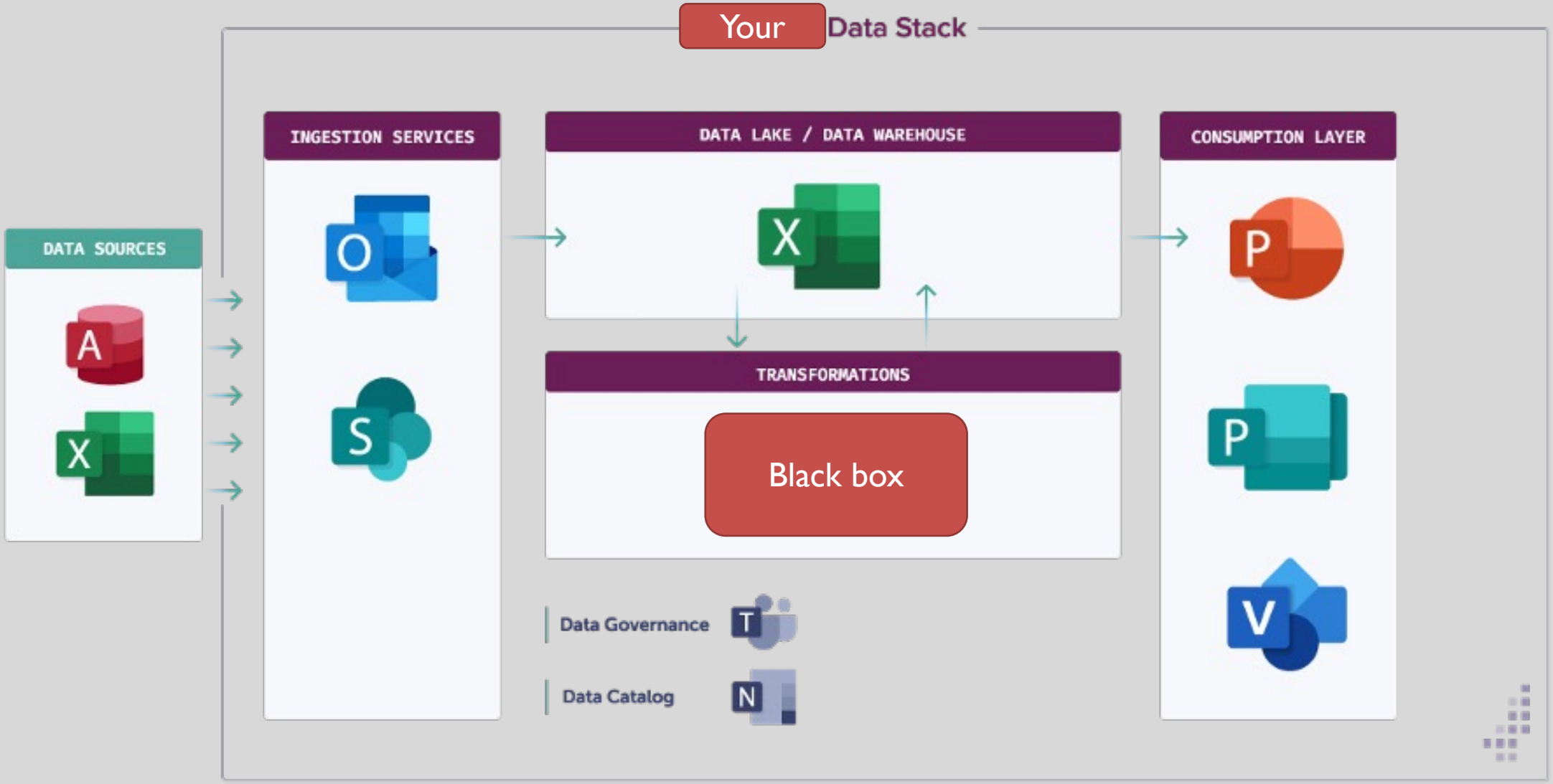
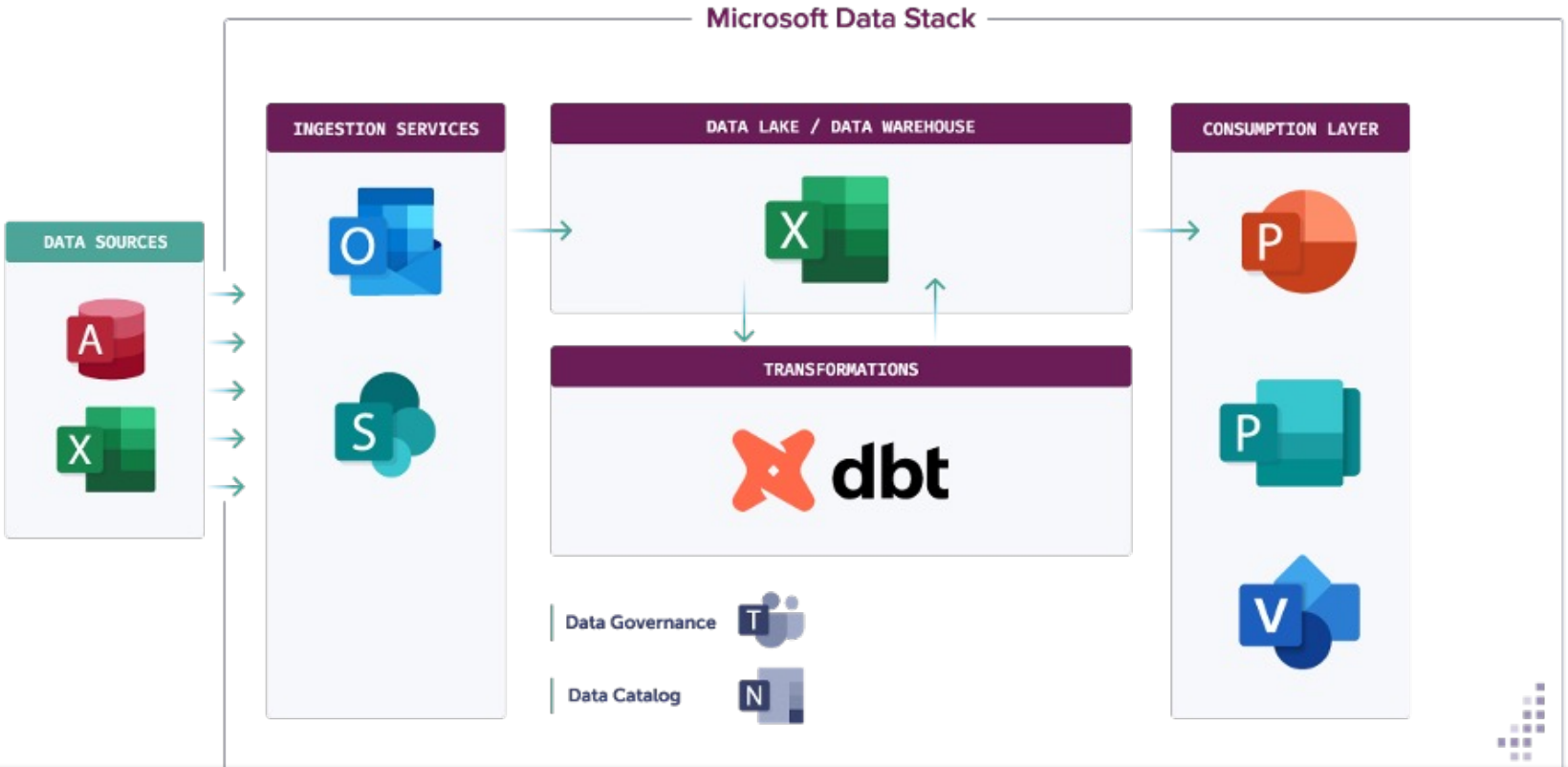
# From 0 6 * * * to 🚀

Georg Heiler | 2023

from cron-based distributed data monoliths
towards lazily-coupled data products

T Magenta


Your
Data Stack
DATA SOURCES
INGESTION SERVICES
DATA LAKE / DATA WAREHOUSE
CONSUMPTION LAYER
TRANSFORMATIONS
Black box
Data Governance
Data Catalog


Microsoft Data Stack
DATA SOURCES
INGESTION SERVICES
DATA LAKE / DATA WAREHOUSE
CONSUMPTION LAYER
TRANSFORMATIONS
dbt
Data Governance
Data Catalog

# DBT-EXCEL : )

# About me:
# georgHeiler.com

Lecturer, speaker, meetup organizer (VDSG)

Senior software engineer with a specialization in data

PhD in informatics (TU Wien, CSH)

# Agenda

TRADITIONAL PLATFORM OVERVIEW

FUTURE STACK

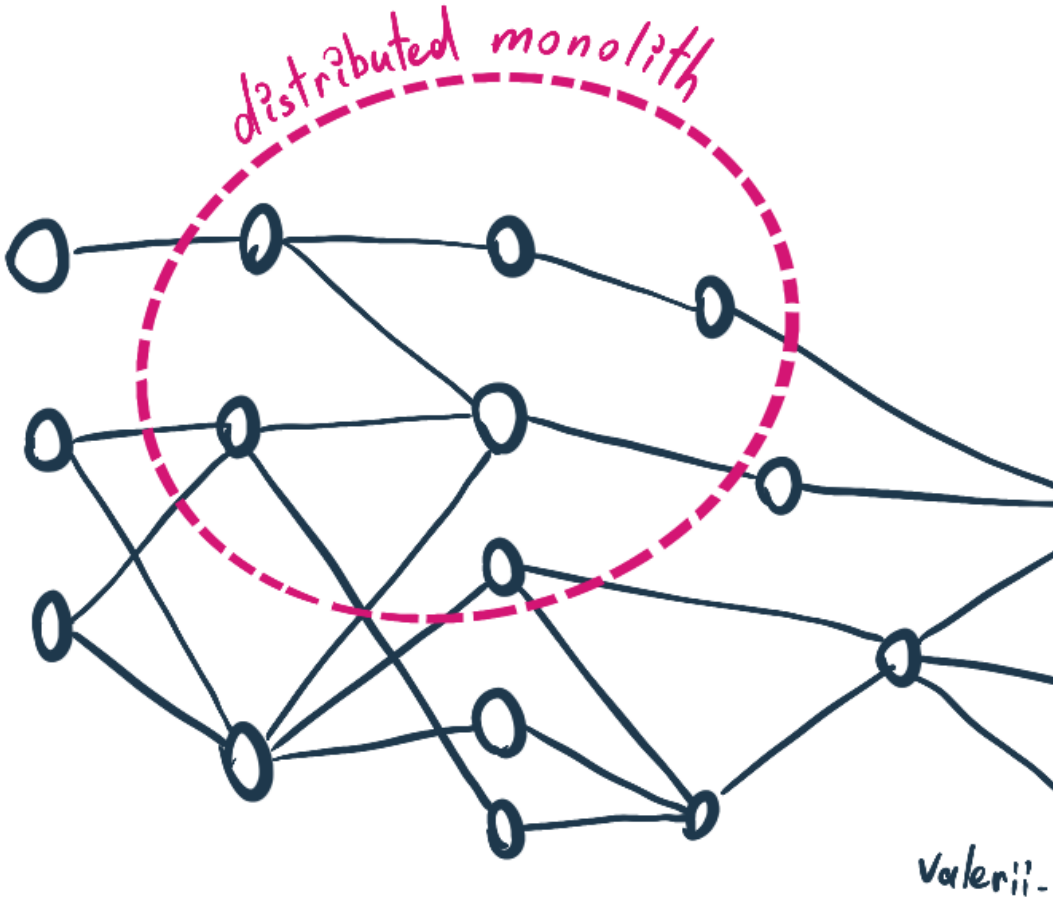
DEMO

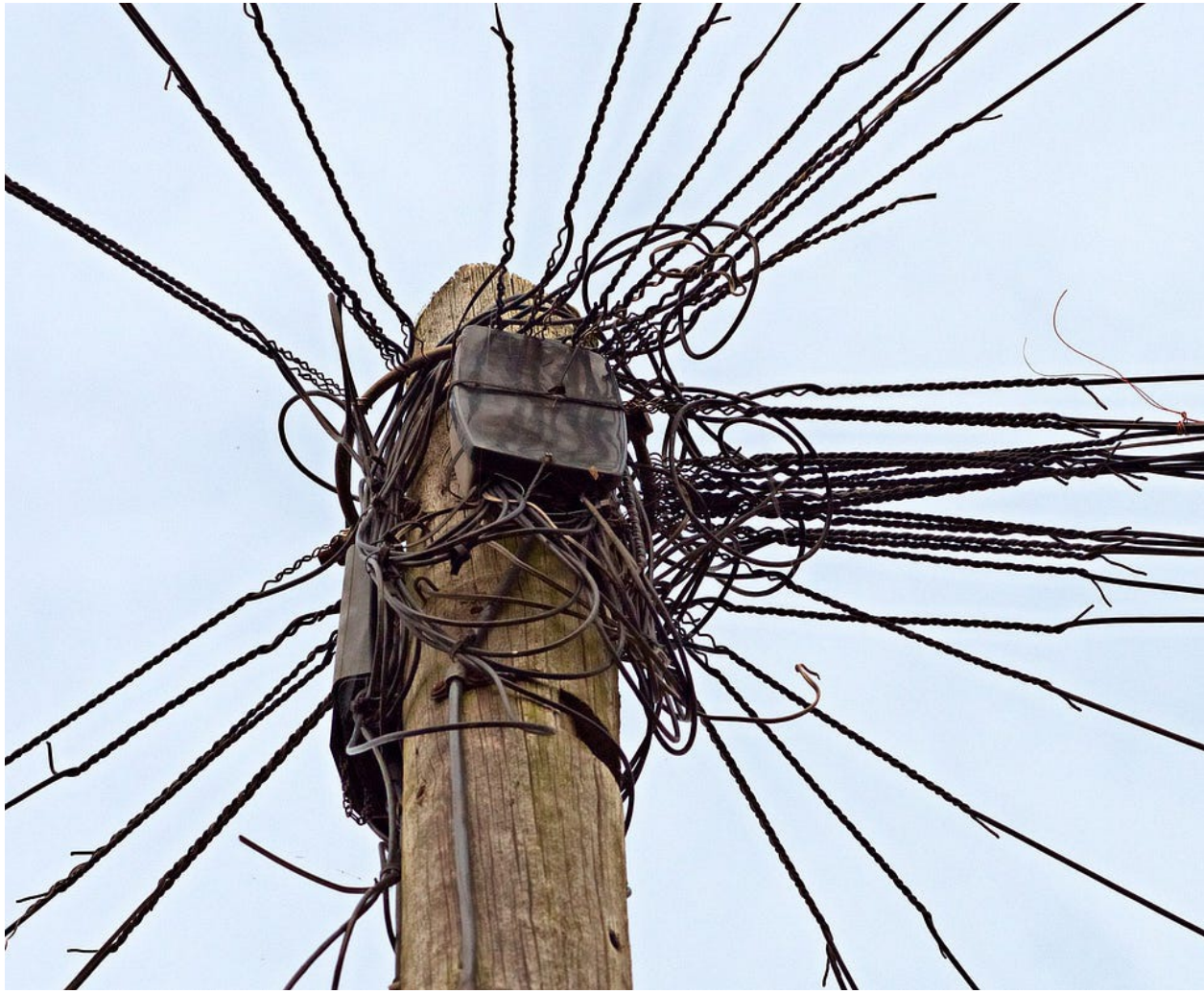
# Traditional data stack (ETL)

- (Often) custom ingestion processes
- Data warehouse transformations with proprietary tools (Informatica, Talend, plsql, …)
  - Not a single source of truth
  - Not a single E2E lineage
  - Multiple separate transformation tools (used by various departments)
  - No central scheduler
- No clear separation of layers and domains
- Layer over layer – starving in abstractions

- Data mart presentation layer consumed by tools like Tableau, Qlick, …
- Way too often (hot) fixes are deployed in analytical systems
- Tools chosen where some other joinable data resides (not DDD)

# Distributed (data) monolith

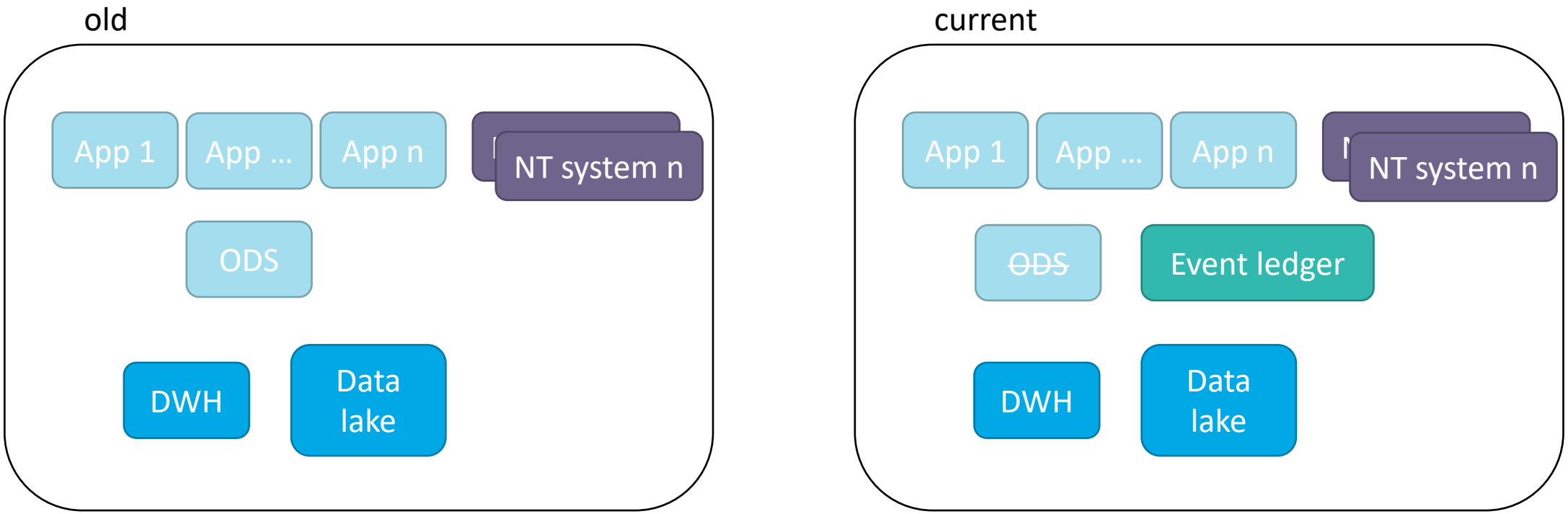


Sources: https://www.google.com/imgres?imgurl=https%3A%2F%2Fvalerii-udodov.com%2Fimages%2Fposts%2Fmonolithic-architecture%2Fdistributed_monolith.png&tbnid=QcgGICV_J3xE6M&vet=12ahUKEwjZ14P6hbb-AhVYkScCHUMBArIQMygNegUIARDRAQ..i&imgrefurl=https%3A%2F%2Fvalerii-udodov.com%2Fposts%2Fmonolithic-architecture%2F&docid=L2B334OolhZMAM&w=1076&h=646&q=distributed%20monolith&client=firefox-b-d&ved=2ahUKEwjZ14P6hbb-AhVYkScCHUMBArIQMygNegUIARDRAQ

https://www.google.com/imgres?imgurl=https%3A%2F%2Fmiro.medium.com%2Fv2%2Fresize%3Afit%3A1400%2F1*TqsfHvBEfvFqXdbSS-0Ypg.jpeg&tbnid=OB9PNkUyixC4CM&vet=12ahUKEwjZ14P6hbb-AhVYkScCHUMBArIQMygzegQIARBB..i&imgrefurl=https%3A%2F%2Fblog.devgenius.io%2Fdistributed-monolith-1d2d9f86a68f&docid=gzcJzqtPSDtojM&w=1400&h=794&q=distributed%20monolith&client=firefox-b-d&ved=2ahUKEwjZ14P6hbb-AhVYkScCHUMBArIQMygzegQIARBB

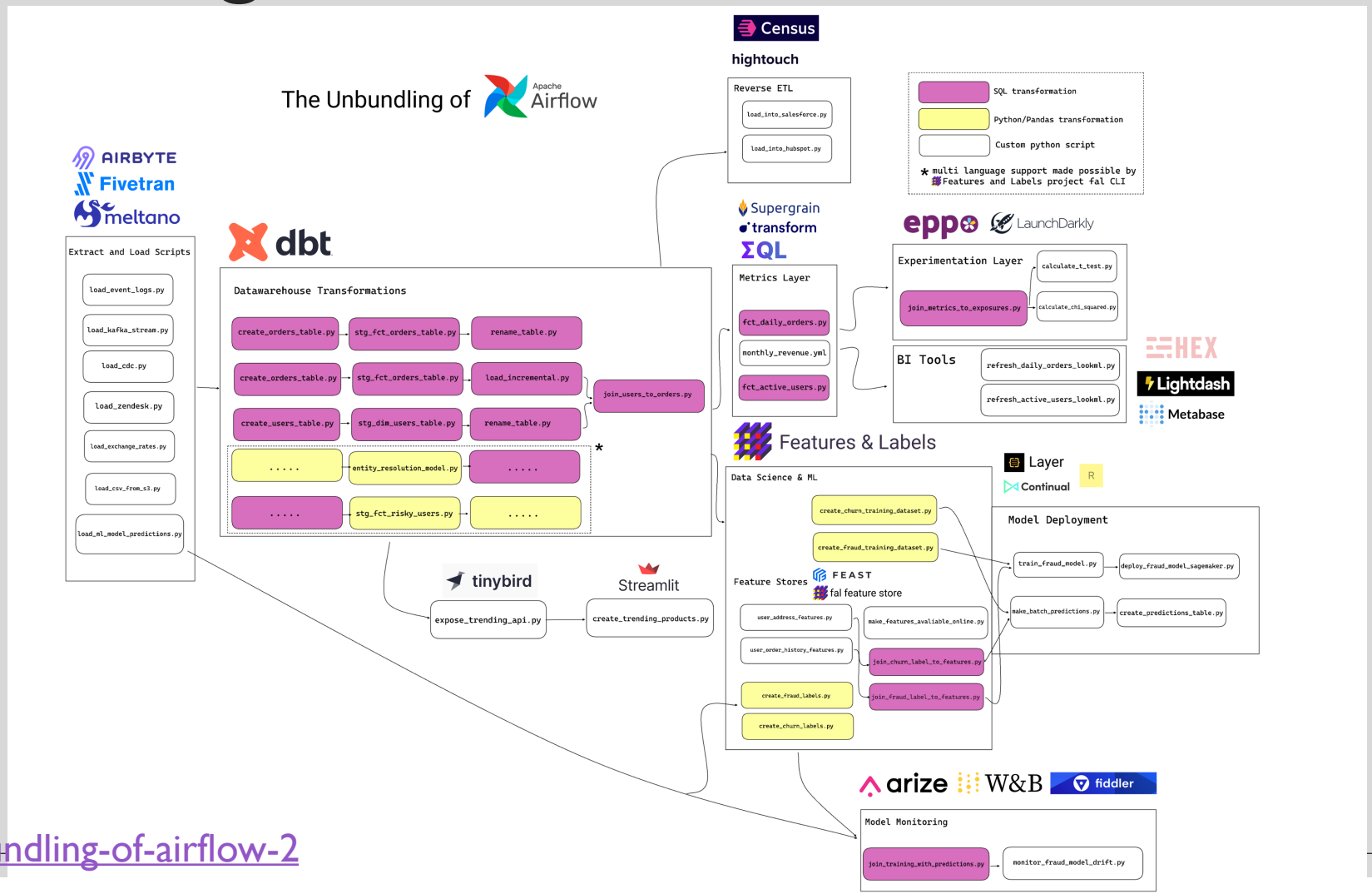
# Data platform



# TECH | MDS

How well are *modern* tools integrated?

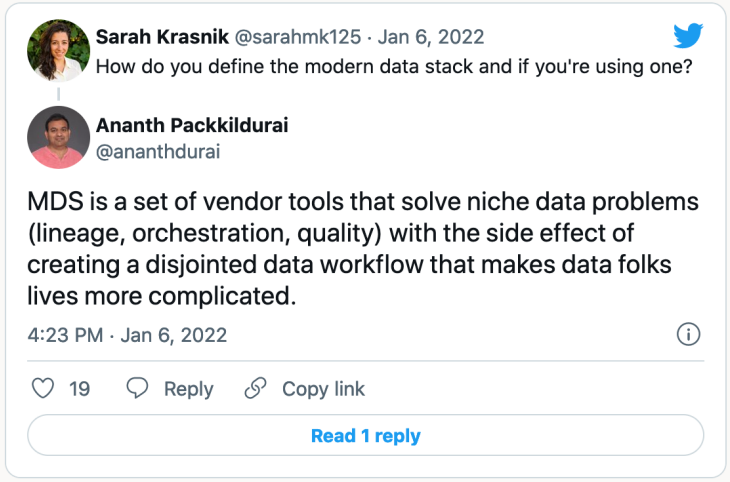
# Unbundling Airflow: Silos? Orchestration?

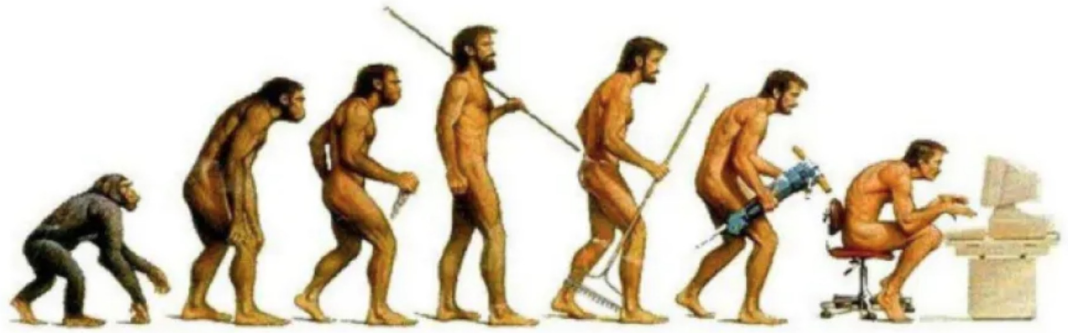



blog.fal.ai/the-unbundling-of-airflow-2


WHAT
MANAGES THE
MANAGED
SERVICE?

Ananth Packkildurai, the author of the Data Engineering Weekly newsletter, summarizes this state of affairs well:



Overlapping Crons | Workflow Engines | Overlapping Crons in MDS

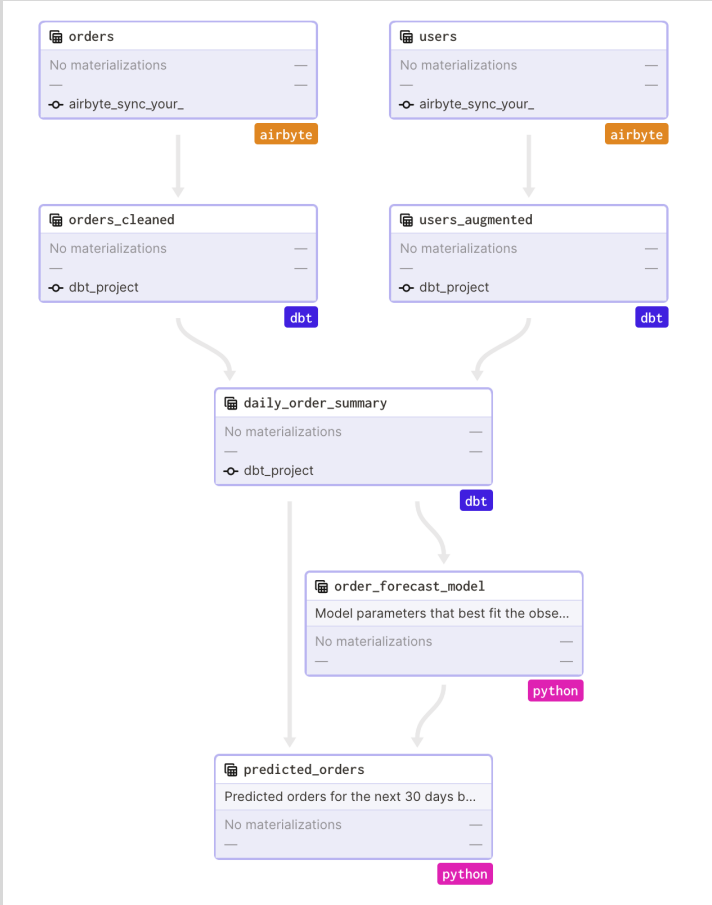
https://dagster.io/blog/rebundling-the-data-platform

# UNBUNDLING AIRFLOW:

# SILOS? ORCHESTRATION?

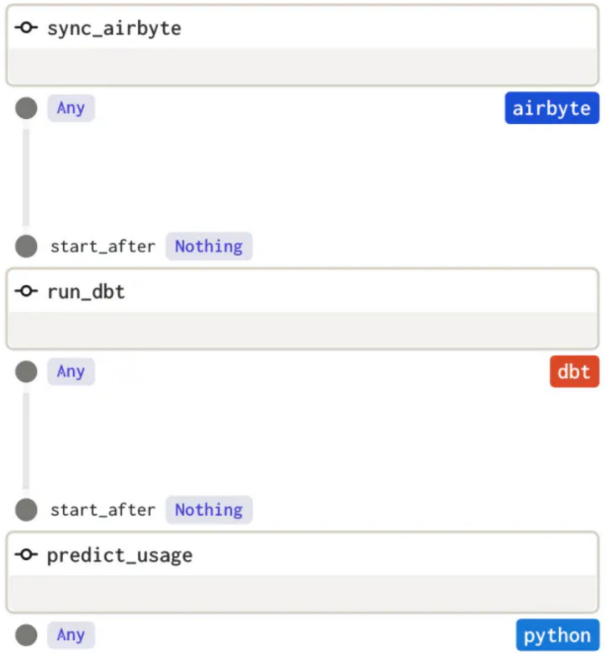
# Dagster: Overall orchestration

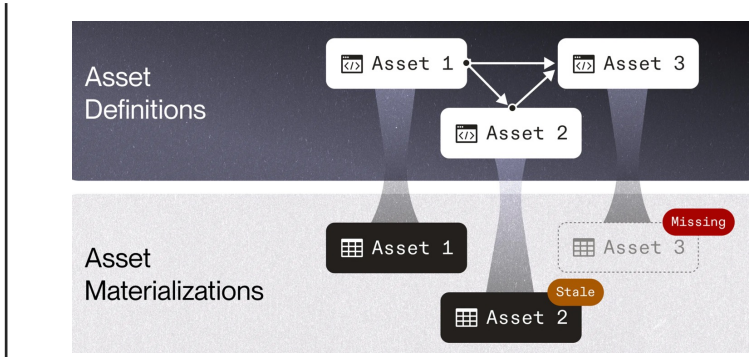
- Alleviate Airflow`s problems
  - Testability first (resources allow for separation of business logic and IO, cloud services, APIs)
  - Increase developer productivity (i.e. locally E2E DEV-test the pipeline with local resources
  - Native data dependencies
  - E2E orchestration (ingest, transform, ML)
    - Lineage first – improve governance
- Assets: Turning the pipeline inside out ➔ Rebundling

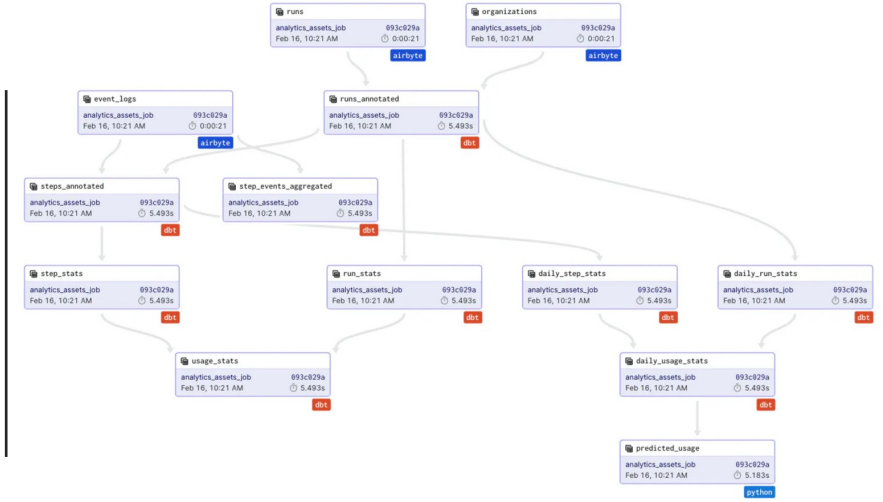



georgheiler.com/2022/03/04/connector-goodness-from-airbyte-e2e-lineage

# REBUNDLING WITH DAGSTER

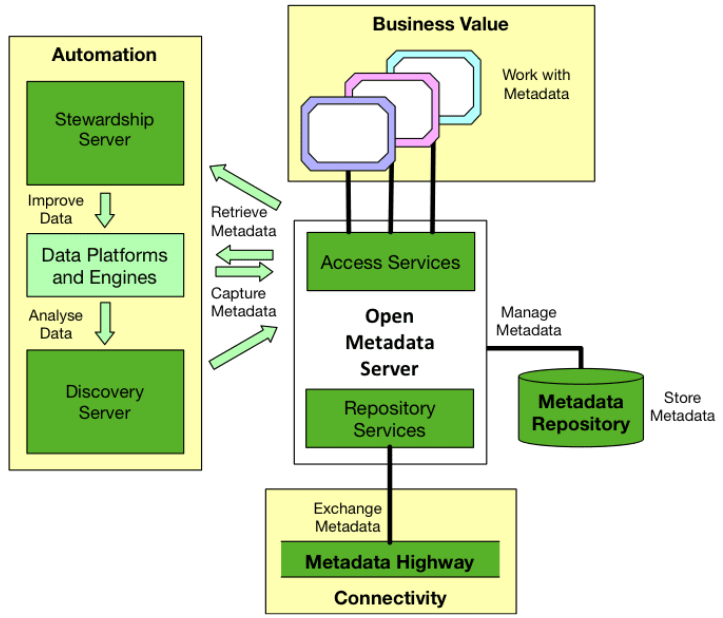



https://dagster.io/blog/rebundling-the-data-platform

ODPi EGERIA




Apache Atlas

Collibra

Open Metadata

OpenLineage

AN OPEN FRAMEWORK FOR DATA LINEAGE COLLECTION AND ANALYSIS

Data lineage is the foundation for a new generation of powerful, context-aware data tools and best practices. OpenLineage enables consistent collection of lineage metadata, creating a deeper understanding of how data is produced and used.

DataHub

GOVERNANCE

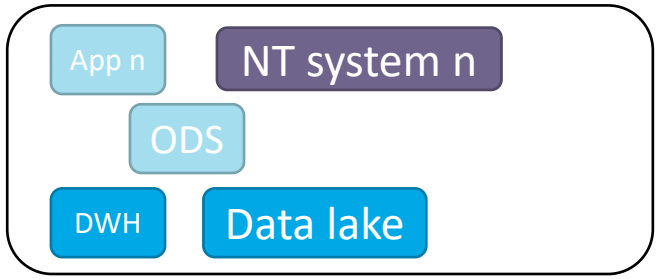
- Simple: write documentation into a file Intake
- Medium: BI docs as code (evidence.dev)
- Comlex (distributed, enterprise) EGERIA
- **MDS? → Open Metadata, DataHub**

# FUTURE PLATFORM

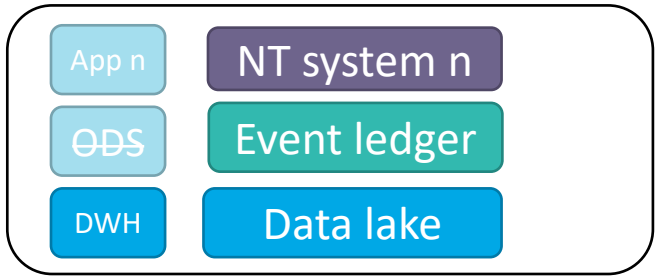
>> work in progress >>

# Future data platform

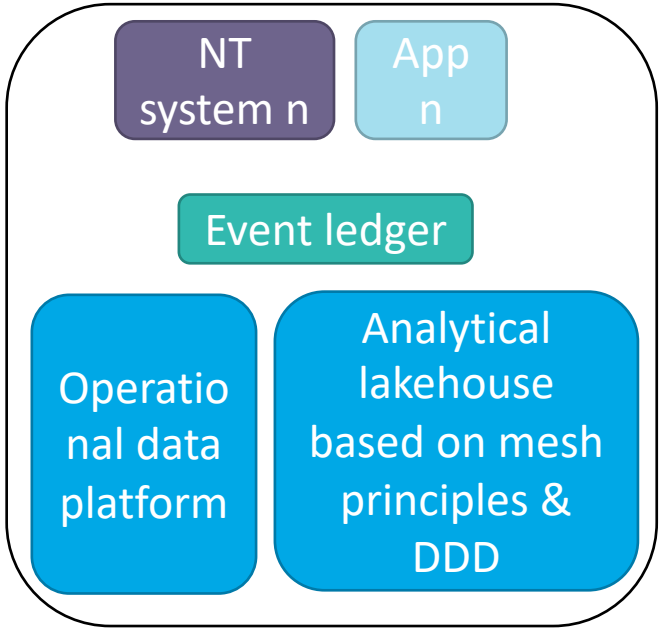
old



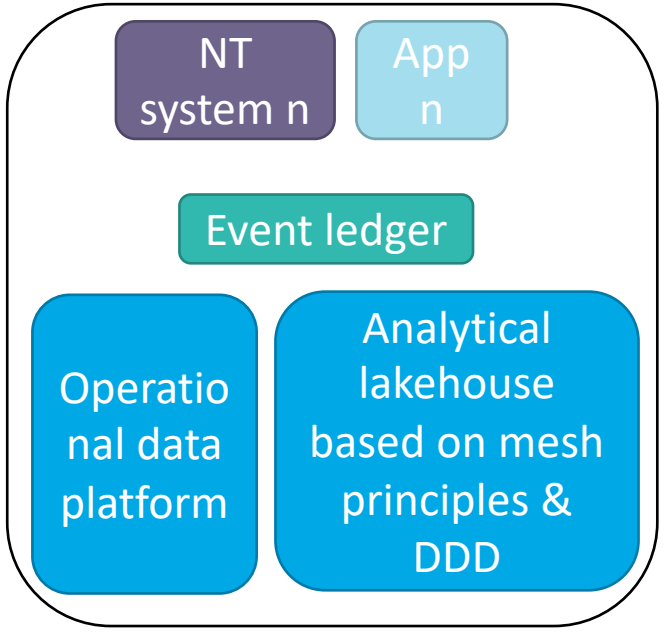

current




On-prem cloud




(public) cloud [n]



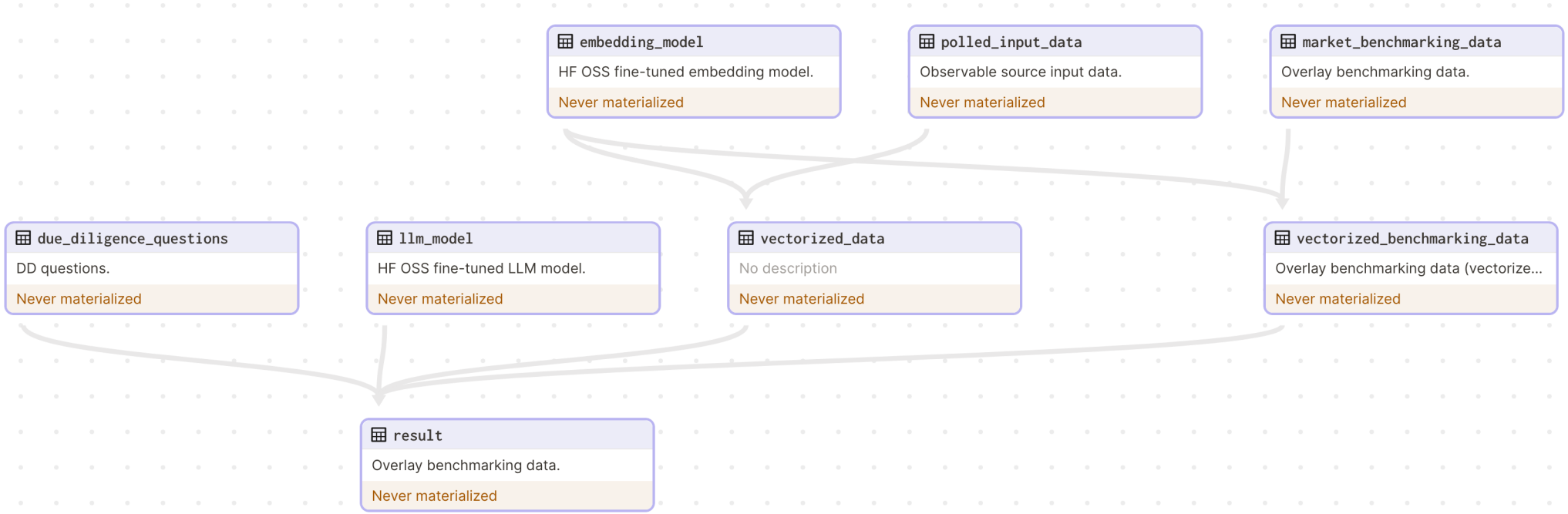

Shared:

- Data governance
- Orchestrator
- IaC
- Domain separation (mesh)
- Unify data access

Various SaaS apps

# DEMO EXAMPLES

Dagster, DBT,  + various storage/compute engines (duckdb and more)




# PROSPECUTS.CHAT

One concrete case: LLM AI app

More ML examples: https://dagster.io/blog/finetuning-llms

# Findings

## 01

DDD

- Clear data contracts
- Domain based data ownership
- E2E ownership Reorganization (data product)

## 02

Foundational pillars

- **Data governance**
- **Central data orchestration tool** with support for lazy coupling of domains
- Revival of SQL

## 03

Enablement

- Hybrid cloud setup (new types of hardware, faster provisioning)
- Use established SaaS tools (less build your own)
- Invest in tooling data engineers love for easier hiring

Foundational principles to help Magenta become more data driven.

# Dagster vs. Airflow resources

Responsible for delivering **advanced analytics & machine learning** solutions.

Data Science team currently consists of 6 people, Data & Insights in total is ~35 internals + ~15 externals.

Central team providing DS solutions to stakeholders across the company.

**Topics include:**

- Airflow's operational complexity: Blogpost outlining many of the reasons Airflow is hard to run at scale in production due to architecture limitations. https://dagster.io/blog/dagster-airflow
- Airflow's developer pain (esp with local testing): Group 1001 testimonial https://www.youtube.com/watch?v=Kf0iURfebdA
- Benefits of declarative scheduling for large, multi-team projects: Whatnot testimonial https://www.youtube.com/watch?v=ZZZO33MEvF4
- Overview D-A comparison https://docs.google.com/spreadsheets/d/1TM-DrqCnMv6QK3jw710GhPU0UP9NQhws/edit#gid=492840621 (request access required)

# Dagster resources

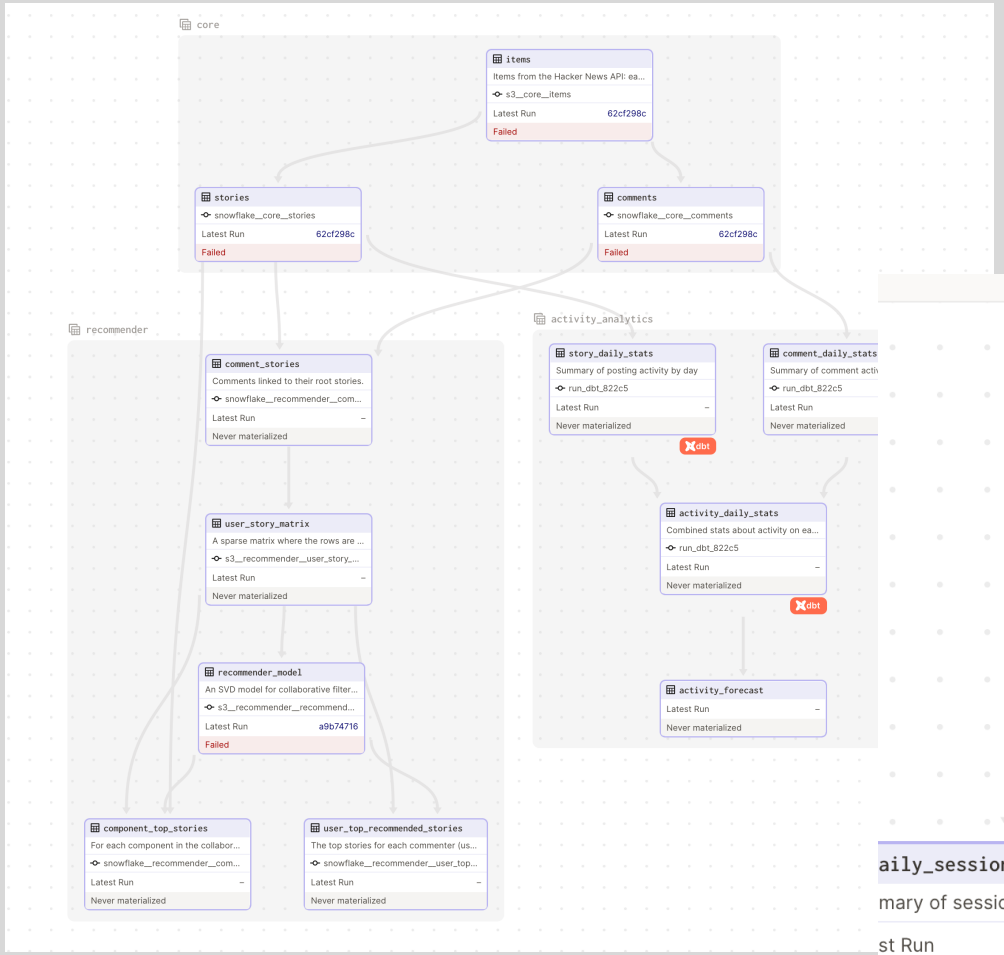
- https://github.com/dagster-io/hooli-data-eng-pipelines/tree/master
- https://github.com/slopp/dagteam
- https://stkbailey.substack.com/p/orchestration-isnt-going-anywhere

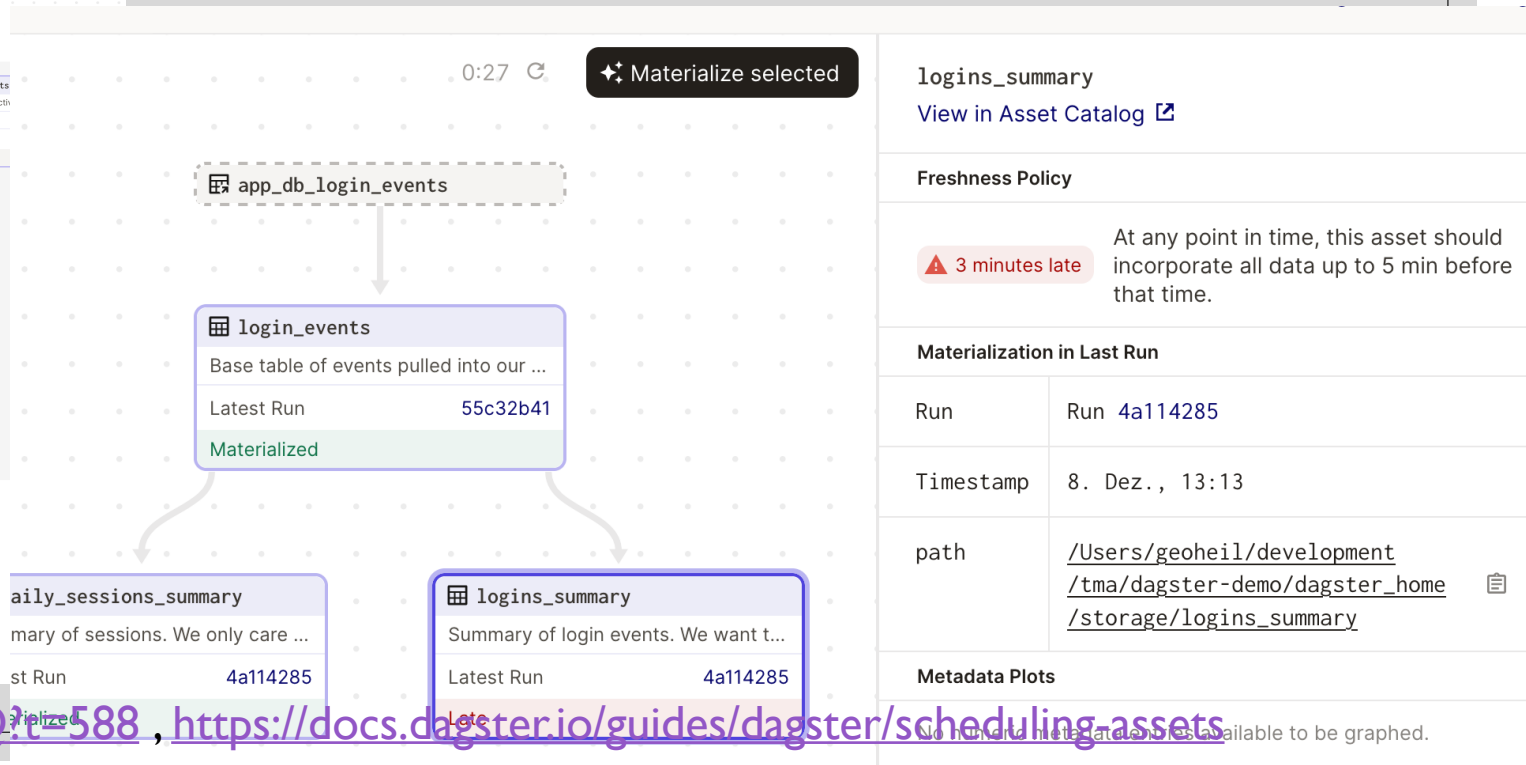
# Hiring Data Engineers

- You measure your data engineer’s customer impact
- Data engineers grow with the company
- Your data ecosystem is easy to navigate for data engineers
- You invest in the tooling data engineers love
- Your data engineers have autonomy

https://www.montecarlodata.com/blog-5-on-obvious-ways-to-make-data-engineers-love-working-for-you

# Demo declarative scheduling

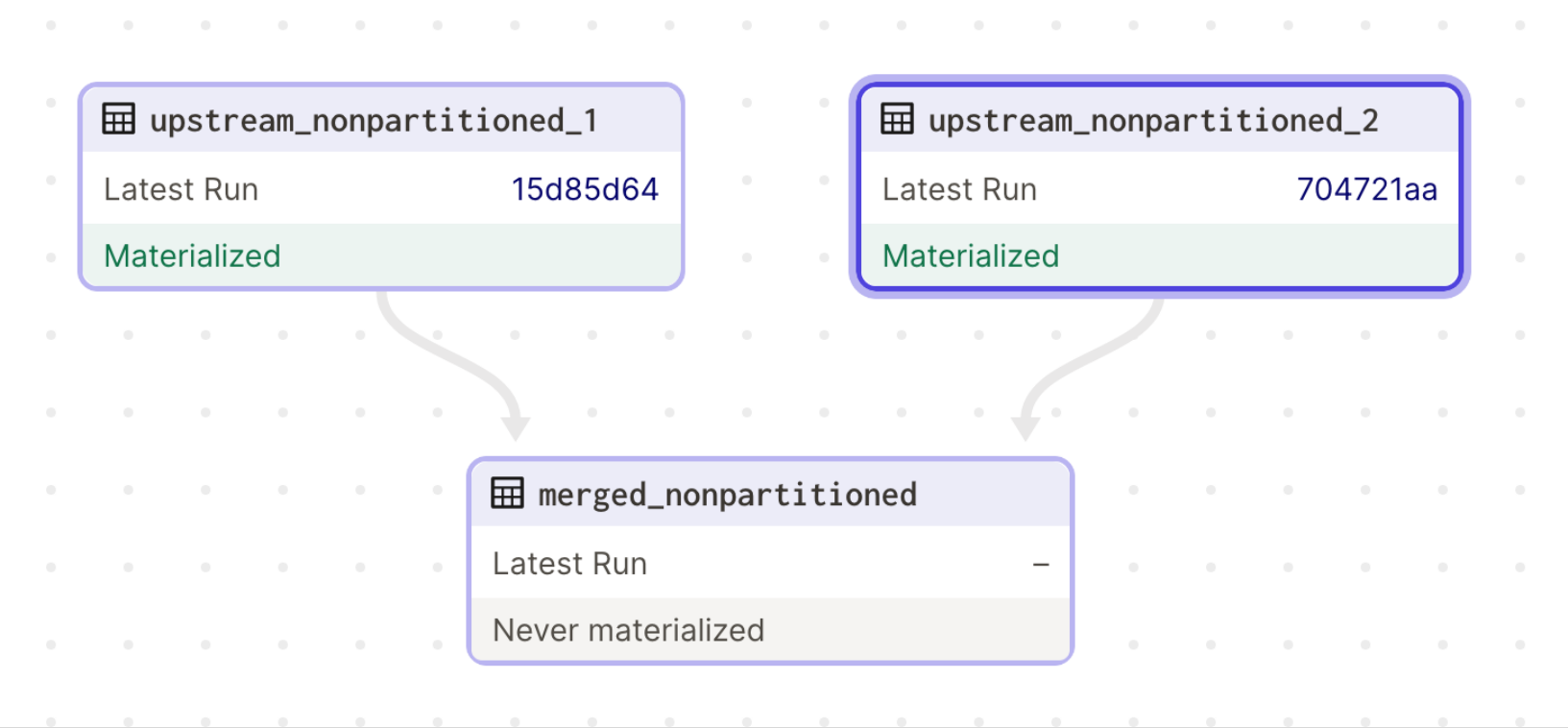



SDA + freshness: https://youtu.be/An78xLxM9zQ?t=588 , https://docs.dagster.io/guides/dagster/scheduling-assets

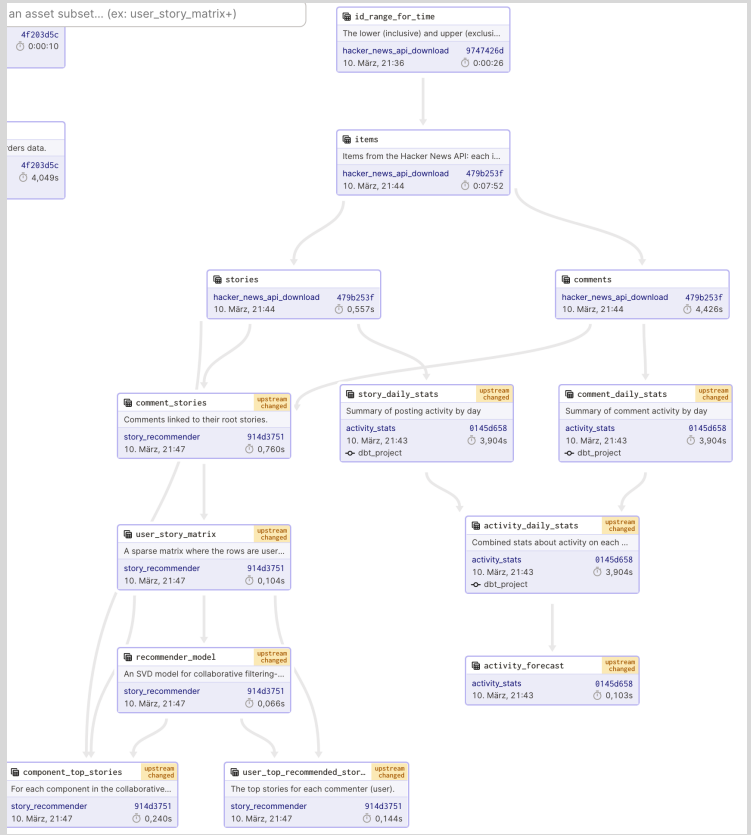
https://dagster.io/blog/declarative-scheduling

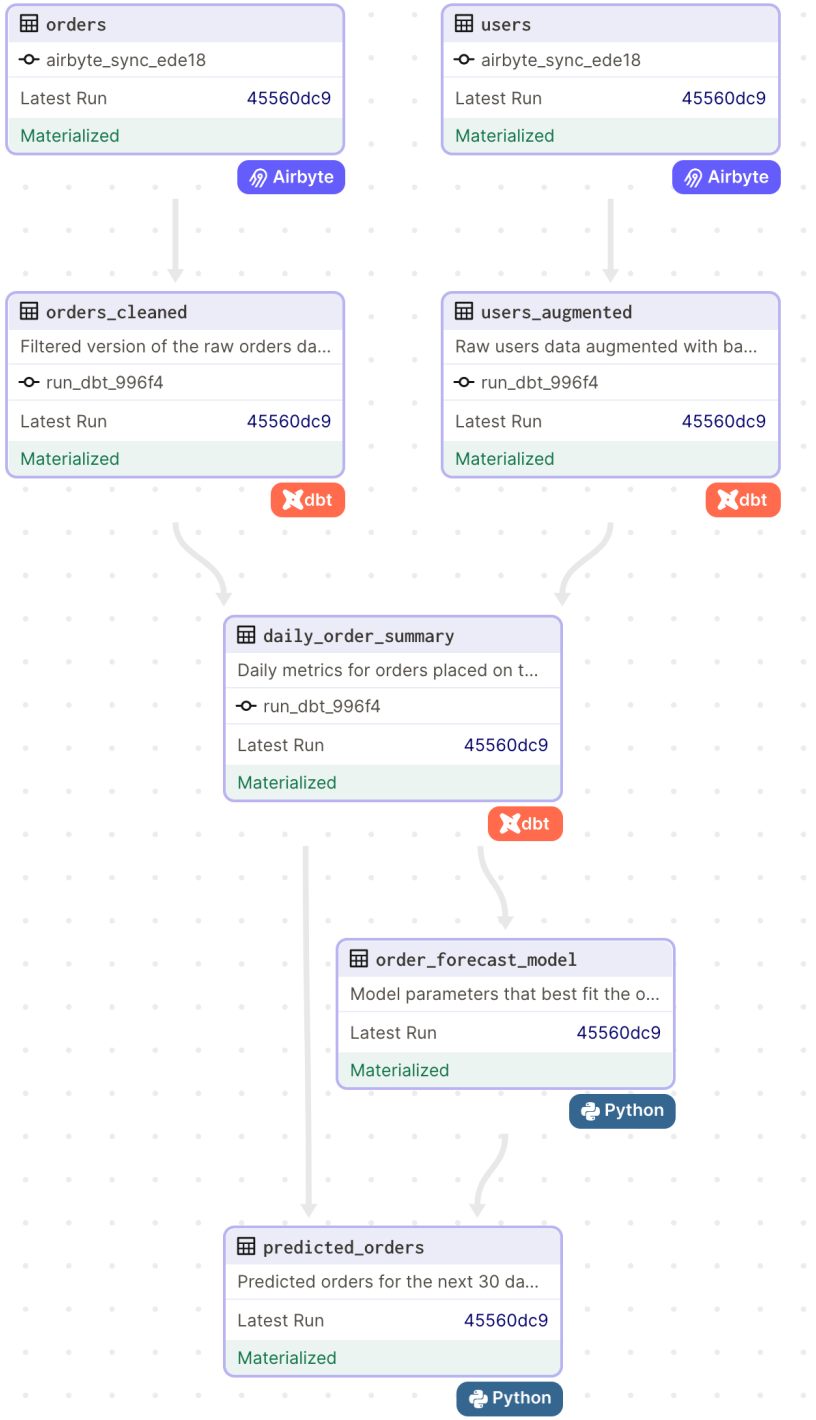
# Demo blocking/reconciliation




https://docs.dagster.io/concepts/partitions-schedules-sensors/asset-sensors#asset-reconciliation-sensor

# Demo E2E governance

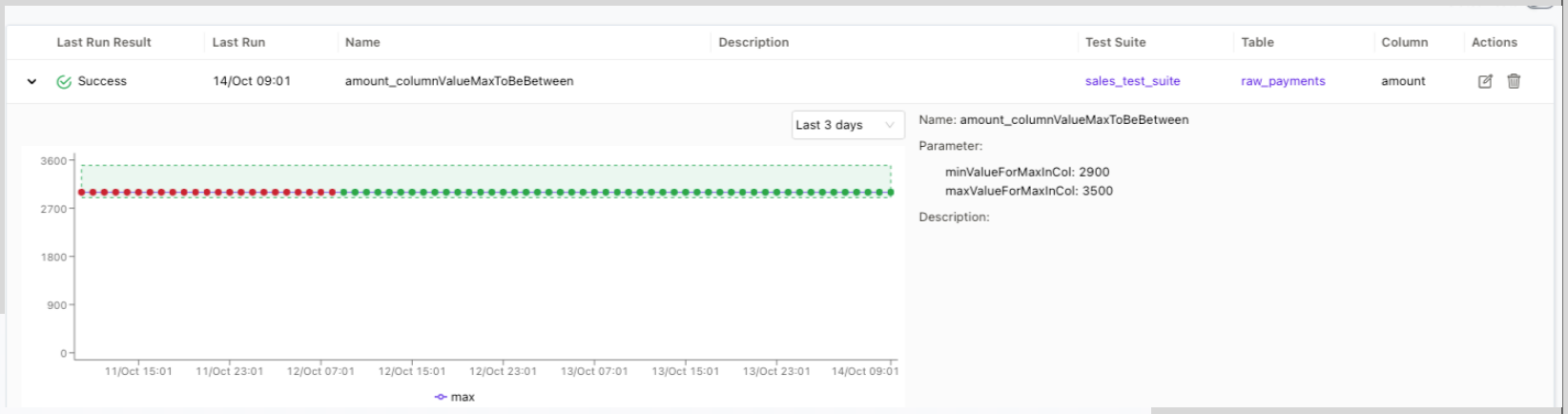



georgheiler.com/2022/03/04/connector-goodness-from-airbyte-e2e-lineage

georgheiler.com/2022/03/04/fully-fledged-example-with-resources

# Temporal metrics & tests




sample_data > ecommerce_db > shopify > dim_address > address_id

No Owner | No Tier | Type: Regular | Usage - 0th pctile | 7 Queries

+ Add tag

Summary | Profiler | Data Quality

Last 3 days | Add Test

**Column summary**

Unique identifier for the address.

Data type

NUMERIC

**Table Metrics Summary**

| Row Count | Column Count | Table Sample % |
|---|---|---|
| 14,567 | 12 | 100% |

**Table Tests Summary**

| Success | Aborted | Failed |
|---|---|---|
| 02 | 00 | 00 |

Distinct Count

14,500

Null Count

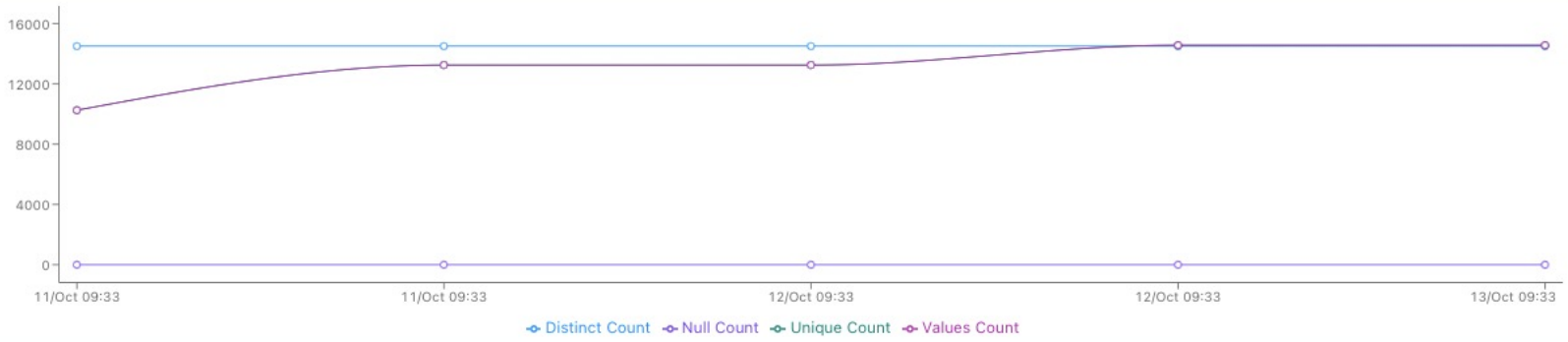
0

Unique Count

14,600

Values Count

14,600



# DBT governance

- https://github.com/dbt-labs/dbt-project-evaluator
- https://github.com/offbi/pre-commit-dbt including workflow example of https://montrealanalytics.notion.site/Coalesce-Workshop-Guide-6382db82046f41599e9ec39afb035bdb and video of the workshop https://www.youtube.com/watch?v=L6ixHejZX5A&list=PL0QYlrC86xQlj9UDGiEwhXQuSjuSyPJHl&index=42
- https://github.com/tobymao/sqlglot
- https://www.sqlfluff.com/
- https://github.com/Montreal-Analytics/dbt-gloss
- Teamstruktur: dbt-labs/dbt-core#5244 community endorsement possible
- A/B (blue/green) deployments for data https://blog.montrealanalytics.com/blue-green-deployment-with-dbt-and-snowflake-922f1c658011
- documentation https://blog.montrealanalytics.com/building-sustainable-dbt-project-documentation-8def88ca67c3
- Enterprise transformation https://www.getdbt.com/analytics-engineering/case-for-elt-workflow/
- Single Source of Truth for metrics https://www.databricks.com/de/dataaisummit/session/democratizing-metrics-airbnb
- DBT + Dagster Demo https://www.youtube.com/watch?v=JHCDgRbWWq0