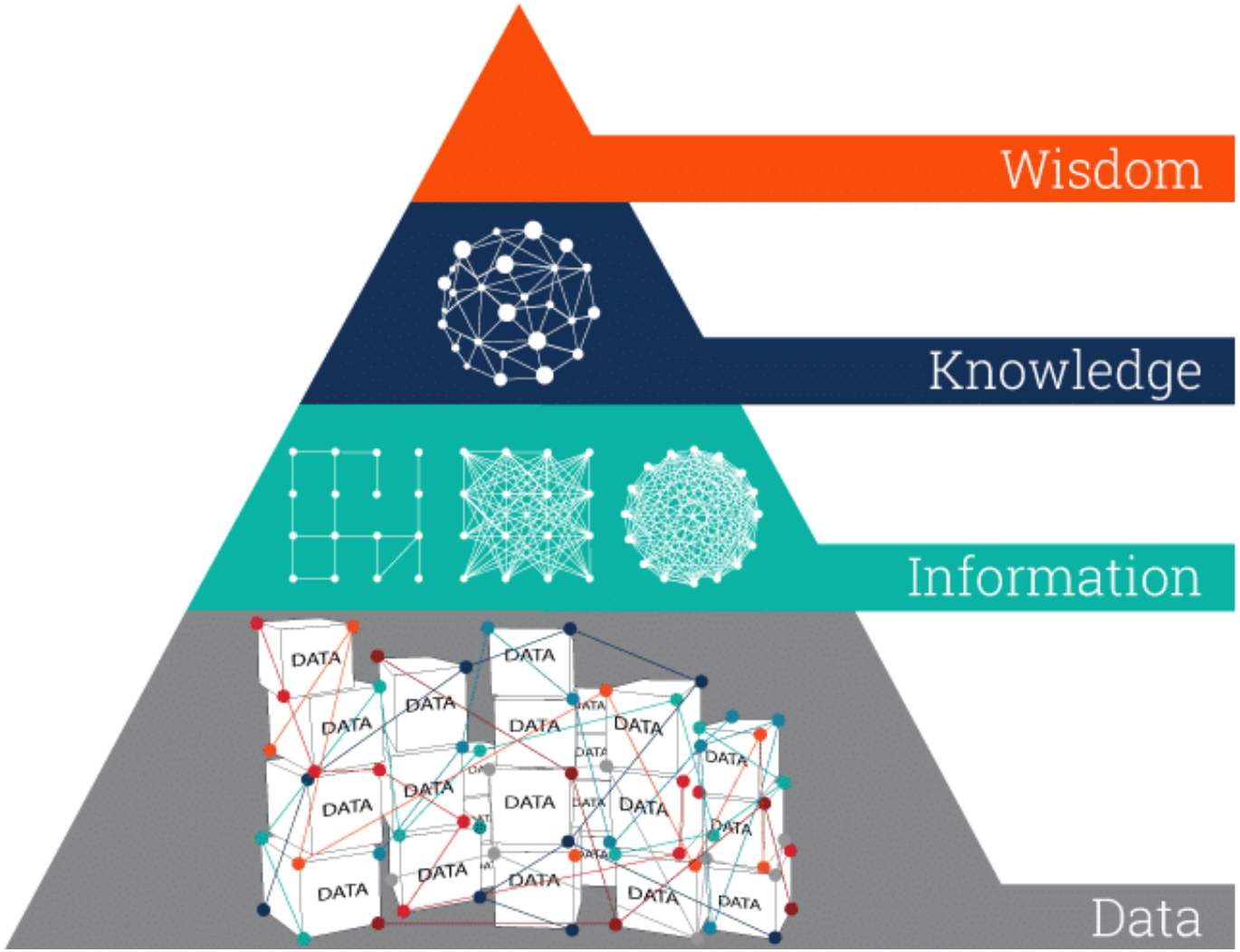
Agenda

- → The DIKW Pyramid!
- → Metadata, Metadata, METADATA!
- → (Some) Tools of the trade
  - → Apache Atlas
  - → LinkedIn Datahub
  - → Current Research
- → Data Mesh

SDU

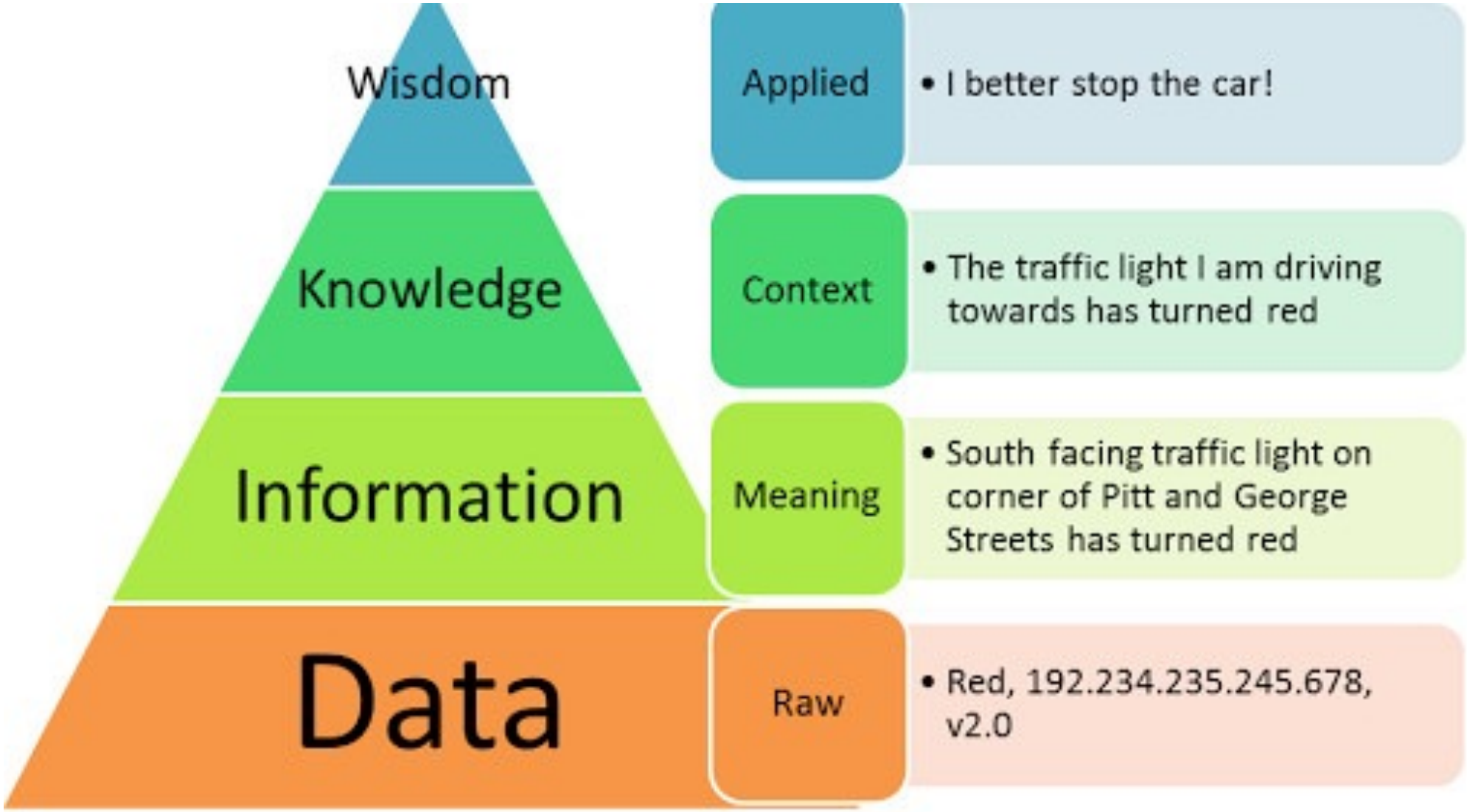
Data vs Information vs Knowledge vs Wisdom


Wisdom
Knowledge
Information
DATA
Data



Each step up
the pyramid
answers
questions
about and
adds value
to the initial data.

SDU


Wisdom
Knowledge
Information
Data
Applied
• I better stop the car!
Context
• The traffic light I am driving towards has turned red
Meaning
• South facing traffic light on corner of Pitt and George Streets has turned red
Raw
• Red, 192.234.235.245.678, v2.0

Metadata, Metadata, METADATA!

SDU

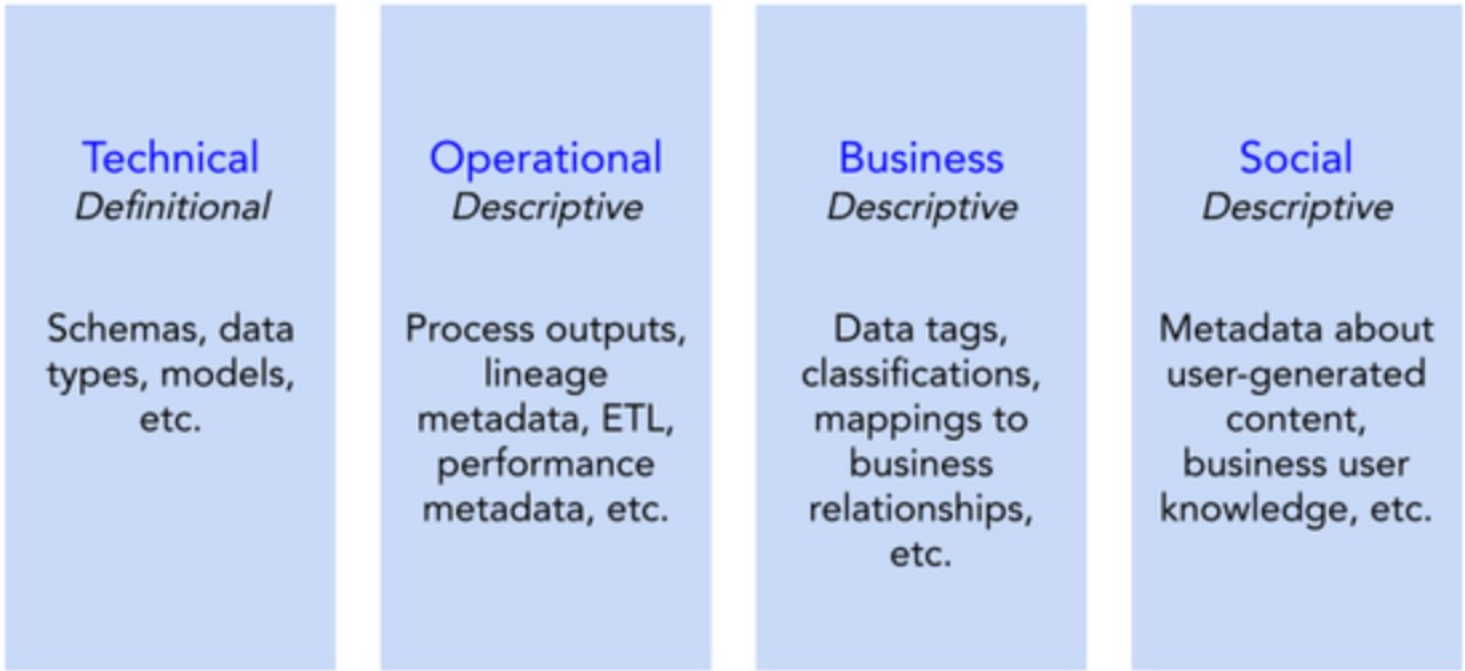
What is Metadata?

SDU

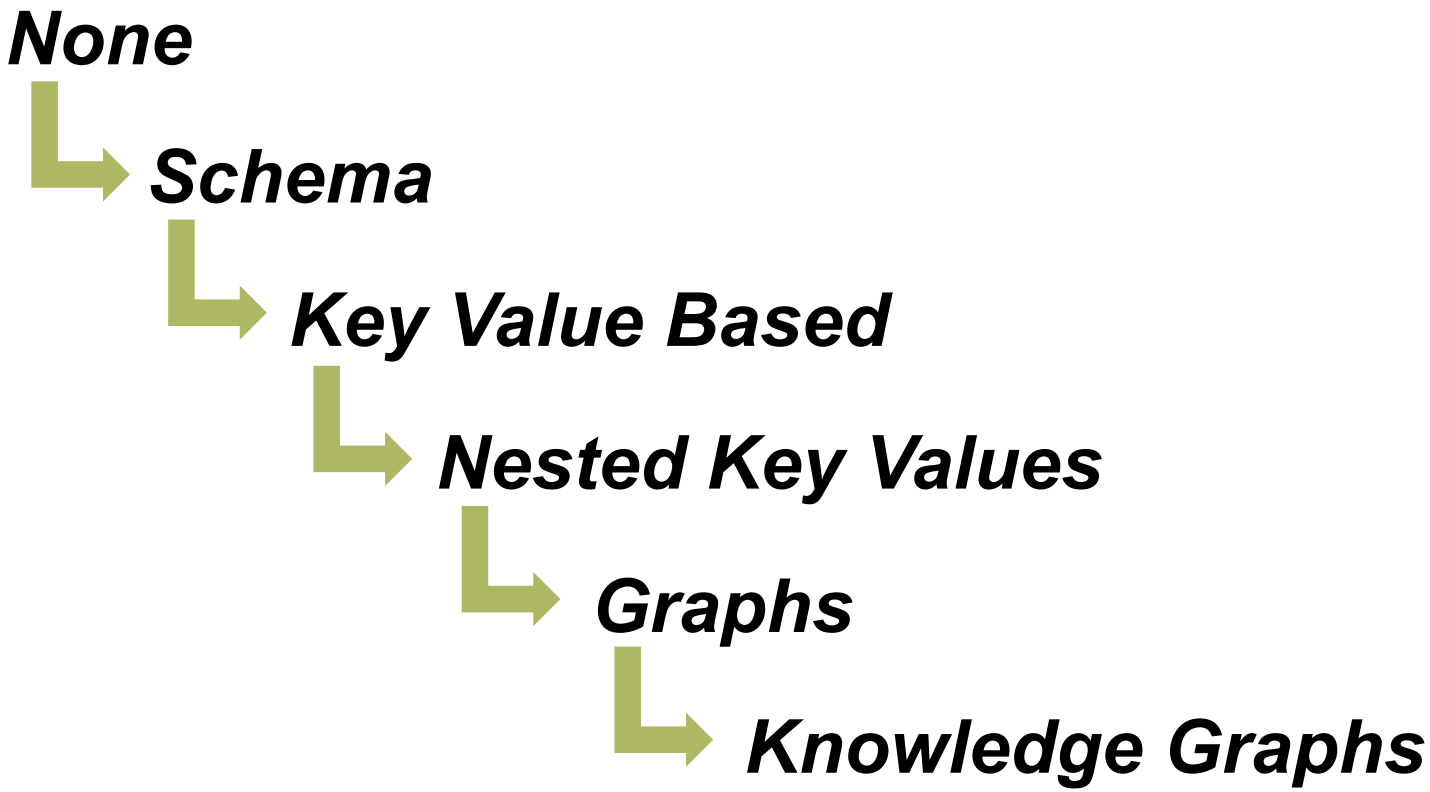
Why have Metadata?

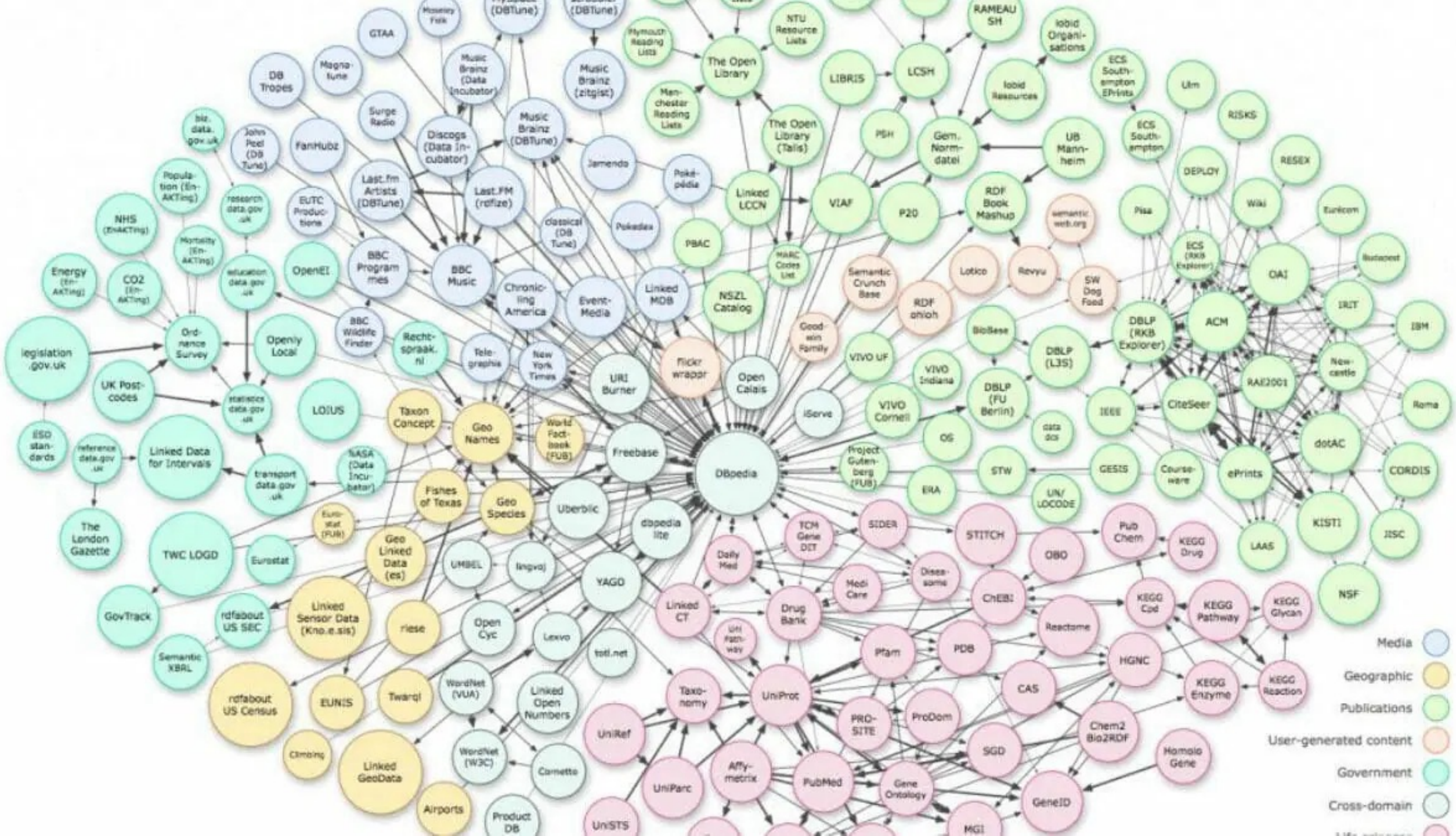
SDU

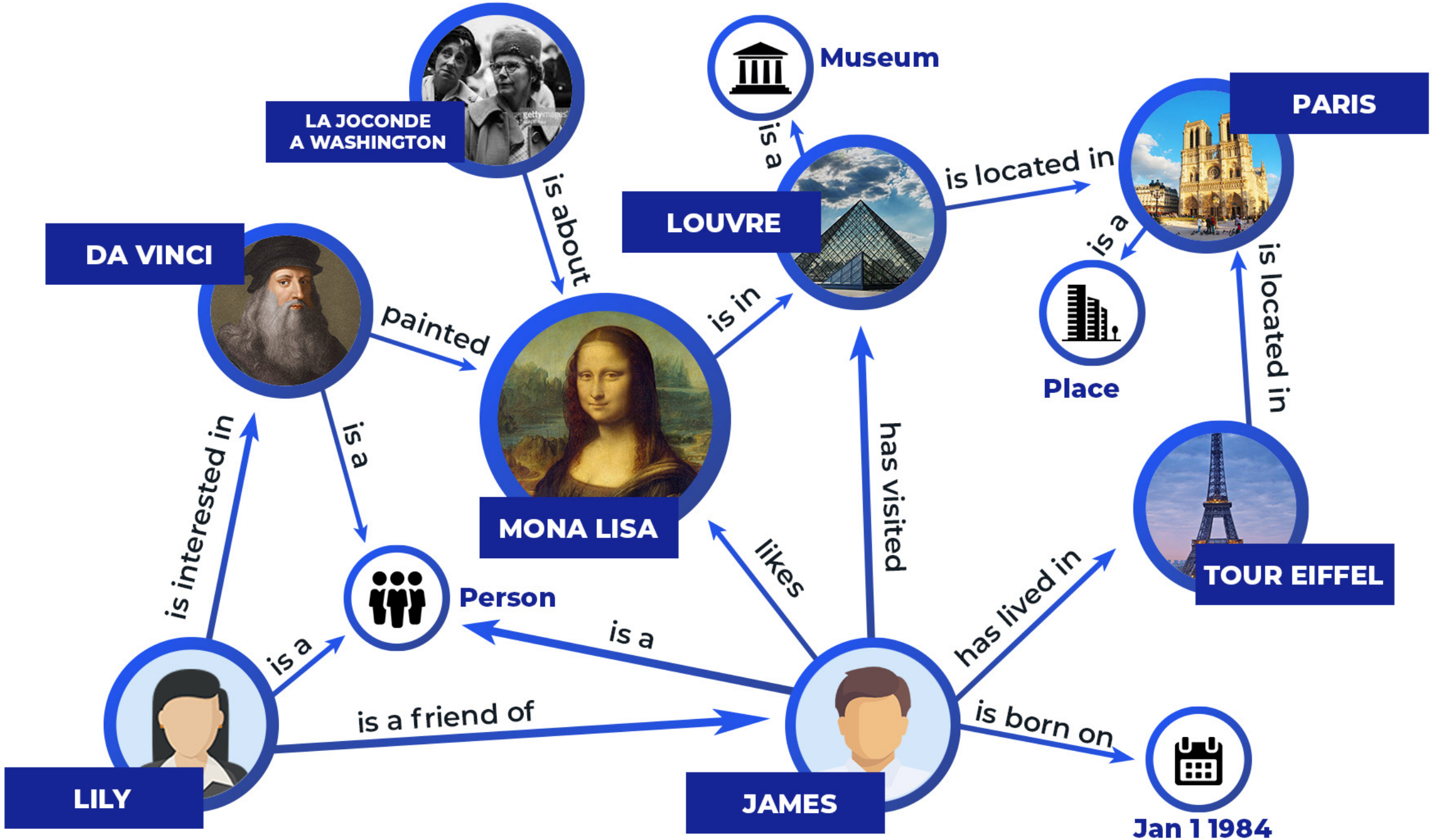
Types of metadata



Typical history of descriptive metadata




DBpedia
Media
Geographic
Publications
User-generated content
Government
Cross-domain


LA JOCONDE A WASHINGTON
is about
Museum
is a
LOUVRE
is located in
PARIS
is a
DA VINCI
painted
is in
Place
is located in
is a
is interested in
is a
MONA LISA
has visited
Person
likes
TOUR EIFFEL
is a
is a friend of
has lived in
LILY
JAMES
is born on
Jan 1 1984

What is Data Provenance?

provenance

/ˈprɒv(ə)nəns/

noun

the place of origin or earliest known history of something.
"an orange rug of Iranian provenance"

Similar: origin | source | place of origin | birthplace | spring | wellspring | fount

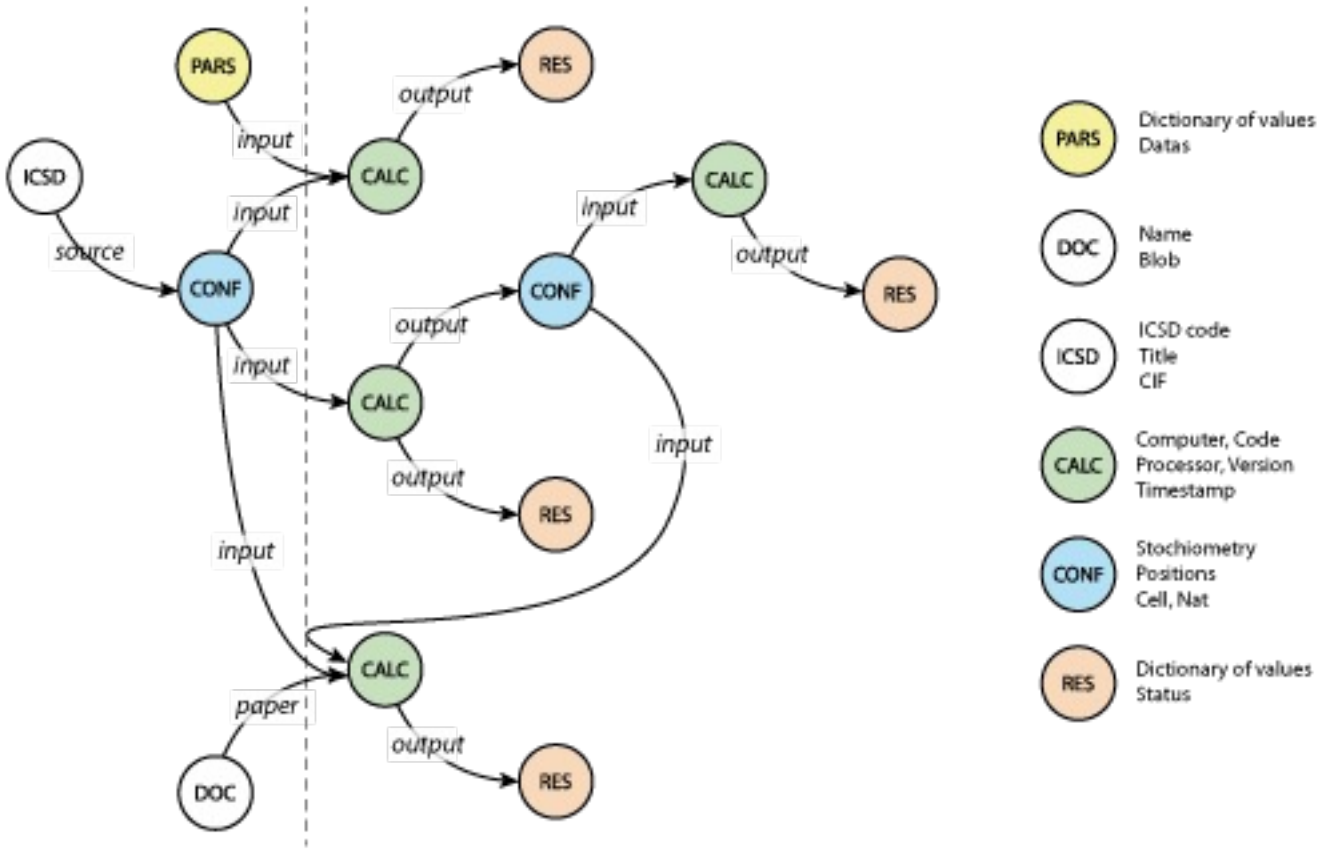
- the beginning of something's existence; something's origin.
  "they try to understand the whole universe, its provenance and fate"
- a record of ownership of a work of art or an antique, used as a guide to authenticity or quality.
  plural noun: **provenances**
  "the manuscript has a distinguished provenance"

Data Provenance in Data Science

- Lineage
- Traceability
- Compliance
- Dependency Analysis
- Data and AI Reproducibility
- AI Explainability

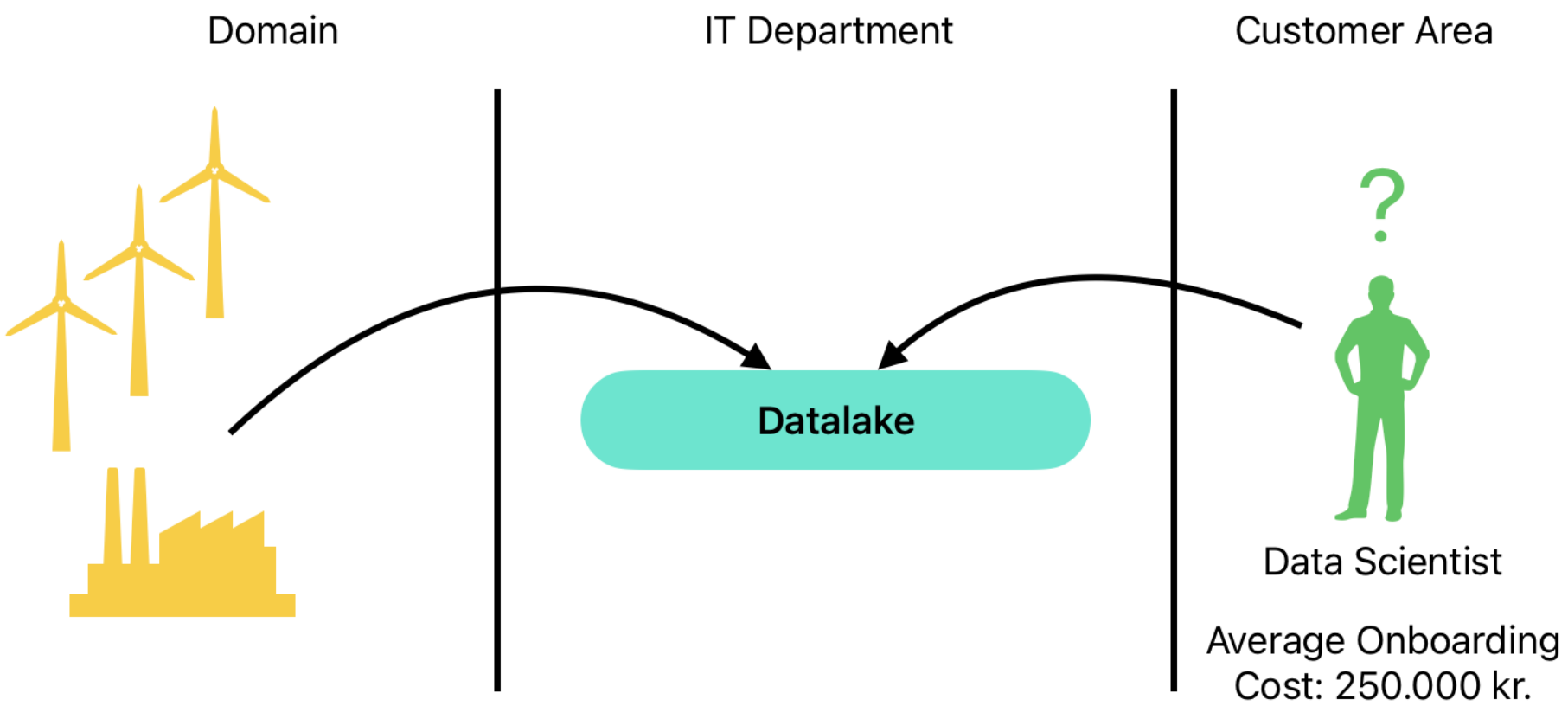
SDU

Lineage, Tracability, Reproducability

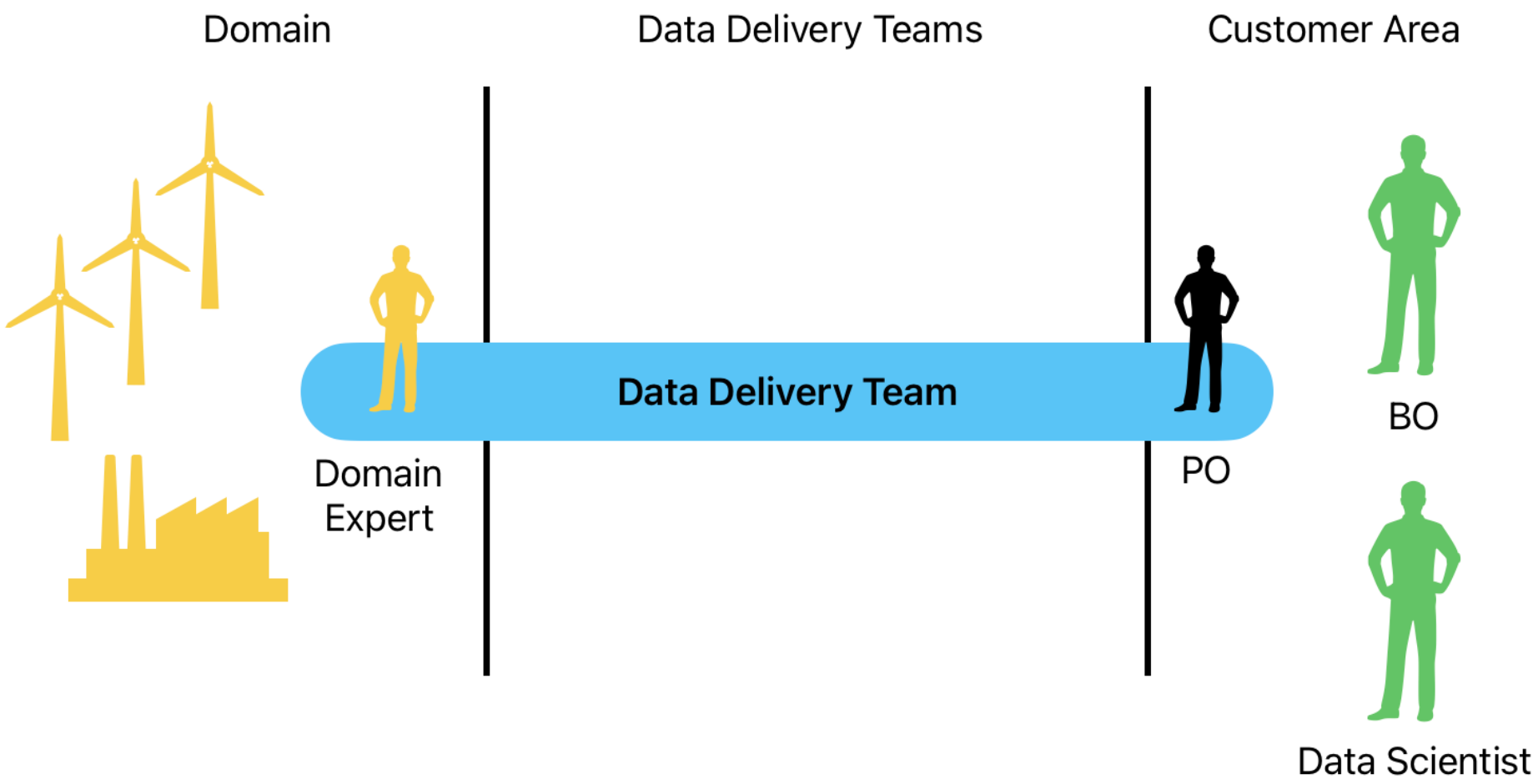


SDU

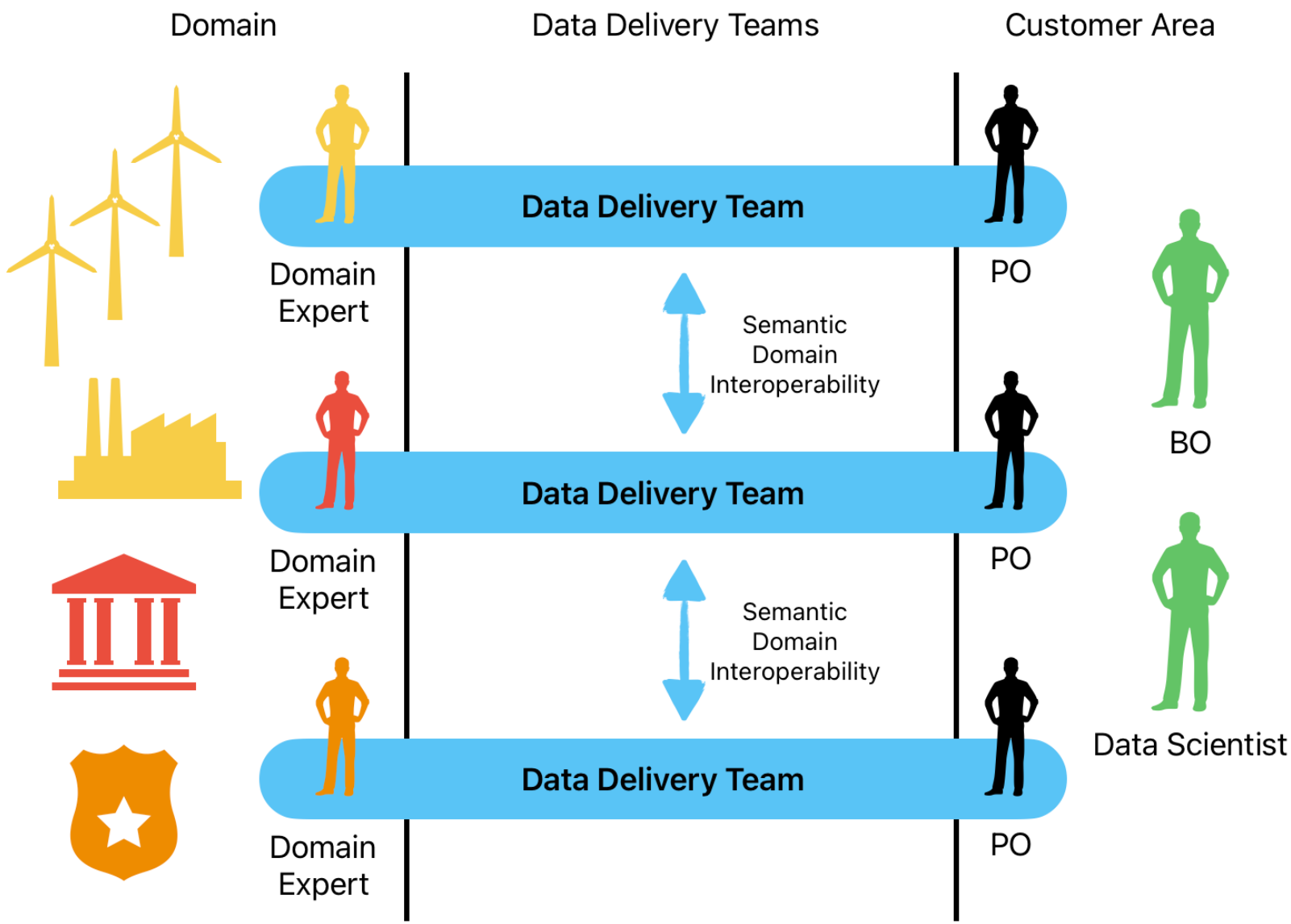
Current State – Domain Knowledge is Impossible to access



Data Delivery Teams Create Access to Domain Knowledge



And Allows Us to Bridge Domains



SDU

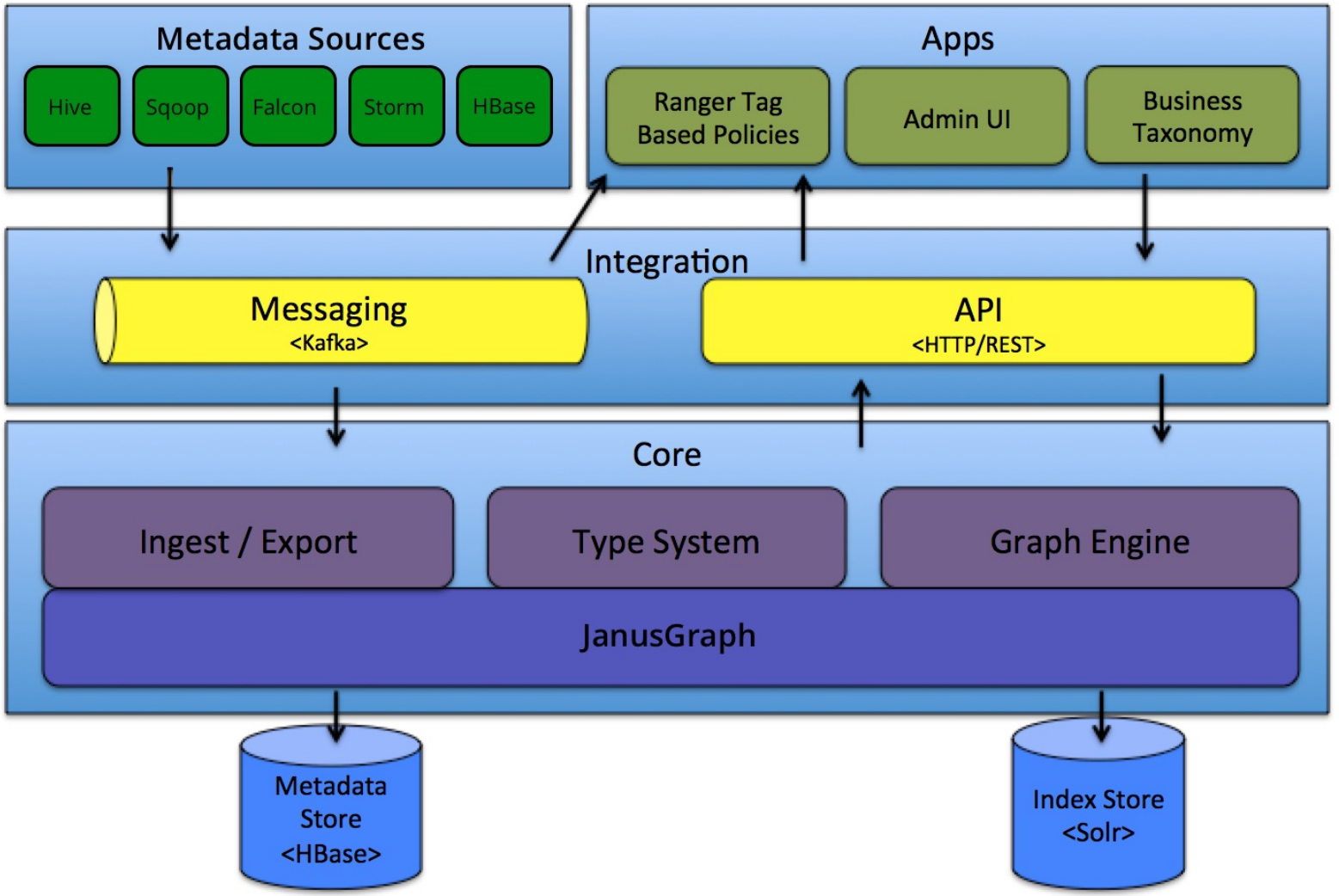
Metadata Implementations: Apache Atlas

SDU

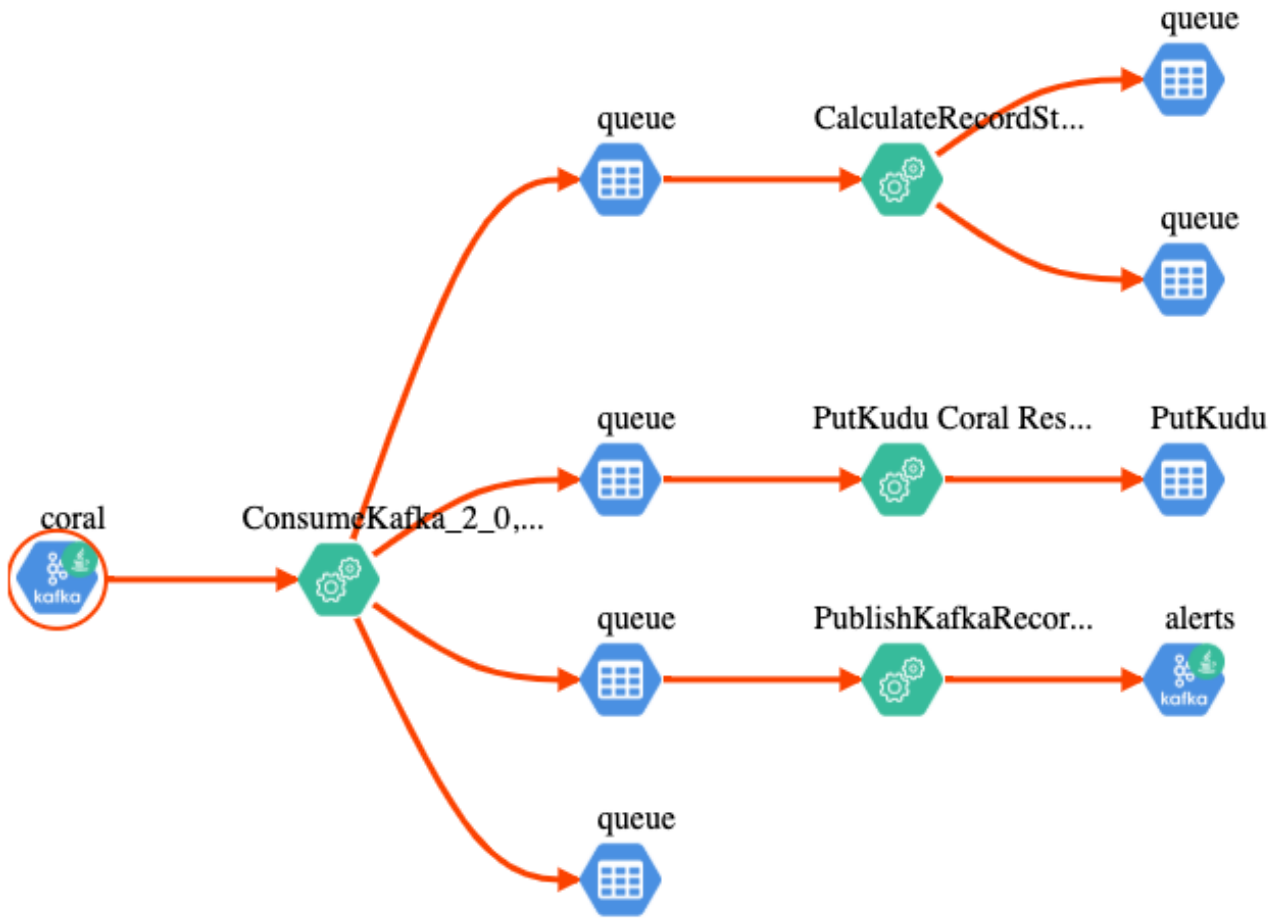
Apache Atlas

- → Features
  - → Types (Schema)
  - → Classification
  - → Lineage
  - → Search/Discovery
  - → Governance (Security and Data Masking)
- → Older metadata alternative
- → Creates sematic connections between terms
- → Relatively simple compared to newer alternatives
- → Integrates well with the Apache ecosystem

Atlas Architecture

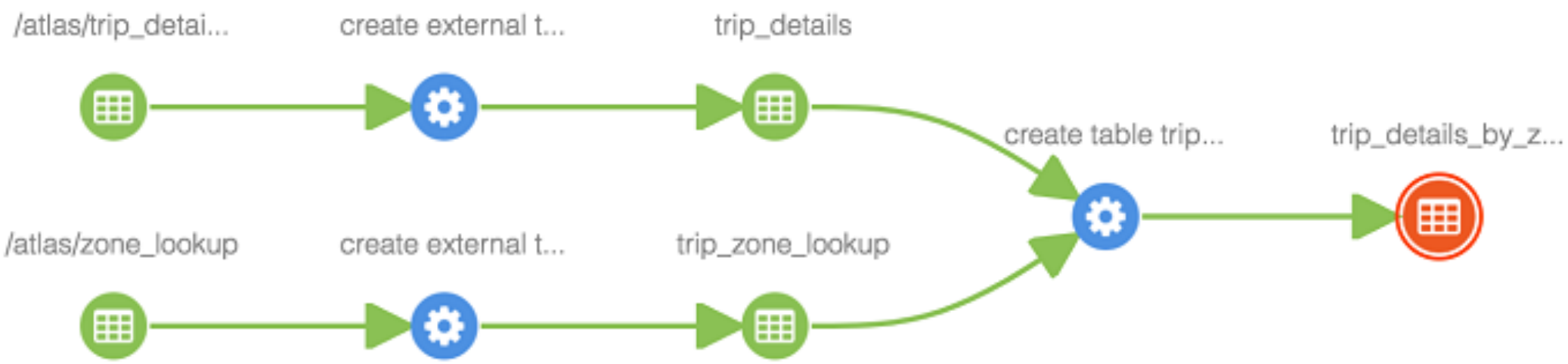


Atlas Lineage – Realtime Data (Kafka)



SDU

Atlas Lineage – Historical Data (Hive)

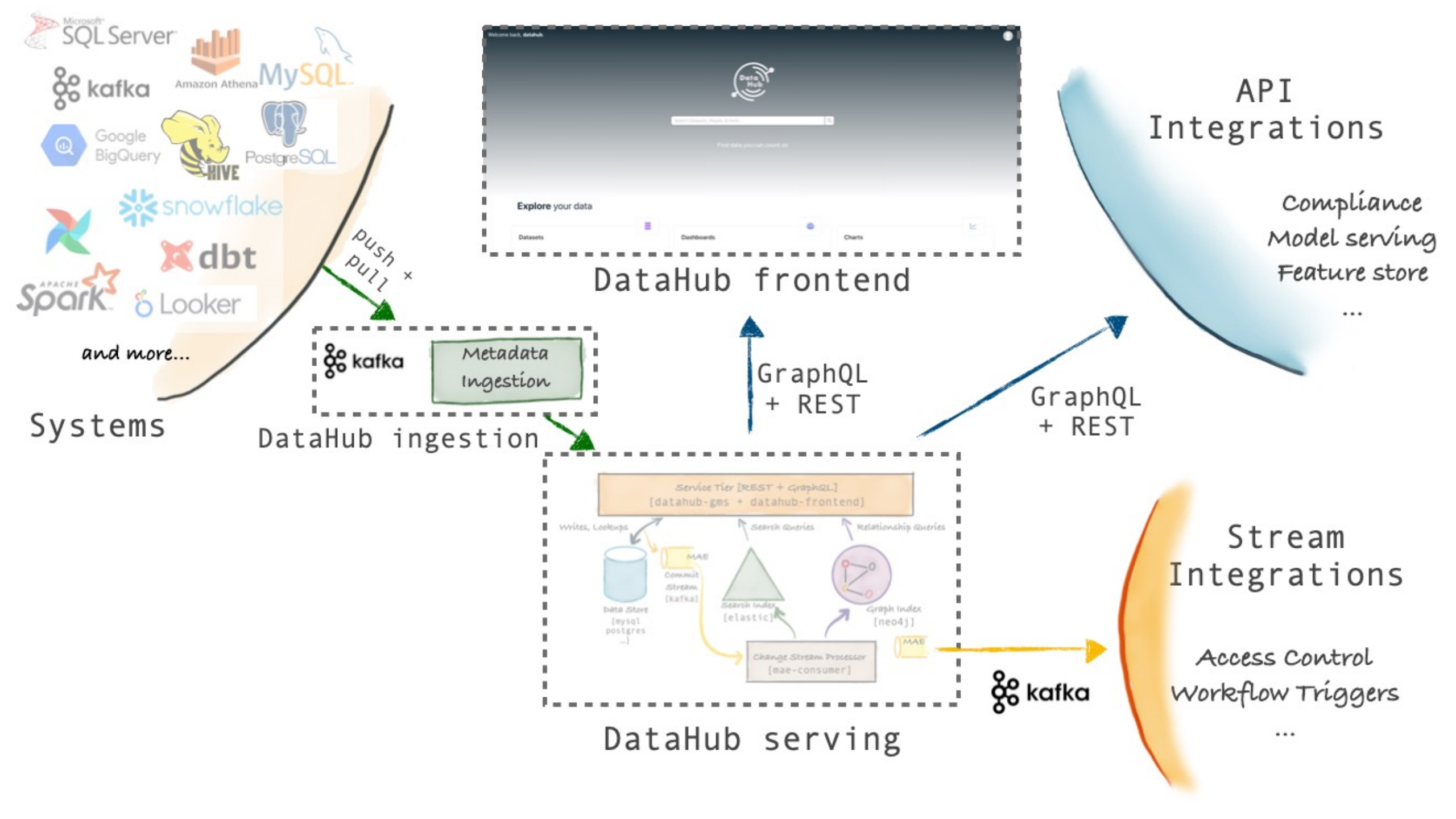


Metadata Implementations: LinkedIn Datahub

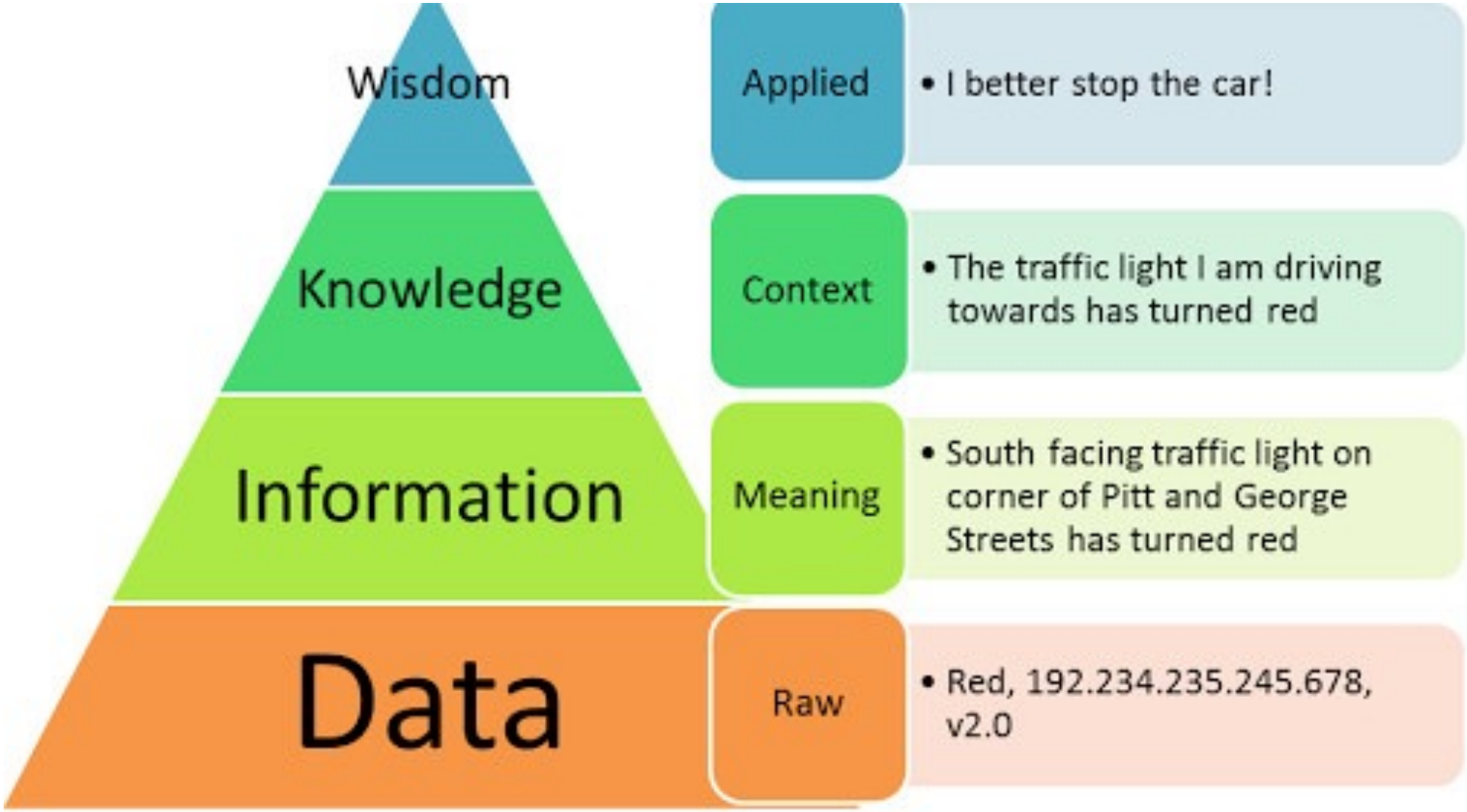
SDU

LinkedIn Datahub

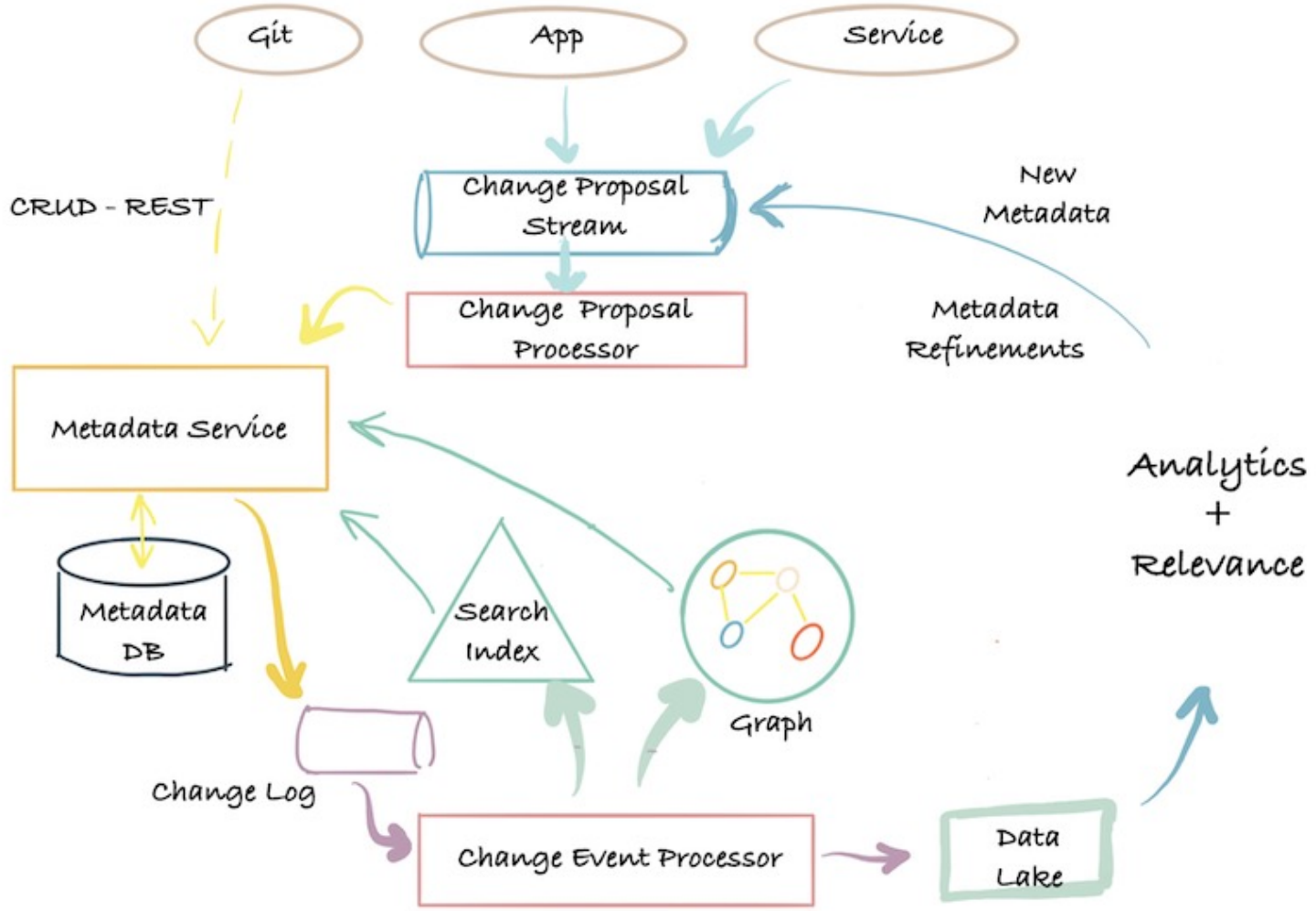
- → Features
  - → Search and Discovery
  - → Access Control
  - → Data Lineage
  - → Compliance
  - → Data management and automated metadata generation
  - → AI Explainability and reproducibility
- → Extremely extensive compatibility with databases and data lakes
- → Not bound to a specific ecosystem

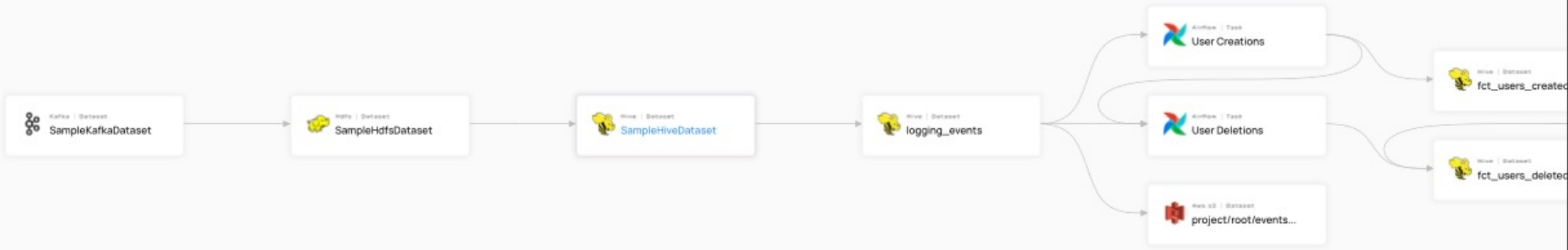

Microsoft SQL Server
kafka
Amazon Athena
MySQL
Google BigQuery
HIVE
PostgreSQL
snowflake
dbt
Looker
APACHE Spark
and more...
Systems
push + pull
Welcome back, datahub.
Explore your data
Datasets
Dashboards
Charts
DataHub frontend
API Integrations
Compliance
Model serving
Feature store
...
kafka
Metadata Ingestion
DataHub ingestion
GraphQL + REST
GraphQL + REST
Service Tier [REST + GraphQL]
[datahub-gms + datahub-frontend]
Writes, Lookups
Search Queries
Relationship Queries
MAE
Commit Stream [kafka]
Data Store [mysql postgres ...]
Search Index [elastic]
Graph Index [neo4j]
MAE
Change Stream Processor [mae-consumer]
DataHub serving
kafka
Stream Integrations
Access Control
Workflow Triggers
...

What Meta Data limitations are there?

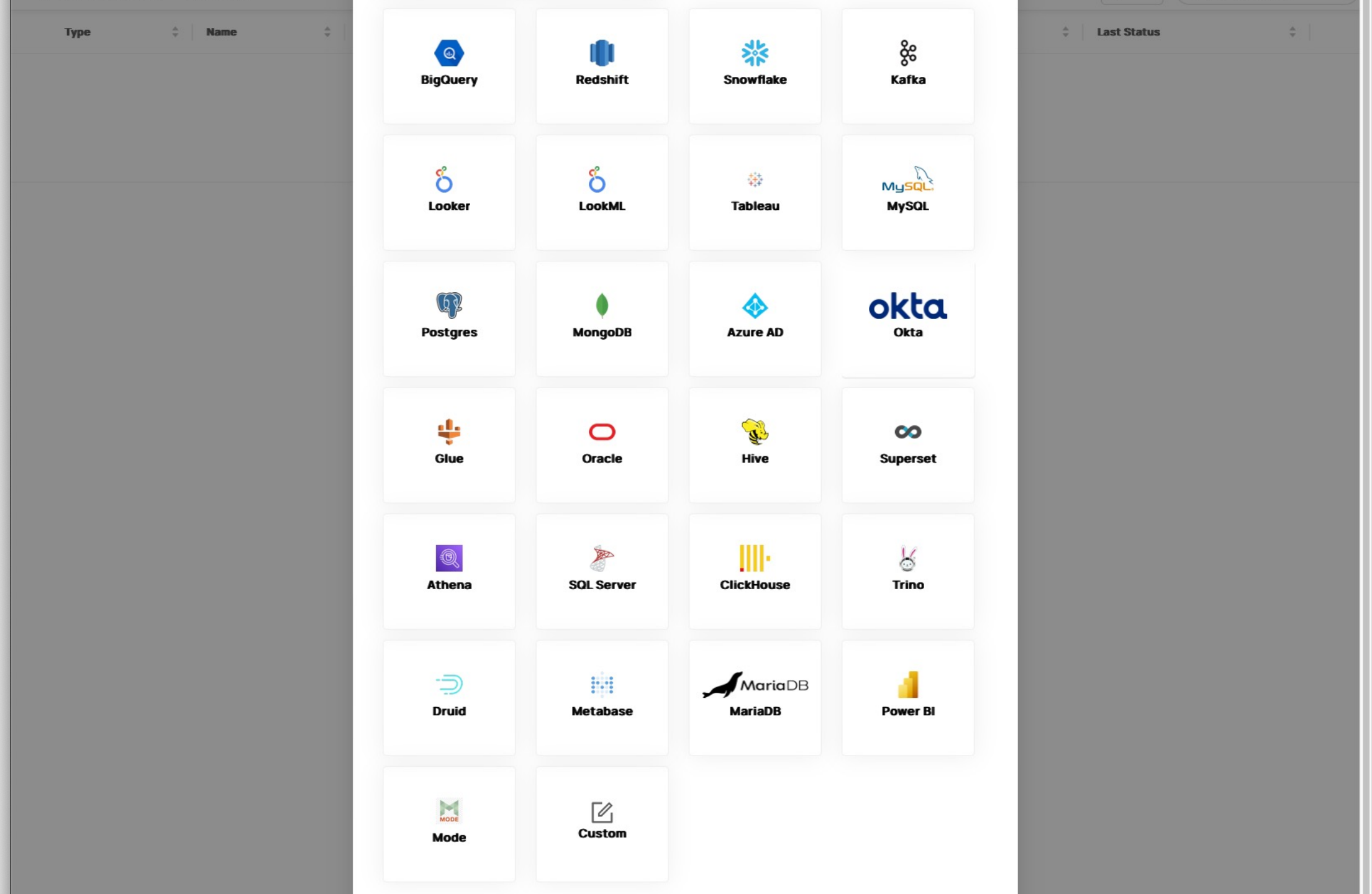

Wisdom
Knowledge
Information
Data
Applied
• I better stop the car!
Context
• The traffic light I am driving towards has turned red
Meaning
• South facing traffic light on corner of Pitt and George Streets has turned red
Raw
• Red, 192.234.235.245.678, v2.0

Dynamic Metadata Changes

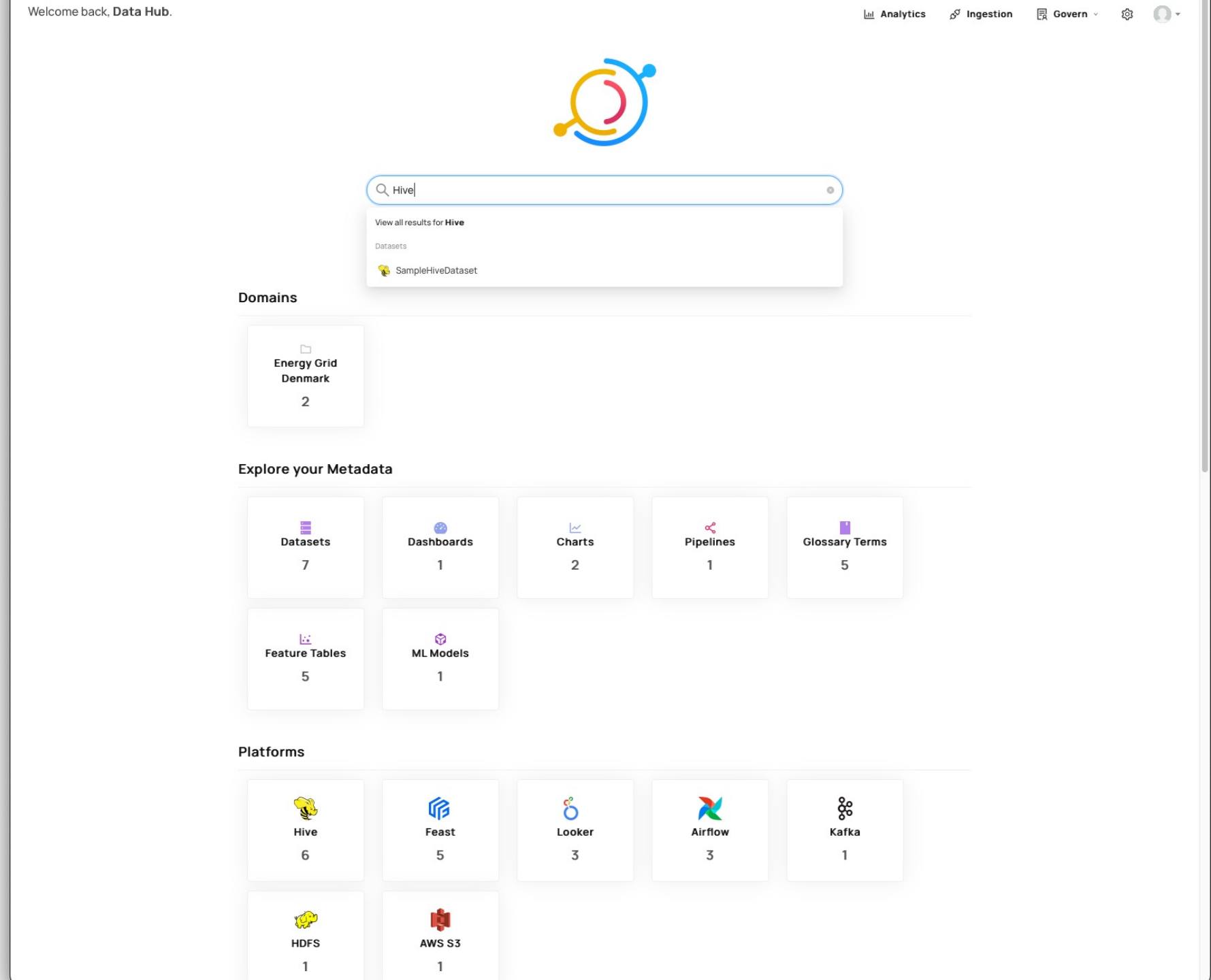



SampleKafkaDataset
SampleHdfsDataset
SampleHiveDataset
logging_events
User Creations
User Deletions
project/root/events...
fct_users_created
fct_users_deleted

SDU


Type
Name
Last Status
BigQuery
Redshift
Snowflake
Kafka
Looker
LookML
Tableau
MySQL
Postgres
MongoDB
Azure AD
Okta
Glue
Oracle
Hive
Superset
Athena
SQL Server
ClickHouse
Trino
Druid
Metabase
MariaDB
Power BI
Mode
Custom

SDU


Welcome back, Data Hub.
Analytics
Ingestion
Govern
Hive
View all results for Hive
Datasets
SampleHiveDataset
Domains
Energy Grid Denmark
2
Explore your Metadata
Datasets
7
Dashboards
1
Charts
2
Pipelines
1
Glossary Terms
5
Feature Tables
5
ML Models
1
Platforms
Hive
6
Feast
5
Looker
3
Airflow
3
Kafka
1
HDFS
1
AWS S3
1

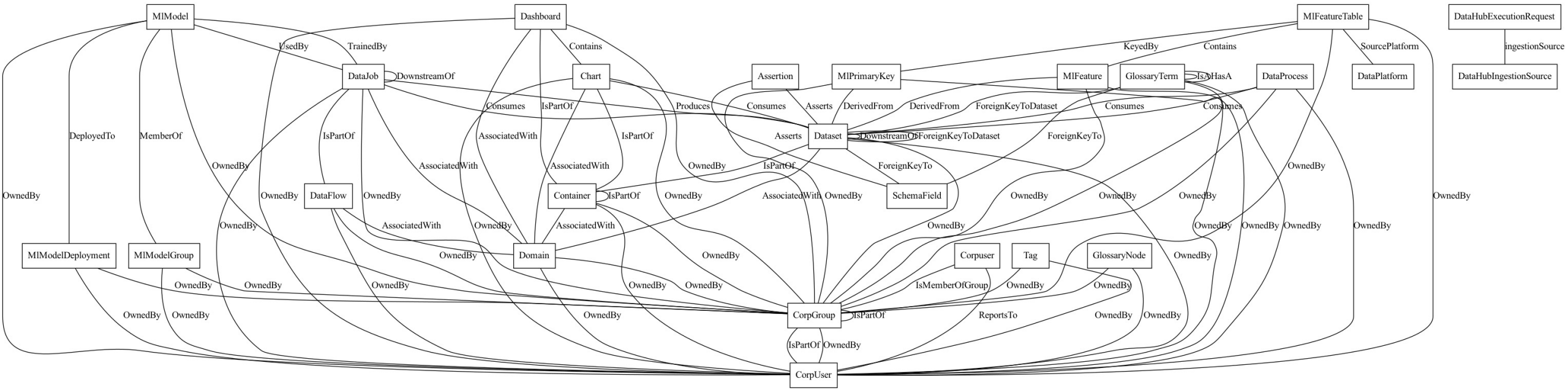
SDU

The Core Entities

DataHub's "core" Entity types model the Data Assets that comprise the Modern Data Stack. They include

1. Data Platform: A type of Data "Platform". That is, an external system that is involved in processing, storing, or visualizing Data Assets. Examples include MySQL, Snowflake, Redshift, and S3.
2. Dataset: A collection of data. Tables, Views, Streams, Document Collections, and Files are all modeled as "Datasets" on DataHub. Datasets can have tags, owners, links, glossary terms, and descriptions attached to them. They can also have specific sub-types, such as "View", "Collection", "Stream", "Explore", and more. Examples include Postgres Tables, MongoDB Collections, or S3 files.
3. Chart: A single data vizualization derived from a Dataset. A single Chart can be a part of multiple Dashboards. Charts can have tags, owners, links, glossary terms, and descriptions attached to them. Examples include a Superset or Looker Chart.
4. Dashboard: A collection of Charts for visualization. Dashboards can have tags, owners, links, glossary terms, and descriptions attached to them. Examples include a Superset or Mode Dashboard.
5. Data Job (Task): An executable job that processes data assets, where "processing" implies consuming data, producing data, or both. Data Jobs can have tags, owners, links, glossary terms, and descriptions attached to them. They must belong to a single Data Flow. Examples include an Airflow Task.
6. Data Flow (Pipeline): An executable collection of Data Jobs with dependencies among them, or a DAG. Data Jobs can have tags, owners, links, glossary terms, and descriptions attached to them. Examples include an Airflow DAG.

See the **Metadata Modeling/Entities** section on the left to explore the entire model.


MlModel
Dashboard
MlFeatureTable
DataHubExecutionRequest
UsedBy
TrainedBy
Contains
KeyedBy
Contains
SourcePlatform
ingestionSource
DataJob
DownstreamOf
Chart
Assertion
MlPrimaryKey
MlFeature
GlossaryTerm
IsA/HasA
DataProcess
DataPlatform
DataHubIngestionSource
Consumes
IsPartOf
Produces
Consumes
Asserts
DerivedFrom
DerivedFrom
ForeignKeyToDataset
Consumes
Consumes
DeployedTo
MemberOf
IsPartOf
AssociatedWith
IsPartOf
Asserts
Dataset
DownstreamOf
ForeignKeyToDataset
ForeignKeyTo
OwnedBy
AssociatedWith
AssociatedWith
OwnedBy
IsPartOf
ForeignKeyTo
OwnedBy
OwnedBy
OwnedBy
DataFlow
OwnedBy
Container
IsPartOf
OwnedBy
AssociatedWith
OwnedBy
SchemaField
OwnedBy
OwnedBy
OwnedBy
OwnedBy
OwnedBy
AssociatedWith
OwnedBy
AssociatedWith
OwnedBy
OwnedBy
OwnedBy
OwnedBy
OwnedBy
MlModelDeployment
MlModelGroup
Domain
Corpuser
Tag
GlossaryNode
OwnedBy
OwnedBy
OwnedBy
OwnedBy
OwnedBy
OwnedBy
OwnedBy
IsMemberOfGroup
OwnedBy
OwnedBy
OwnedBy
OwnedBy
CorpGroup
IsPartOf
ReportsTo
OwnedBy
OwnedBy
OwnedBy
IsPartOf
OwnedBy
CorpUser

SDU

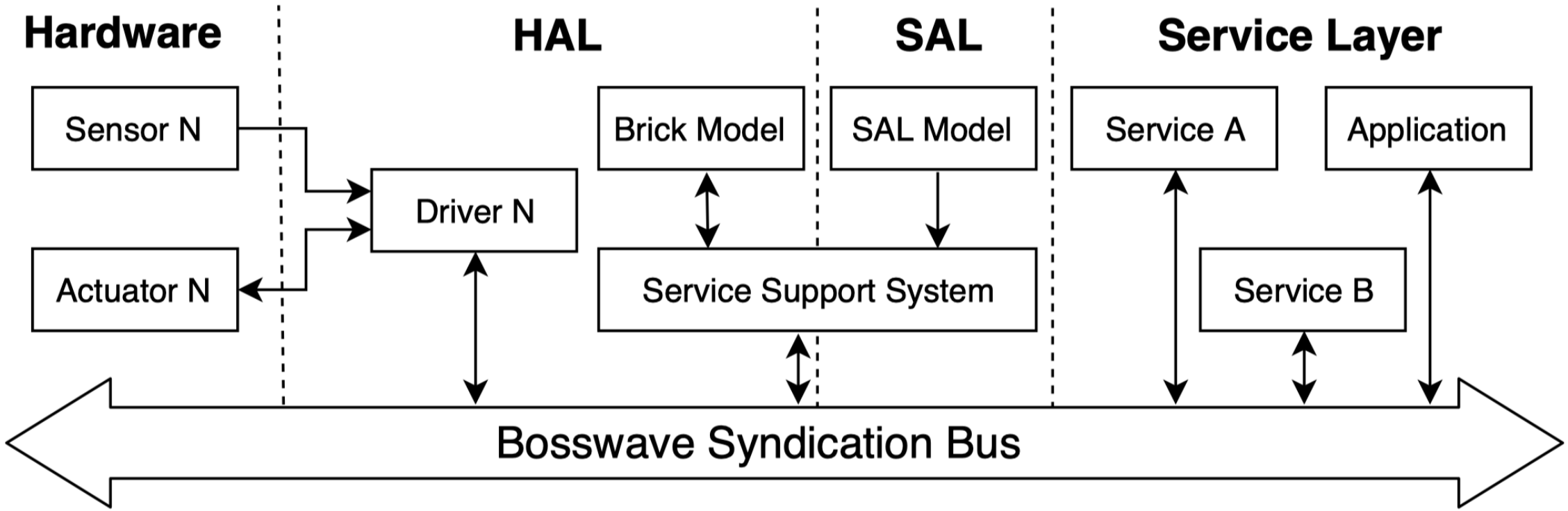
Atlas, Datahub, and others, only uses graphs for Lineage, Dependencies, Ownership and Data Flow, not to argument the metadata context itself.

For a short explanation of DataHub, watch this:
https://www.youtube.com/watch?v=VY57iRdG-Us

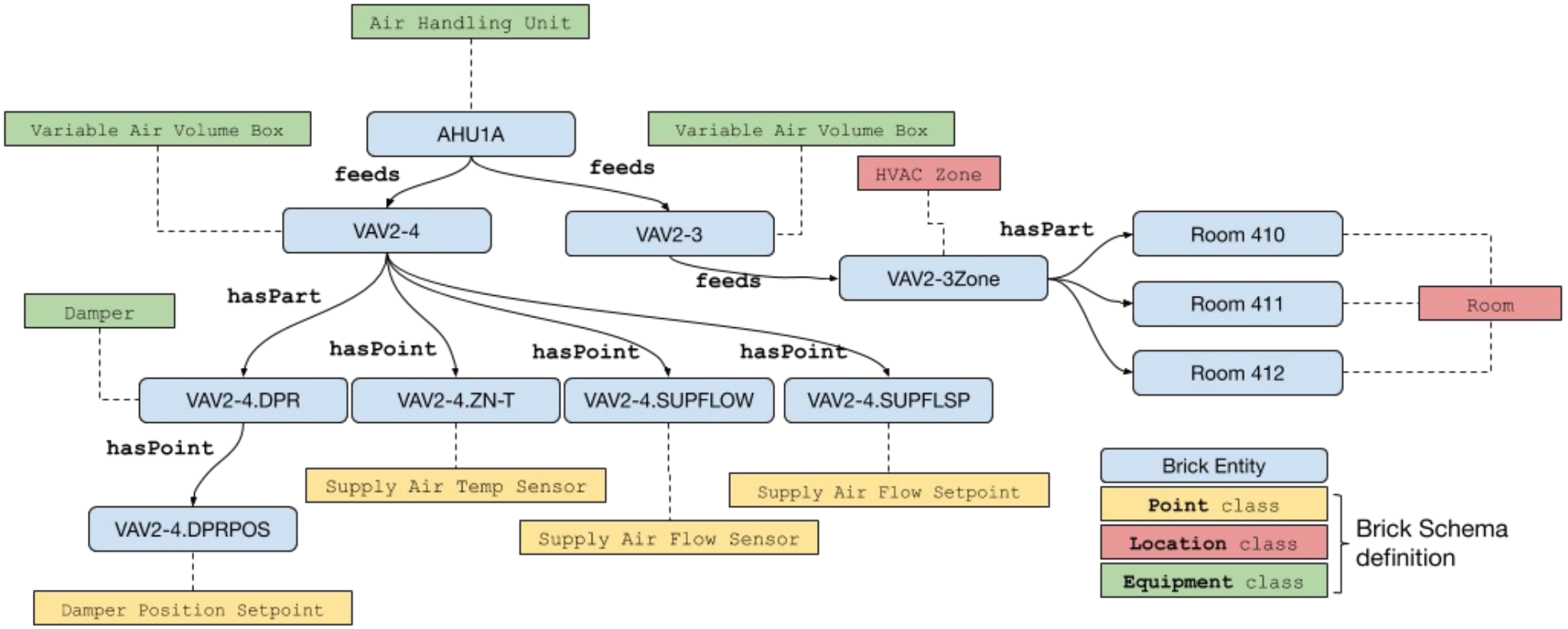
Metadata Research:
Brick and SAL (Knowledge Graphs)

SDU

Why are knowledge graphs important?

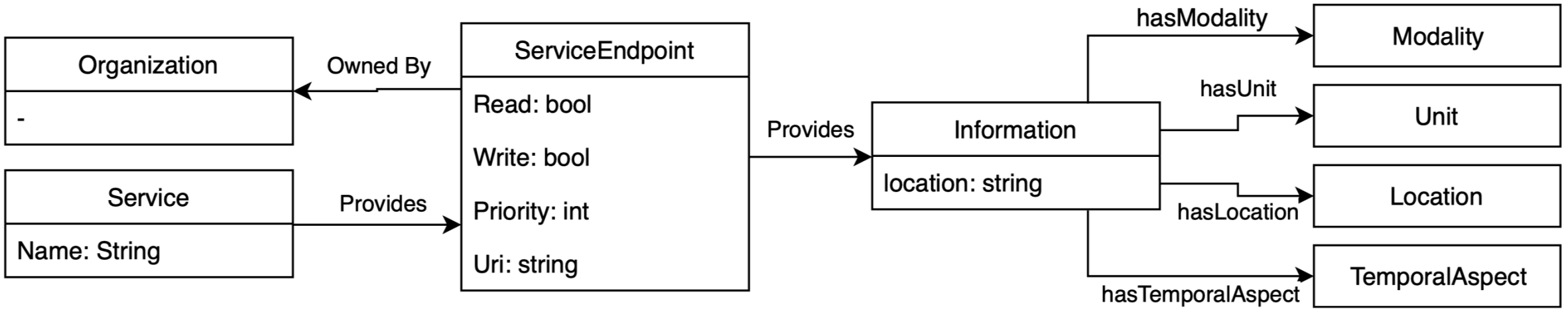


Brick instantiation example



SDU

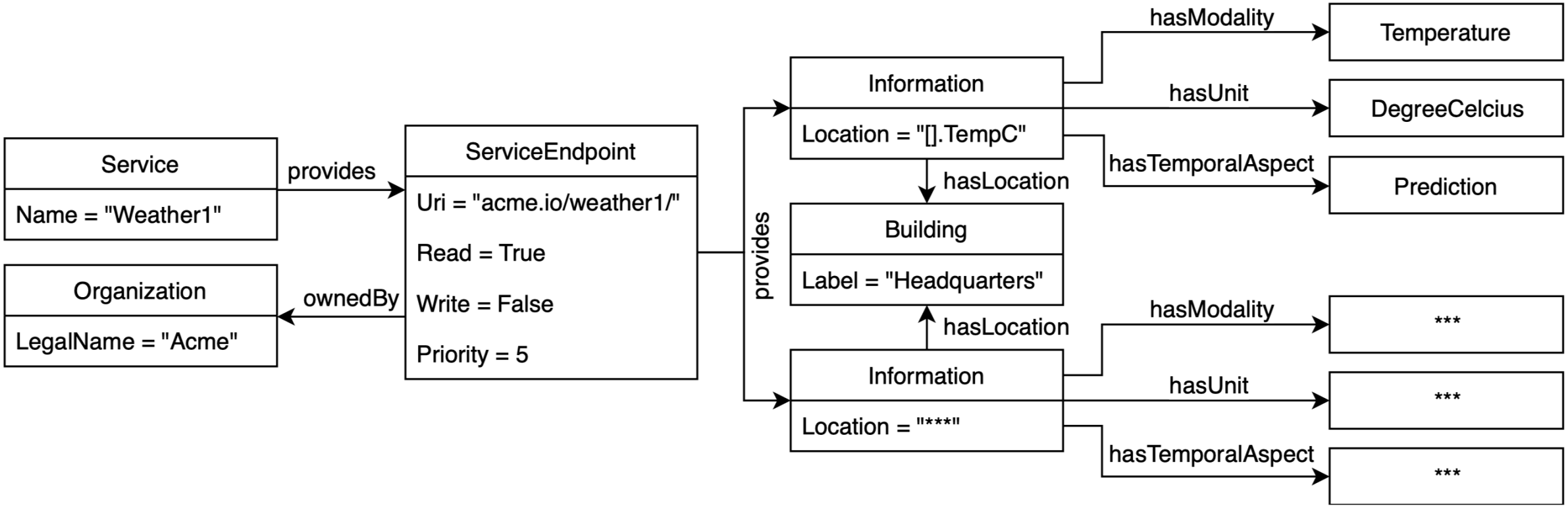
SAL Ontology Classes



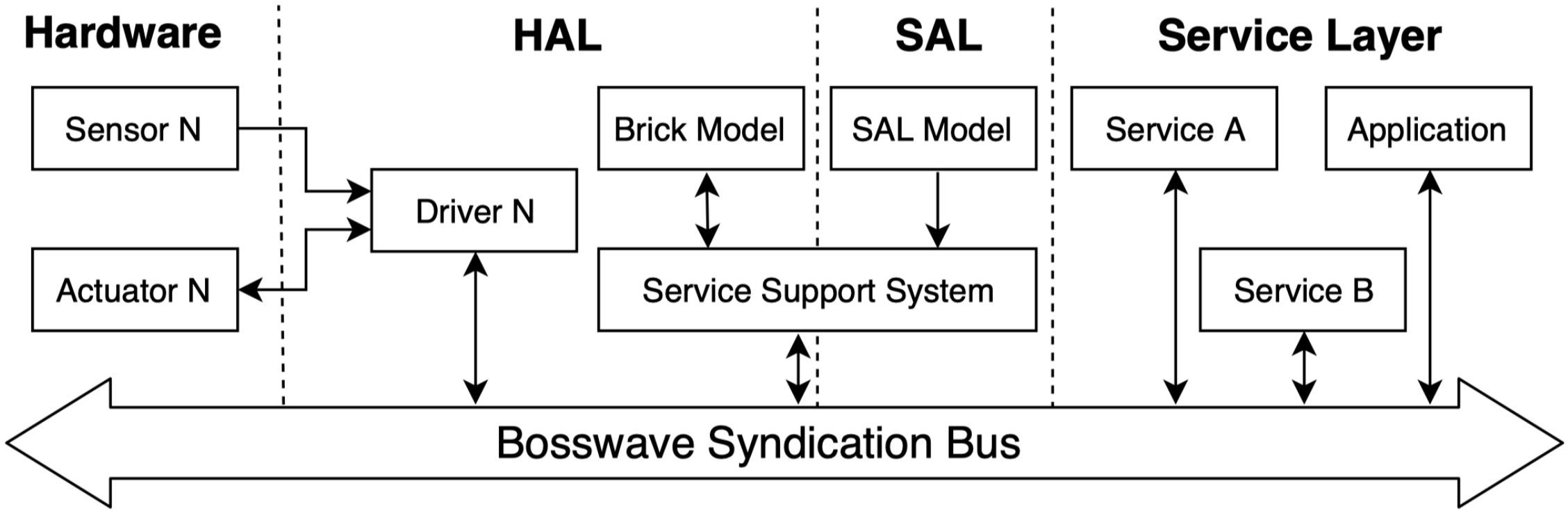
SAL Types

Modality	Unit	TemporalAspect
Angle, CO_2, Presence, Flow, Illuminance, Power, Pressure, Rain, Humidity, Temperature, Wind, AbsoluteTime, RelativeTime, PowerFlexibility, Performance, Energy, Certainty, Time	Boolean, Count, CubicMeters, CubicMetersPerHour, DegreeCelsius, DegreeFahrenheit, Degrees, GigaByte, KiloByte, Hours, Hertz, Joules, JoulesPerCubicMeter, Kelvin, KiloJoulesPerSquareMeter, KiloMeters, KiloWatts, KiloWattHours, Lux, CubicMeters, CubicMetersPerHour, CubicMetersPerSecond, MilliAmperes, Minutes, MilliMeters, MilliSeconds, MetersPerSecond, MilliVolts, MilliWatts, MilliWattHours, Pascal, Percent, PartsPerMillion, RotationsPerMinute, Volts, Watts, Unitless, Time, DateTime, Date	Prediction RealTime Archival

SAL Instantiation Example



How are the knowledge graphs applied?



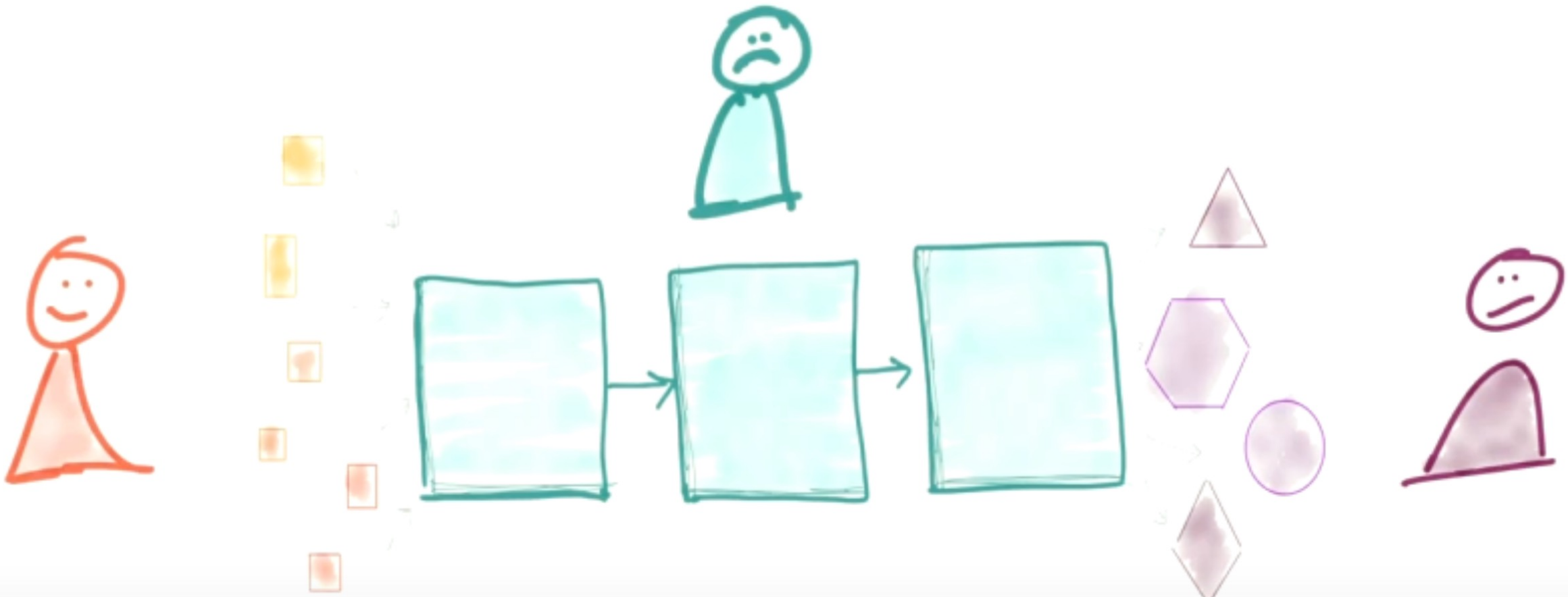
Looking to the future: Data Mesh

This section will give you a quick look into the future, and ensure you have a concept of what Data Mesh means if you are confronted with the term in the industry.

Slides on loan from Zhamak Dehghani

For a comprehensive walkthrough, watch:
https://www.youtube.com/watch?v=52MCFe4v0UU

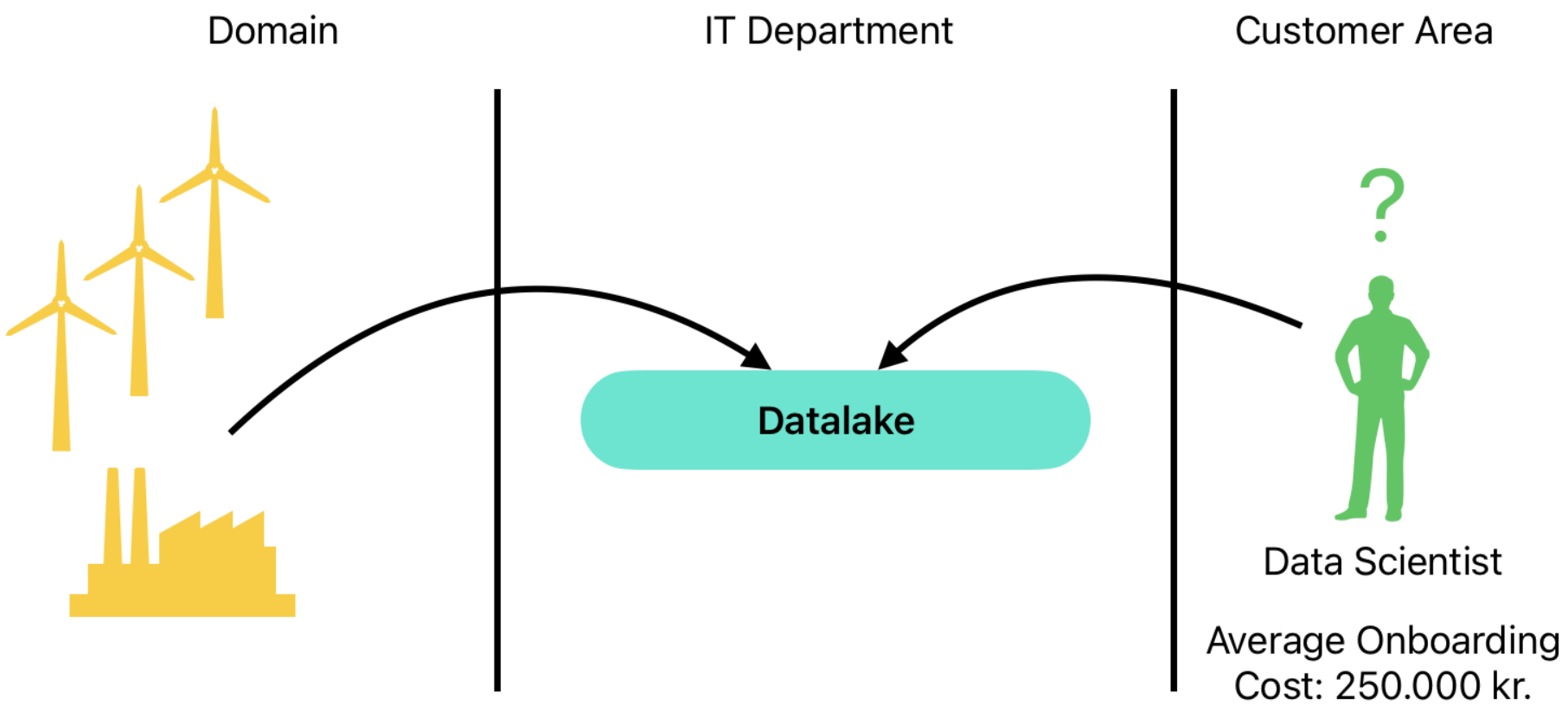
Data Platform Engineers



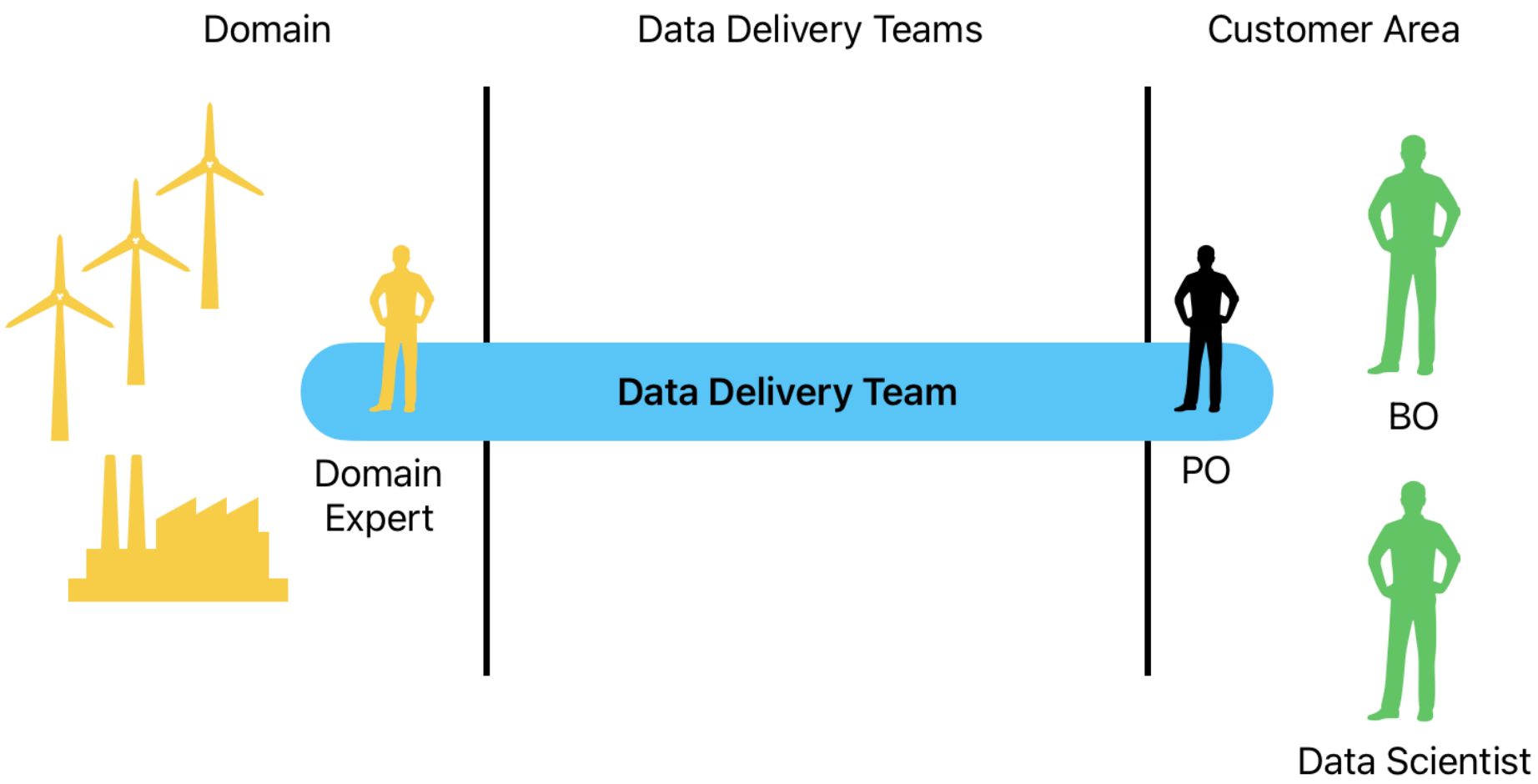
Domains' Operational systems Teams

Data Scientists, BI teams, ...

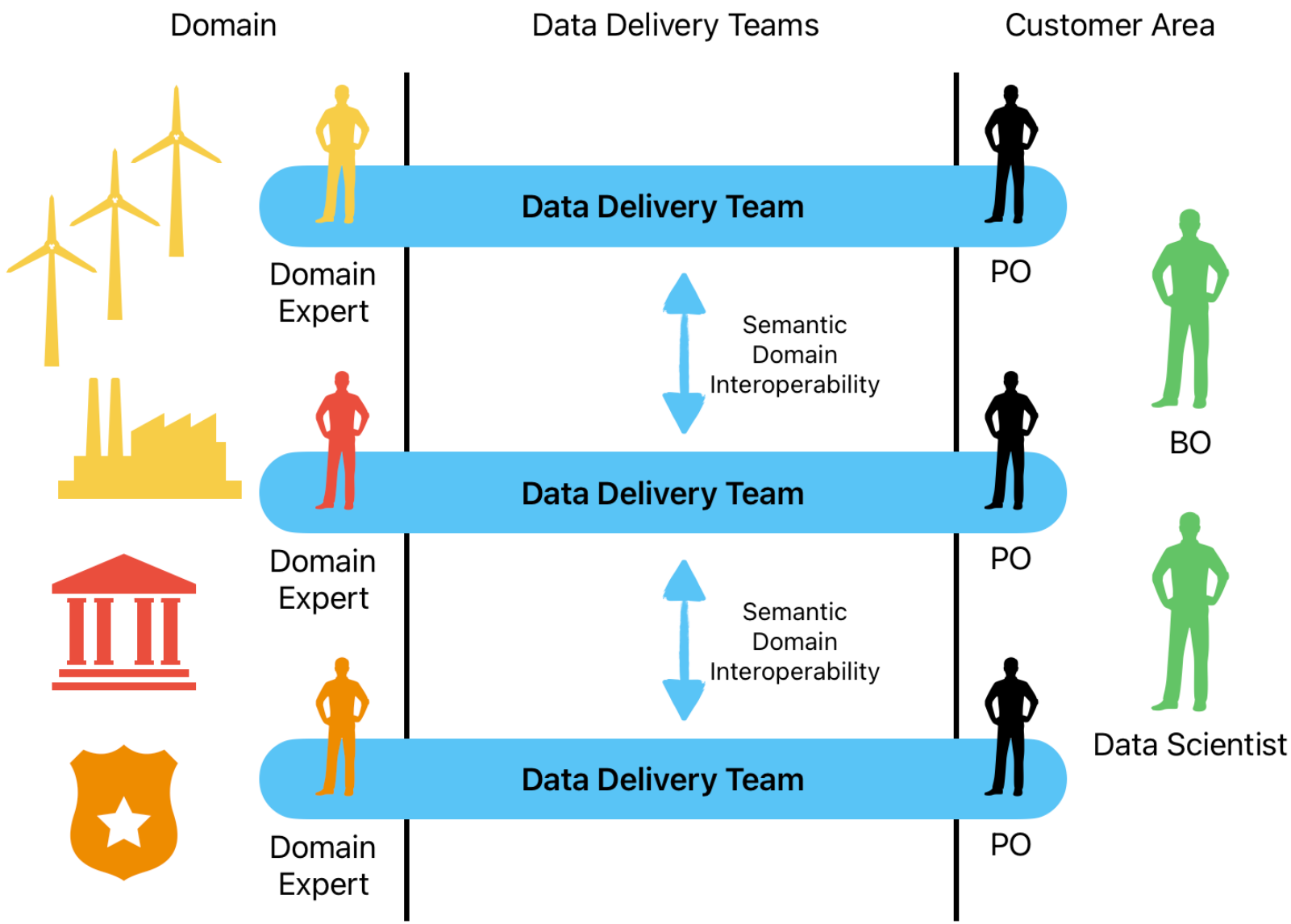
Current State – Domain Knowledge is Impossible to access



Data Delivery Teams Create Access to Domain Knowledge

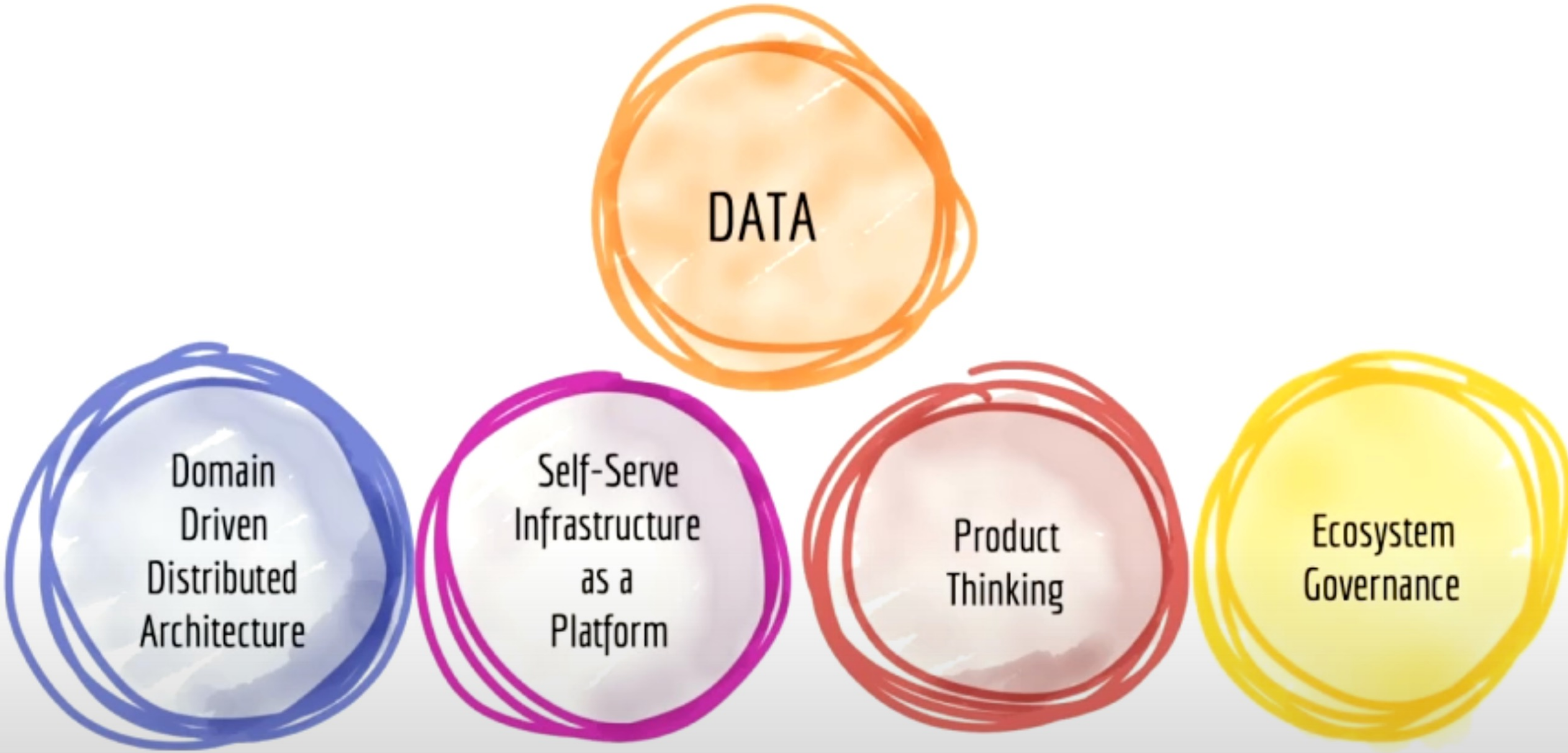


And Allows Us to Bridge Domains

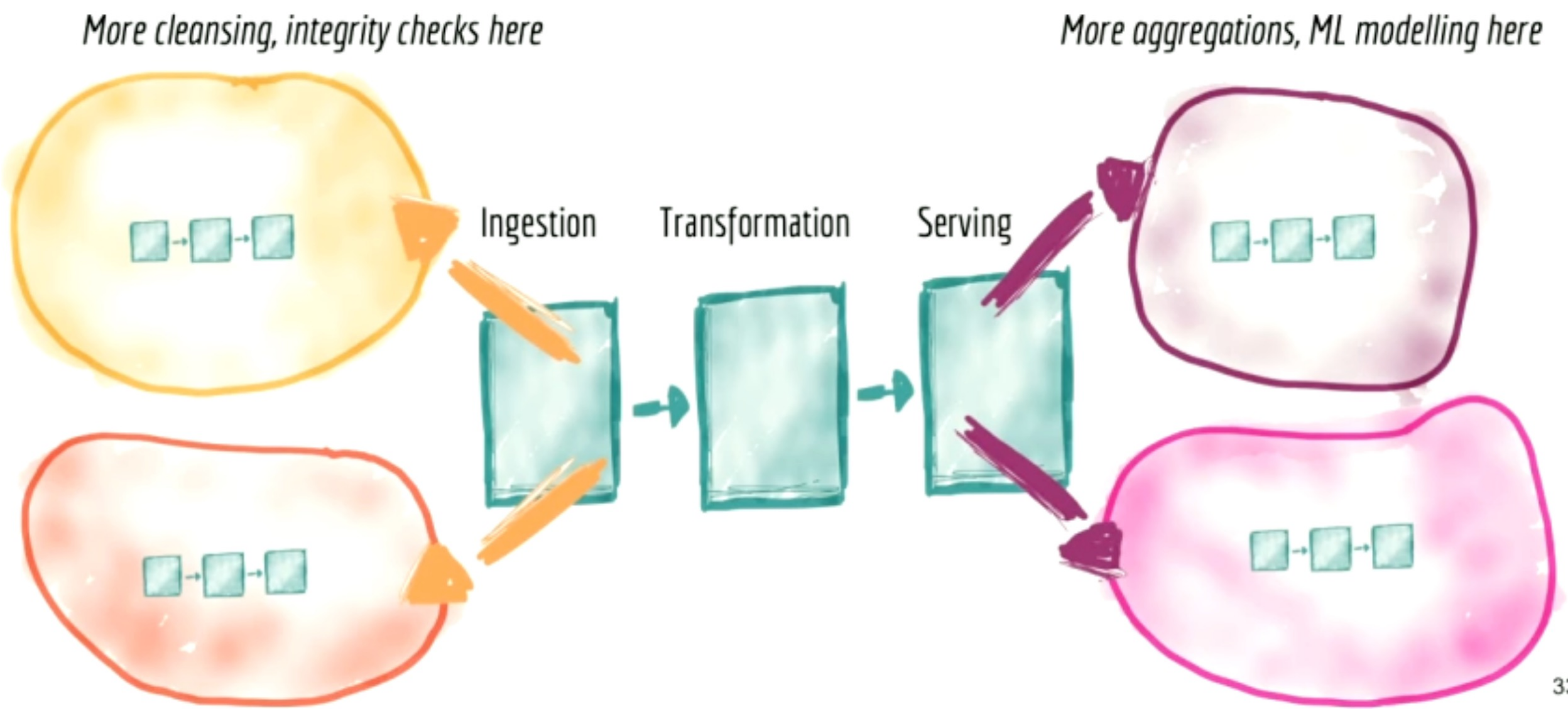


SDU

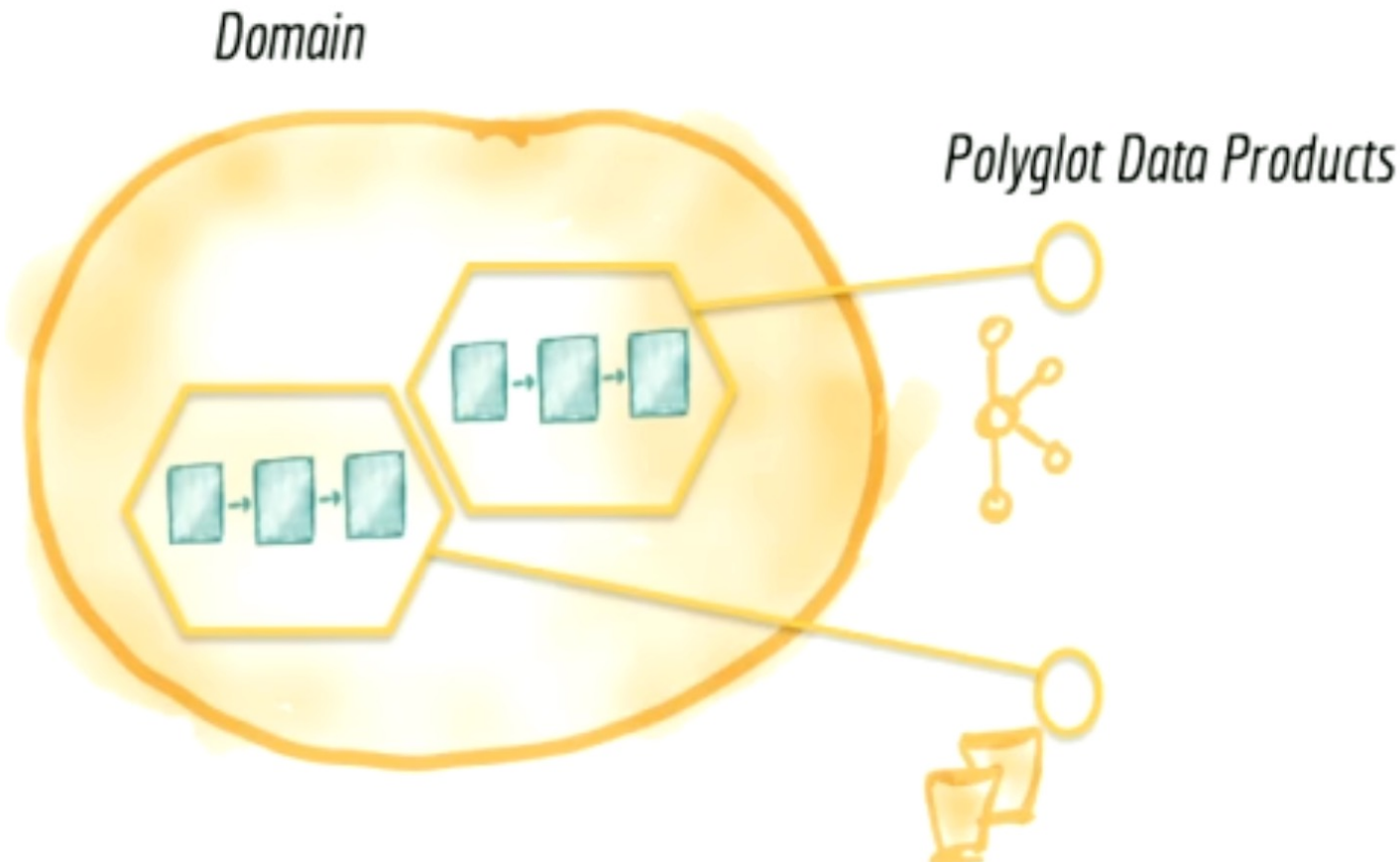
DATA MESH PRINCIPLES



DISTRIBUTED PIPELINES IN DOMAINS



DOMAIN DATA AS A PRODUCT



DOMAIN DATA AS A PRODUCT

Aka Data Products

SHARED | DISCOVERABLE

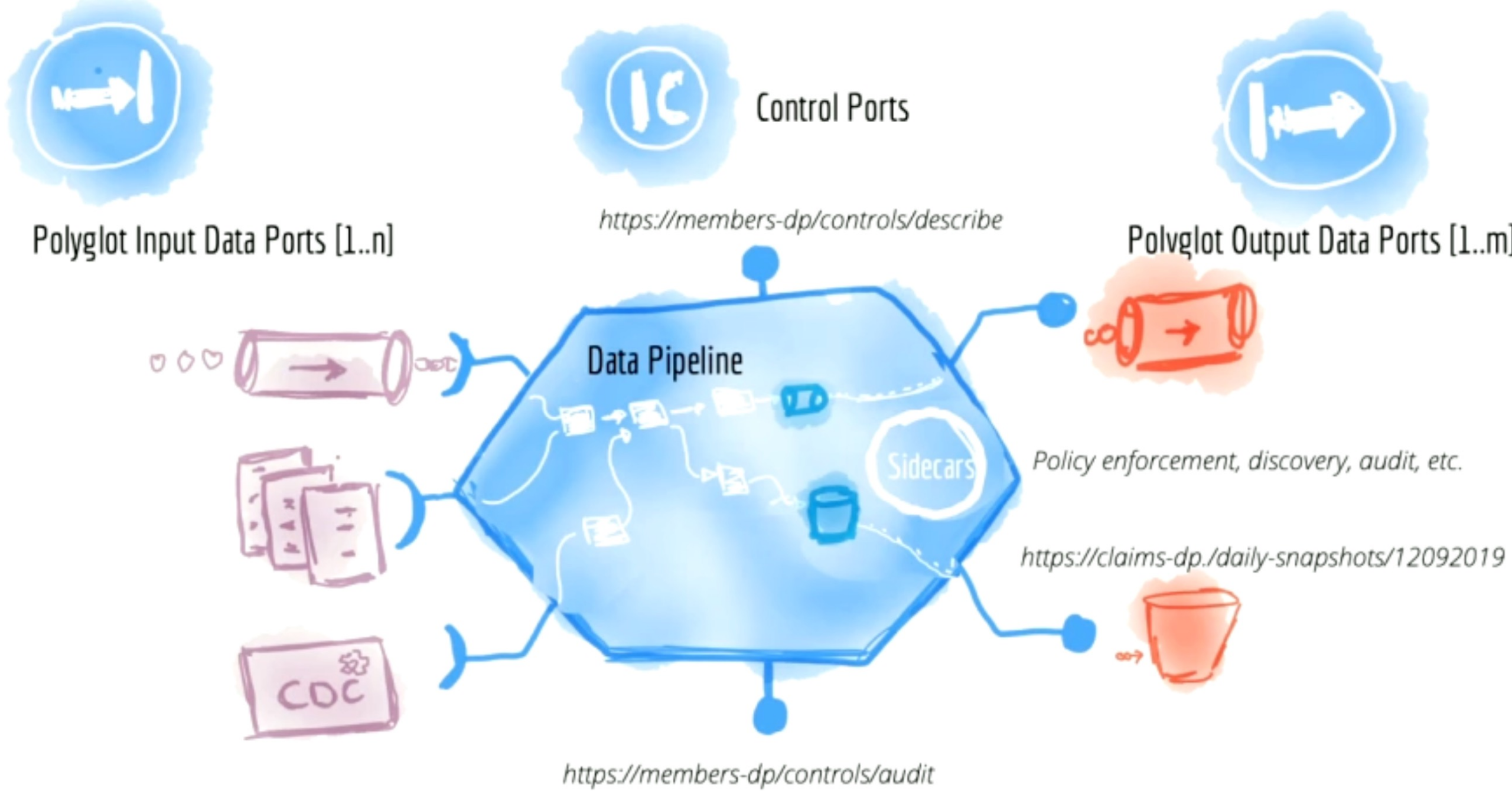
SELF-DESCRIBING

ADDRESSABLE

INTER OPERABLE
(GOVERENED
BY GLOBAL STANDARDS)

TRUSTWORTHY
(DEFINED & MONITORED SLOs)

SECURE
(GOVERENED
BY GLOBAL ACCESS CONTROL)


Control Ports
https://members-dp/controls/describe
Polyglot Input Data Ports [1..n]
Polvglot Output Data Ports [1..m]
Data Pipeline
Sidecars
Policy enforcement, discovery, audit, etc.
https://claims-dp./daily-snapshots/12092019
COC
https://members-dp/controls/audit
On-prem to cloud
On-prem to cloud

DATA MESH PARADIGM SHIFT

FROM	TO
Centralized ownership	Decentralized ownership
Monolithic	Distributed
Pipeline first class concern	Domain data first class concern
Data as a by-product	Data as a product
Siloed data engineering team	Cross-functional data domain teams

ADOPT A NEW LANGUAGE

FROM	TO
Ingesting	Serving
Extracting & Loading	Discovering & Consuming
Flowing data through centralized Pipelines	Publishing output data ports
Centralized Data Lake | Warehouse | Platform	Ecosystem of Data Products

Exercises!