Data-Intensive Distributed Computing

CS 451/651 (Fall 2025)

UNIVERSITY OF WATERLOO

Rubber, Meet Road: Data Management
(v1.0)

Week 5: October 2nd

Khaled Ammar
Rocket Innovation Studio

These slides are available at `https://lintool.github.io/cs451-2025f/`

This Week

Tuesday: Orchestrator

DAGs, ML pipelines, Airflow, etc.

Now: Data Management in Production

Data Governance, Metadata, Feature Stores

Is orchestrator enough for data pipeline?

Why Orchestration Alone is not enough?

- Discoverability
- Lineage
- Quality
- Governance

Tables: Transactions_cleaned_v2 & Clients_full_v4

Questions:

1. Who owns this data?
2. What does each column mean?
3. How was the data prepared? cleaning process?
4. What is the refresh rate?
5. Can we legally use all columns?
6. …

Who has chased a “mystery dataset” in a project before?

Data Catalog

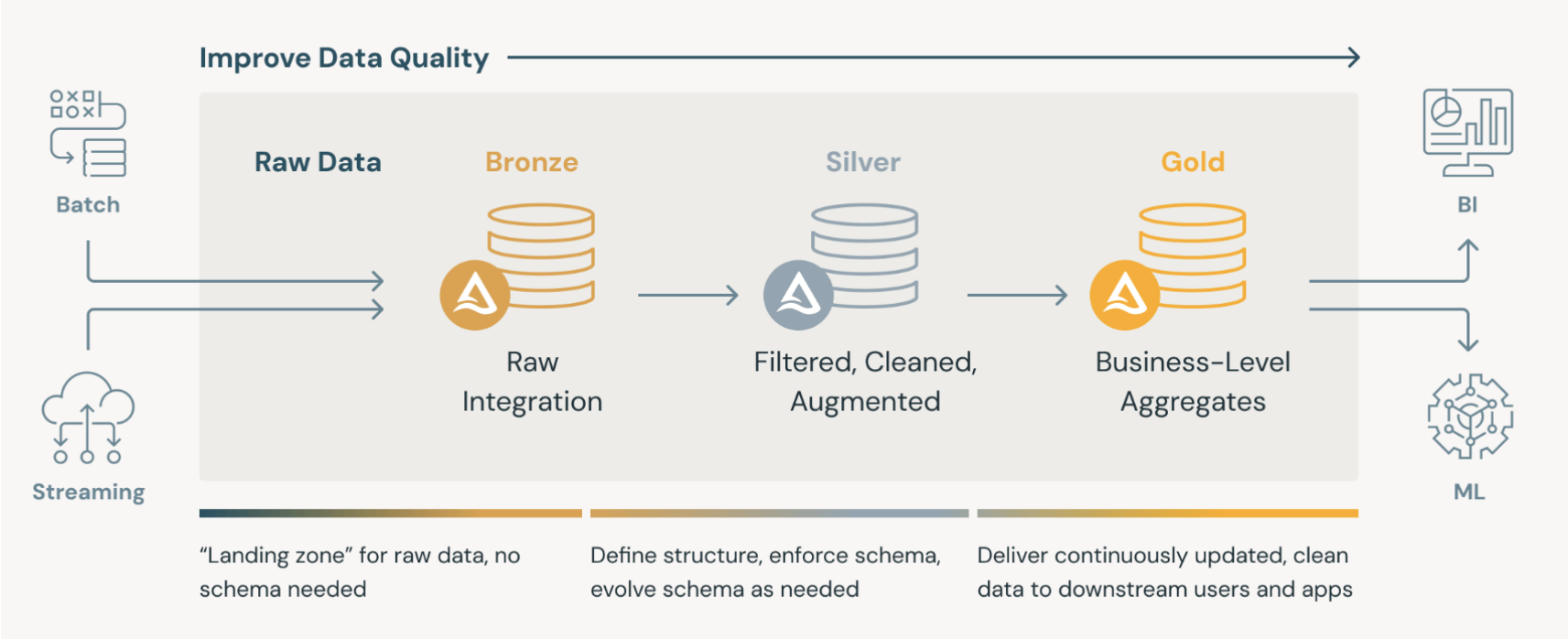
A *data catalog* is the system of record for your data assets.

It tracks:

- **what exists (tables, views, files, ML models),**
- **who owns them,**
- **how they're structured (schema),**
- **how they're used (popularity),**
- **and where they came from (lineage).**

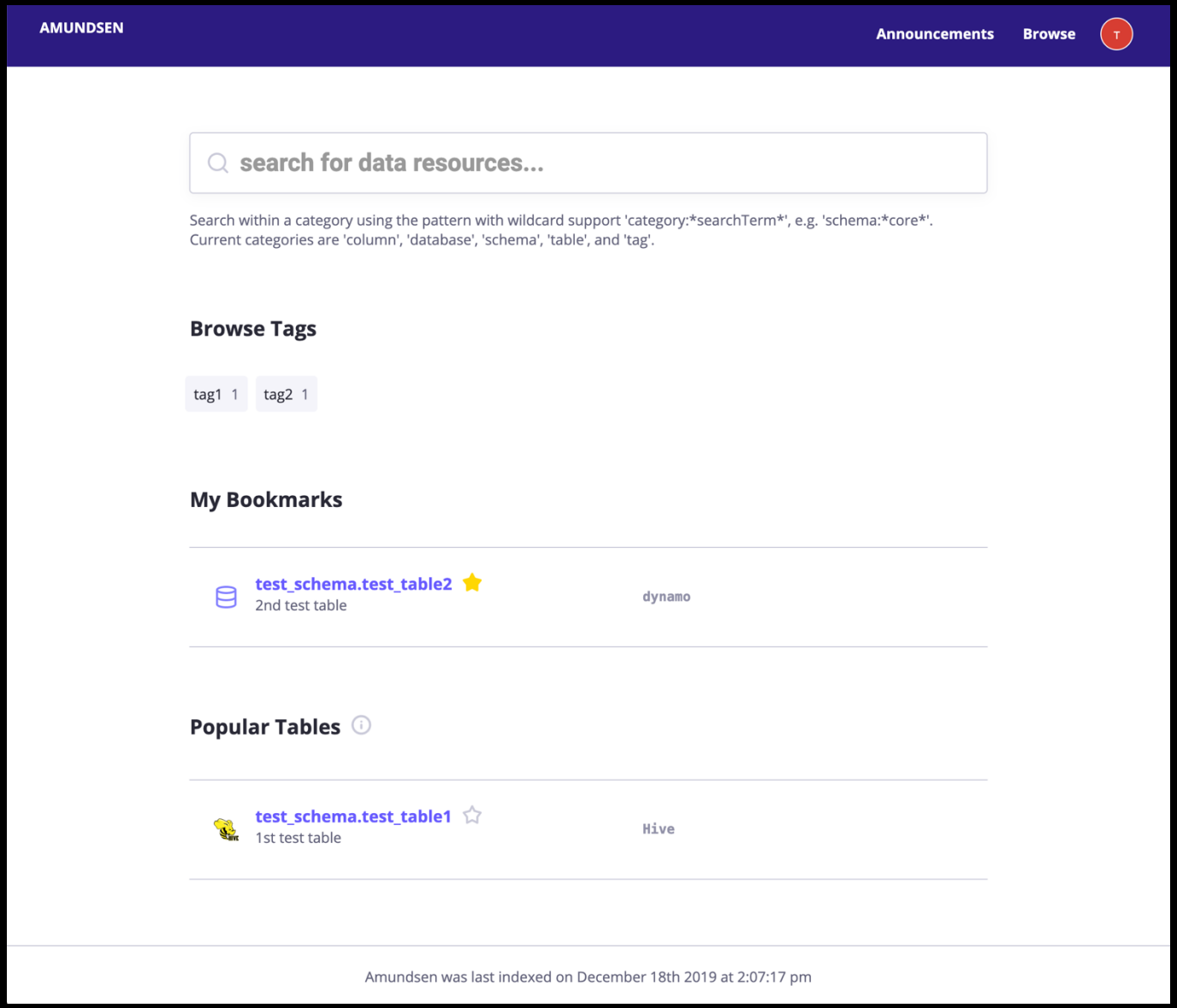
Data Lineage

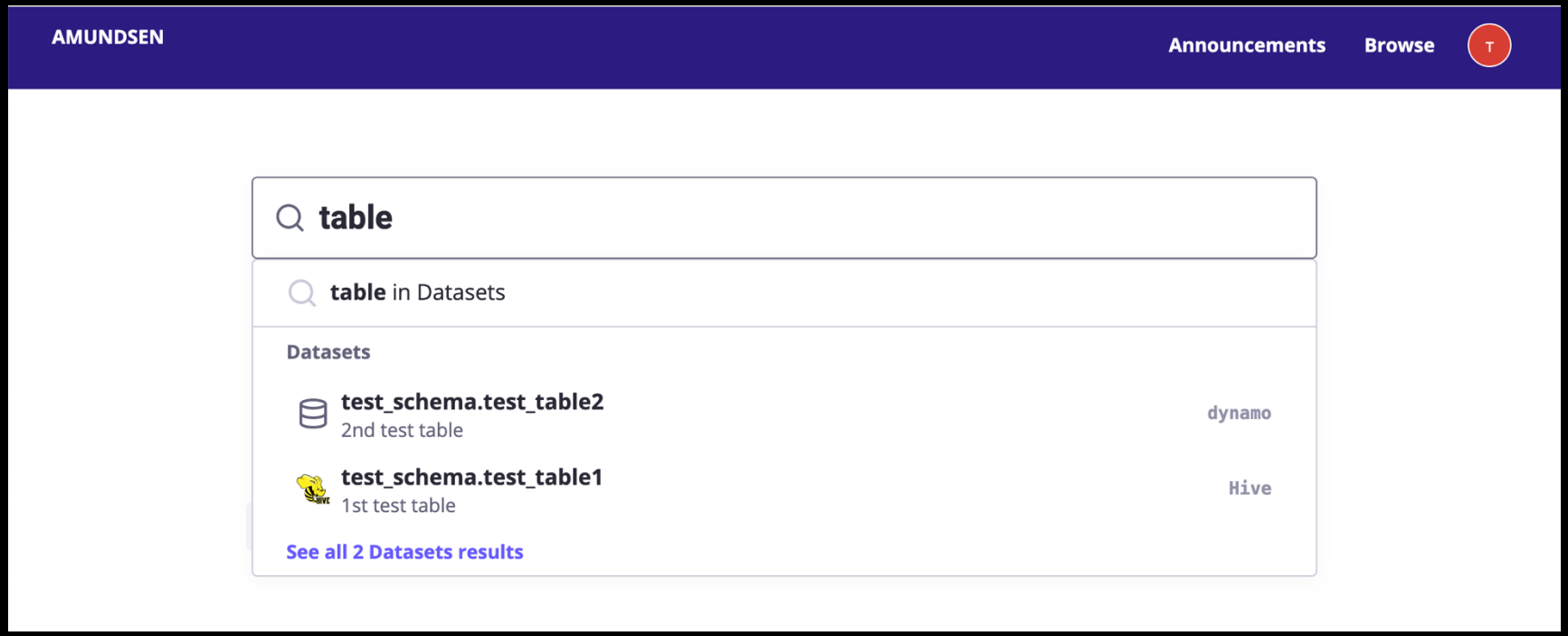
Record how data was constructed – useful for debugging and investigation.

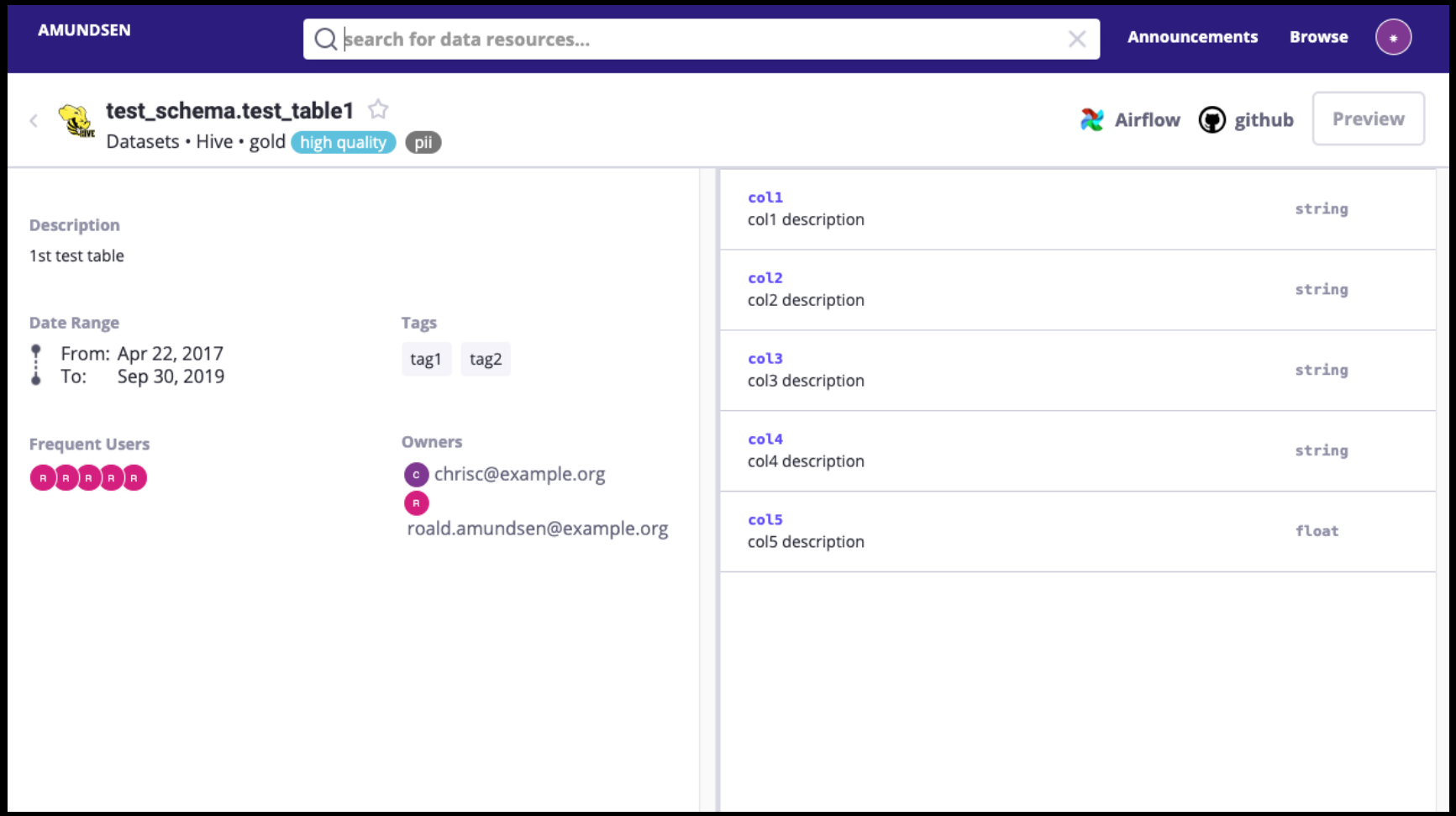


Data Catalog Tools

Tool	Features	Limitations	Best fit
Unity Catalog	• Centralized governance for tables, files, ML models • Fine-grained permissions • Integrates with Delta Lake & MLflow	Locked into Databricks ecosystem	Companies standardized on Databricks lakehouse
AWS Glue Data Catalog	• Central metadata store • Schema inference • Lake Formation integration	weaker UX and less business metadata support.	AWS-native data lakes with Athena/Redshift
DataHub (LinkedIn OSS)	• Active lineage • Rich metadata model (technical + business) • Search & discovery UI	More engineering effort to run & maintain	Heavy engineering-focused companies.
Amundsen (Lyft OSS)	• Metadata search engine + lineage • Simple interface (search-first) • Integrations with warehouses + BI tools	Less feature-rich than DataHub;	Quick-start OSS catalog for BI/analyst teams


AMUNDSEN
Announcements
Browse
T
search for data resources...
Search within a category using the pattern with wildcard support 'category:*searchTerm*', e.g. 'schema:*core*'.
Current categories are 'column', 'database', 'schema', 'table', and 'tag'.
Browse Tags
tag1 1
tag2 1
My Bookmarks
test_schema.test_table2
2nd test table
dynamo
Popular Tables
test_schema.test_table1
1st test table
Hive
Amundsen was last indexed on December 18th 2019 at 2:07:17 pm


AMUNDSEN
Announcements
Browse
T
table
table in Datasets
Datasets
test_schema.test_table2
2nd test table
dynamo
test_schema.test_table1
1st test table
Hive
See all 2 Datasets results


AMUNDSEN
search for data resources...
Announcements
Browse
test_schema.test_table1
Datasets • Hive • gold high quality pii
Airflow
github
Preview
Description
1st test table
Date Range
From: Apr 22, 2017
To: Sep 30, 2019
Tags
tag1
tag2
Frequent Users
Owners
chrisc@example.org
roald.amundsen@example.org
col1
col1 description
string
col2
col2 description
string
col3
col3 description
string
col4
col4 description
string
col5
col5 description
float

col1

`string`

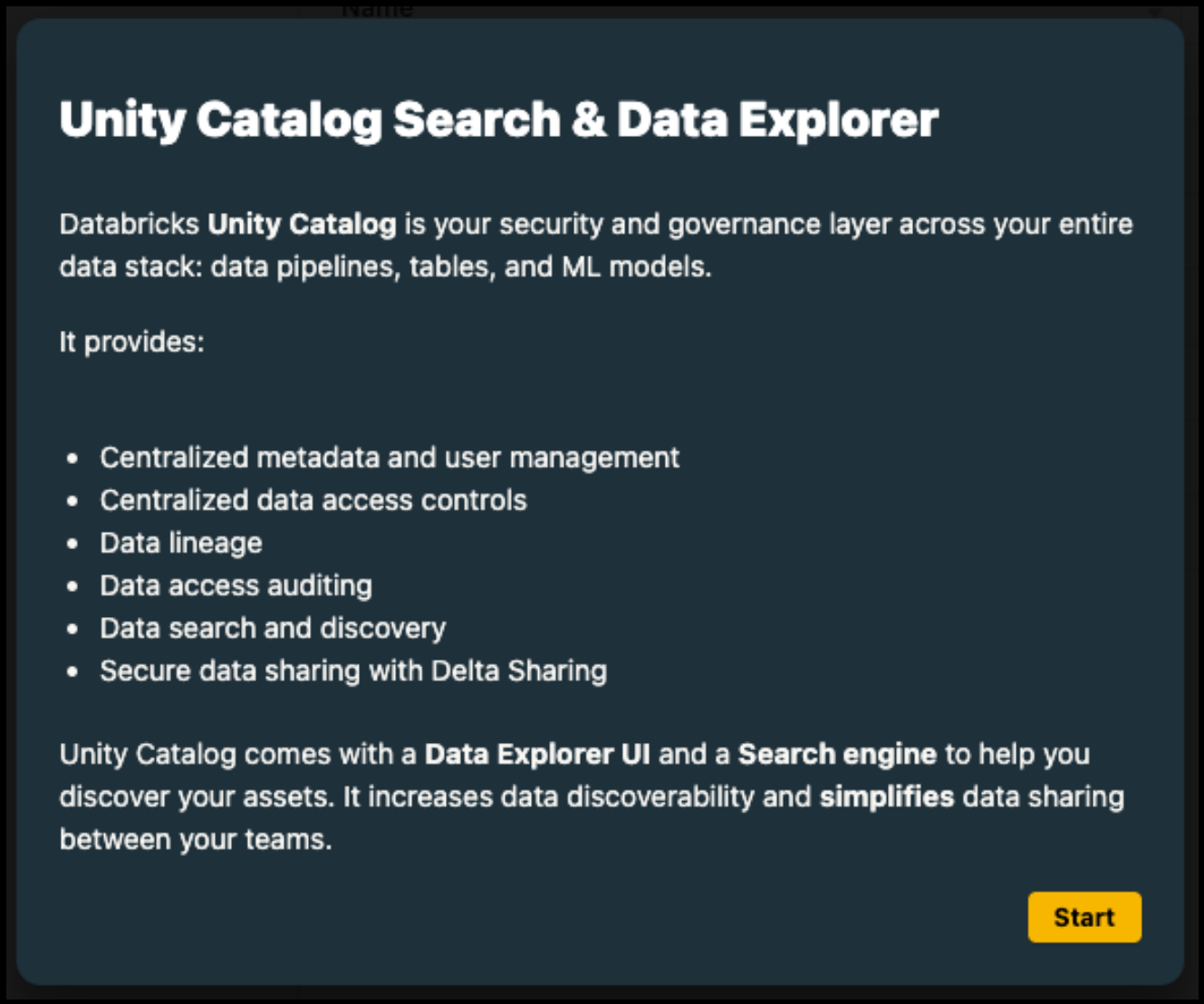
Description

This is an editable test description for the first column. This also supports **Markdown**.

Column Statistics *Stats reflect data collected between May 22, 2015 and Jul 04, 2019.*

distinct values	8	min	aardvark
num nulls	500320	max	zebra
verified	230430		

Unity Catalog Demo



https://app.getreprise.com/launch/96mpAqy/

Data Catalog
Rubber, meet road!

- Catalogs rot without ownership; make “owner” a required field.
- Generate lineage from ETL/code:
  - but validate critical edges; 100% automatic is rare.
- Don’t flood users with noisy assets
  - Categorize data assets as bronze, silver, gold;
  - Most users need “gold” datasets.
- You may need to have multiple catalogs, not ideal but it happens.

Data Quality

Garbage In,
Garbage Out

The cost of bad data represents an astonishing

15% to 25%

of the revenue for most companies

Reference: https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/

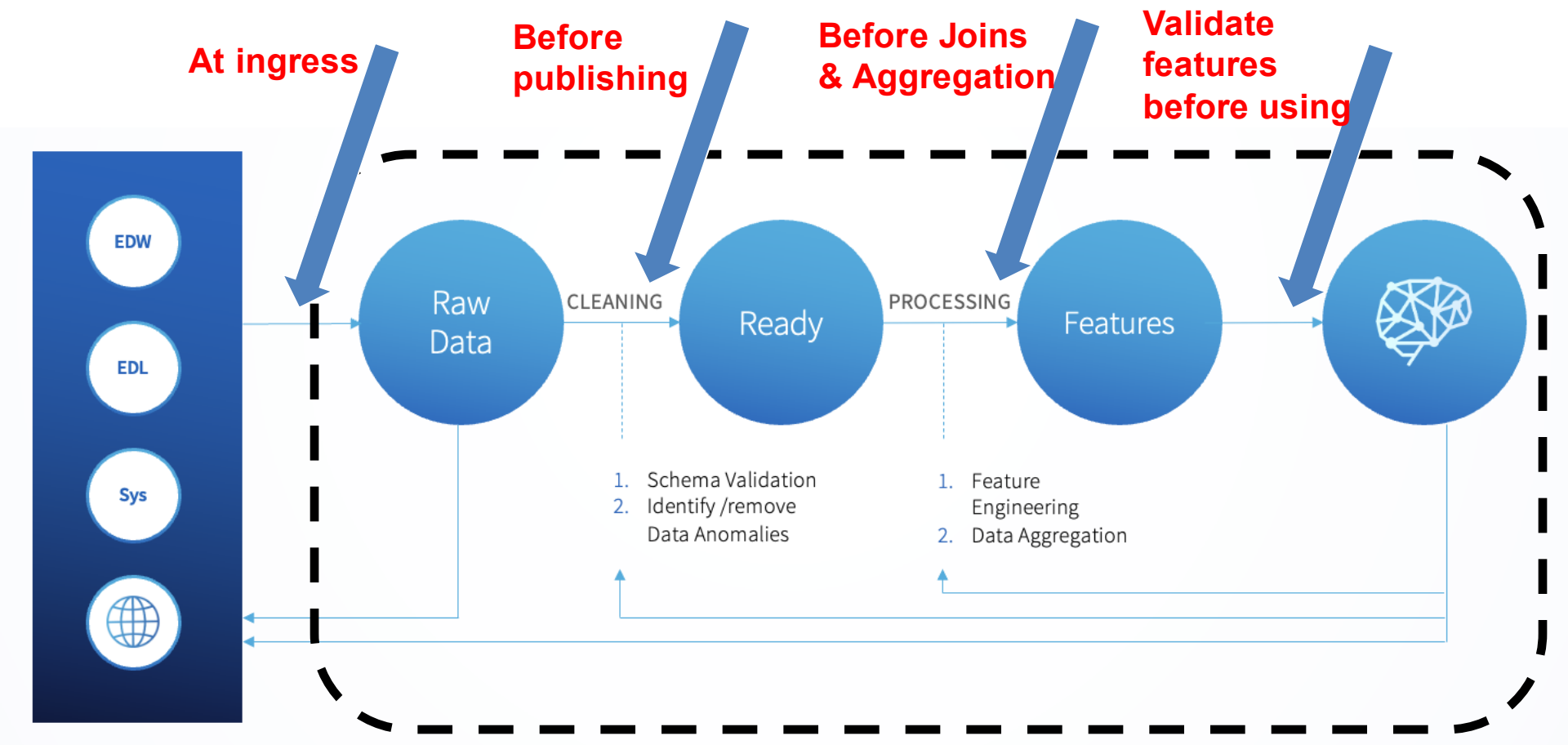
What should we check?

Short answer: everything!

Possible kind of checks:

- Schema checks: data types, not-null columns, unique keys, etc.
- Value checks: ranges, possible values, reference to other tables.
- Distribution checks: monitor drift as system progress, outliers, ratio of null values.
- Freshness: data update frequency.
- Business rules: Same level California employee makes more than Ohio.

When should we check?



Data Cleaning

Exercise:
For an events table like the assignment: (event_time, user_id, event_type, amount)

What expectations (aka quality checks) you should add?

Data Cleaning
Rubber, meet road!

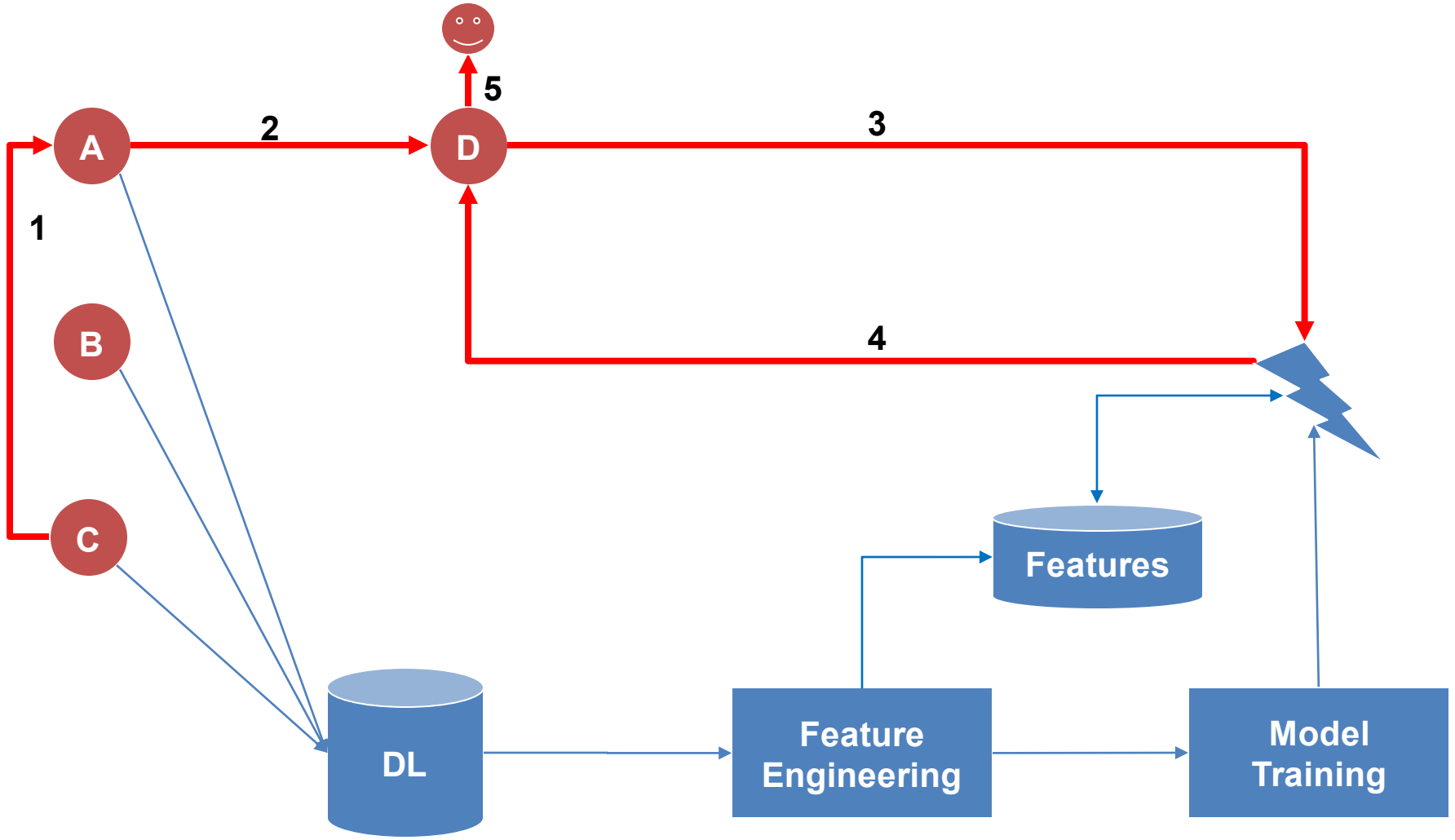
- Same column may mean different things based on context.

Example:
For a company that manufacture toys, and sell them to retailers:
What does revenue mean?

- Very strict rules cause many false alarms and often slow everything, or get ignored.

- Tests hidden in code are forgotten!
  - Centralize checks in configs
  - Surface checks in catalogs
  - Automate checks when possible

Feature Store, Why?

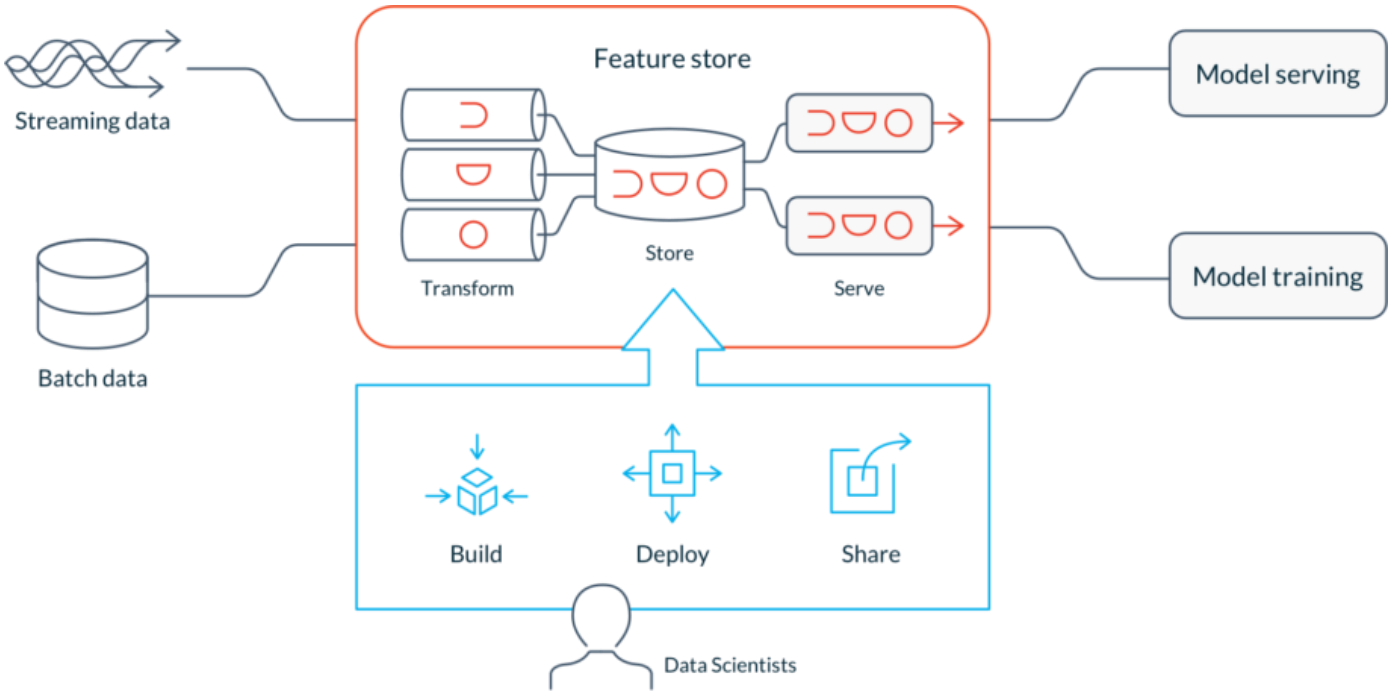


Feature Store

Features must be consistent between training and serving!

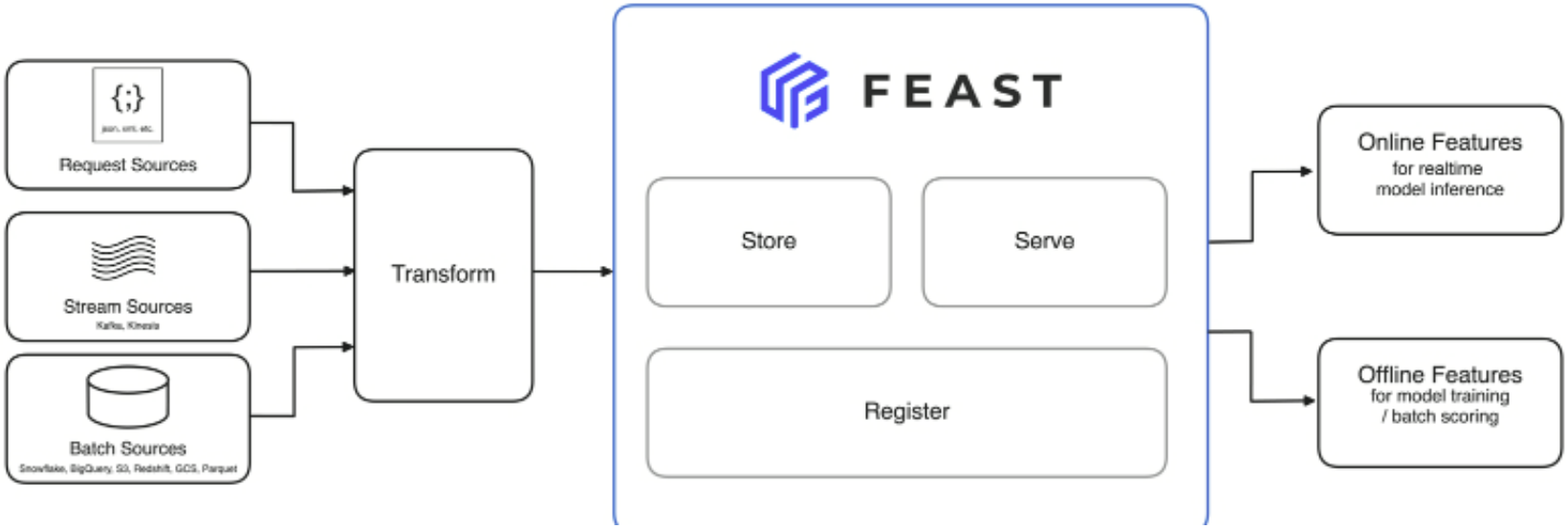
Feature Store

Implementations (Tecton)



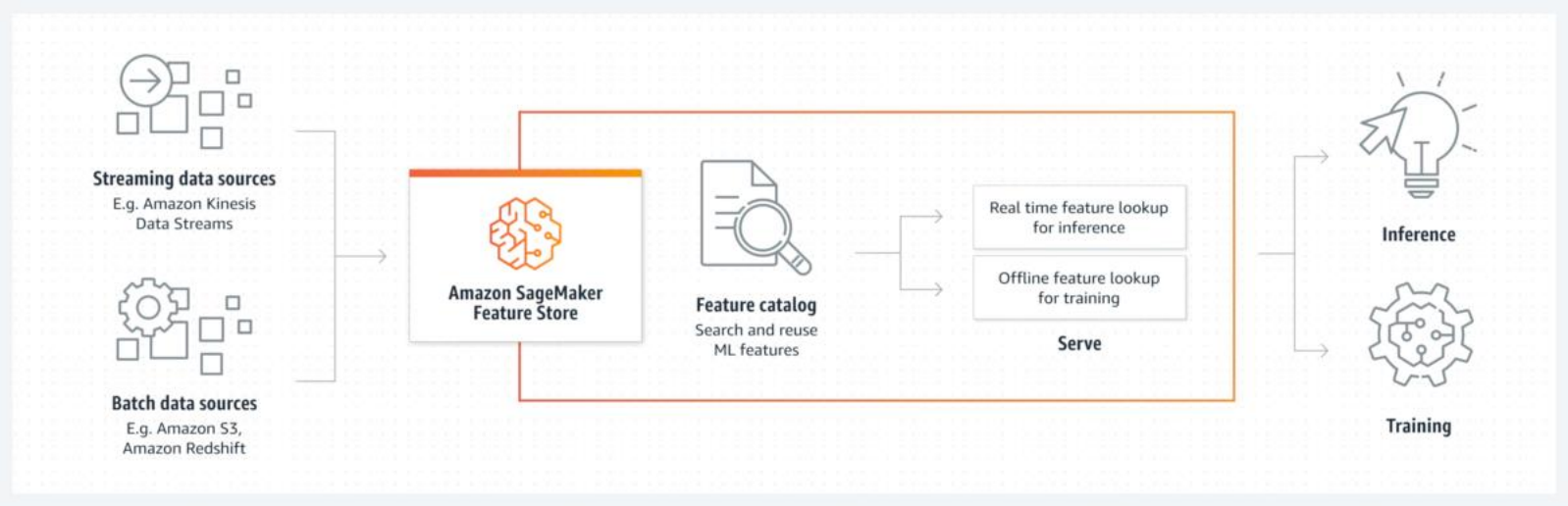
Feature Store

Implementations (FEAST)



Feature Store

Implementations (AWS SageMaker)



Feature Store
Architecture

- **Offline store:** Parquet files, S3 buckets, etc.
- **Online store:** Redis/DynamoDB
- **Feature Catalog & Lineage:** obviously ☺
- **Integration with model registry:** part of lineage

Question: Why do we need an online store instead of using DW?

Feature Store

Increase features quality

- Point-in-time correctness: prevent leakage
- Backfill when feature logic changes
- Cached features expires in online store

Feature Store

Rubber, meet road

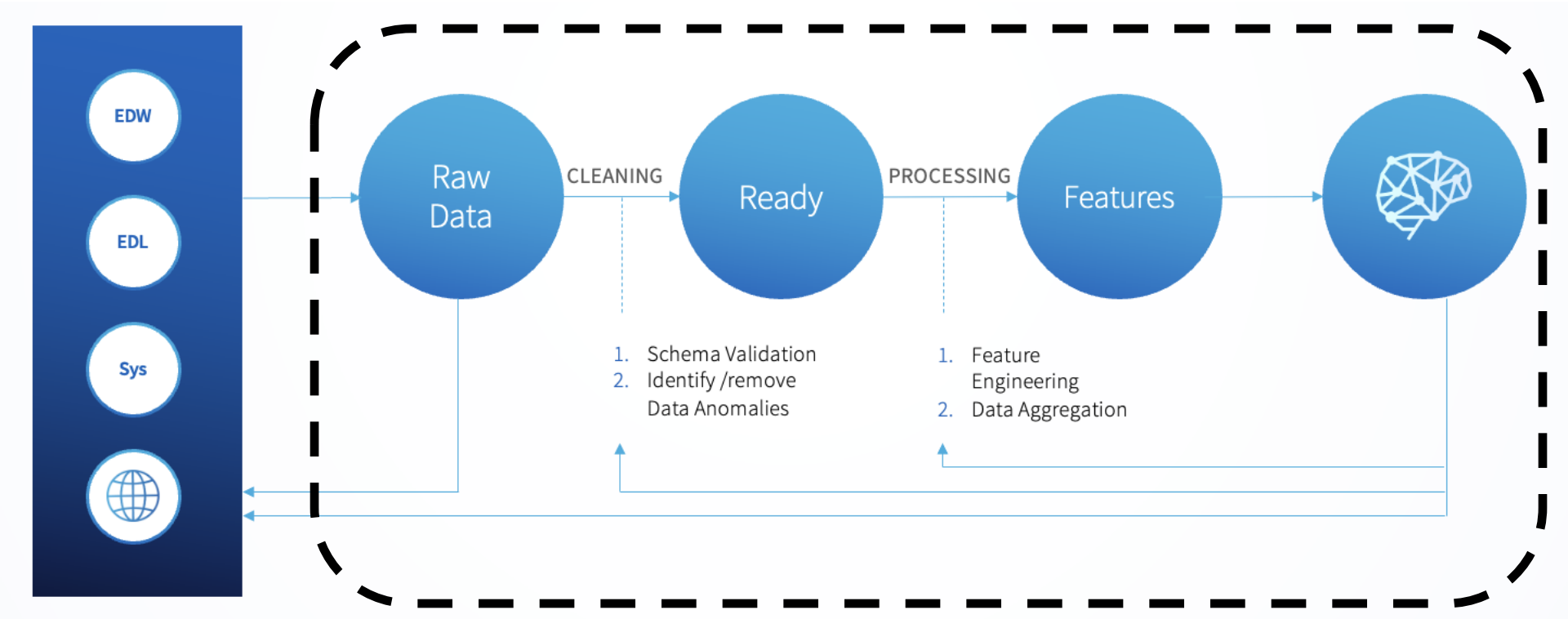
A big value for feature store, is cultural. It helps data teams share definitions and assumptions and start a dialogue about expectations.

Data Governance

- Legal vs Ethical
- Balance between “moving fast breaking things”; and legacy systems!
- You need governance to empower teams!
- PII data is sacred, in many industries.

Data Governance

safe innovation environment



Data Management

Rubber, meet road.

- Data management empower teams to build trustworthy data-enabled systems.
- **Catalog & lineage:** Important for data discovery & governance (owners)
- **Quality:** Bad data is the norm; stop it from being used in your system.
- **Feature Store:** A framework for consistent, low-latency ML features.
- **Governance:** empowers innovators to work in a safe environment.

Thank you!