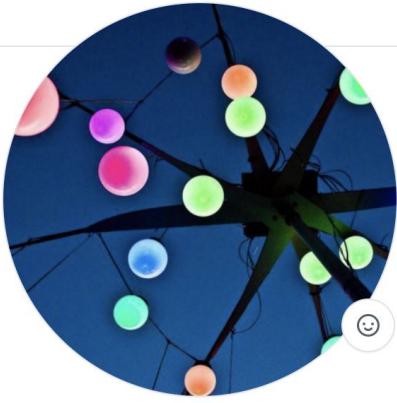




# Navigating the Landscape of a **Fully Open Source Data Stack**

**FY2024 Q4**

11/29/23



Maxime Beauchemin

mistercrunch

creator of Apache Airflow and Apache Superset - founder at Preset

Edit profile

1k followers · 11 following · 139

preset-io

San Mateo, CA

maximebeauchemin@gmail.com

mistercrunch.blogspot.com

#### Organizations

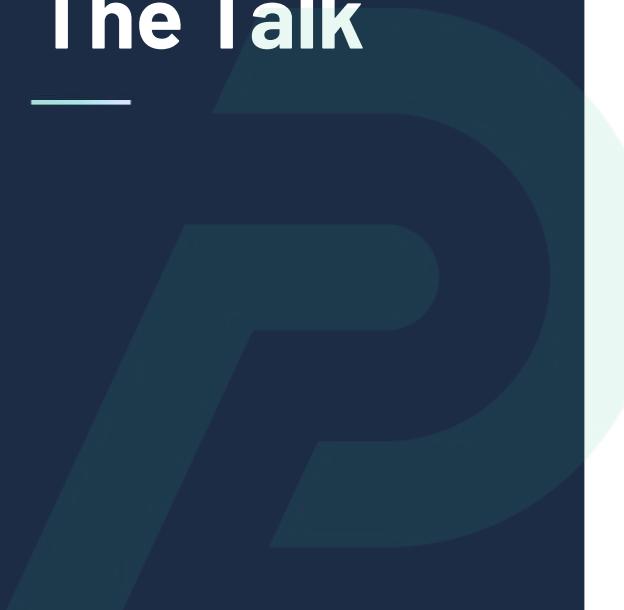


- Passionate about building data tools!
- Started Apache **Airflow** at Airbnb in 2014
- Started Apache **Superset** at Airbnb in 2015
- Started **Preset** - The Apache Superset company in 2019

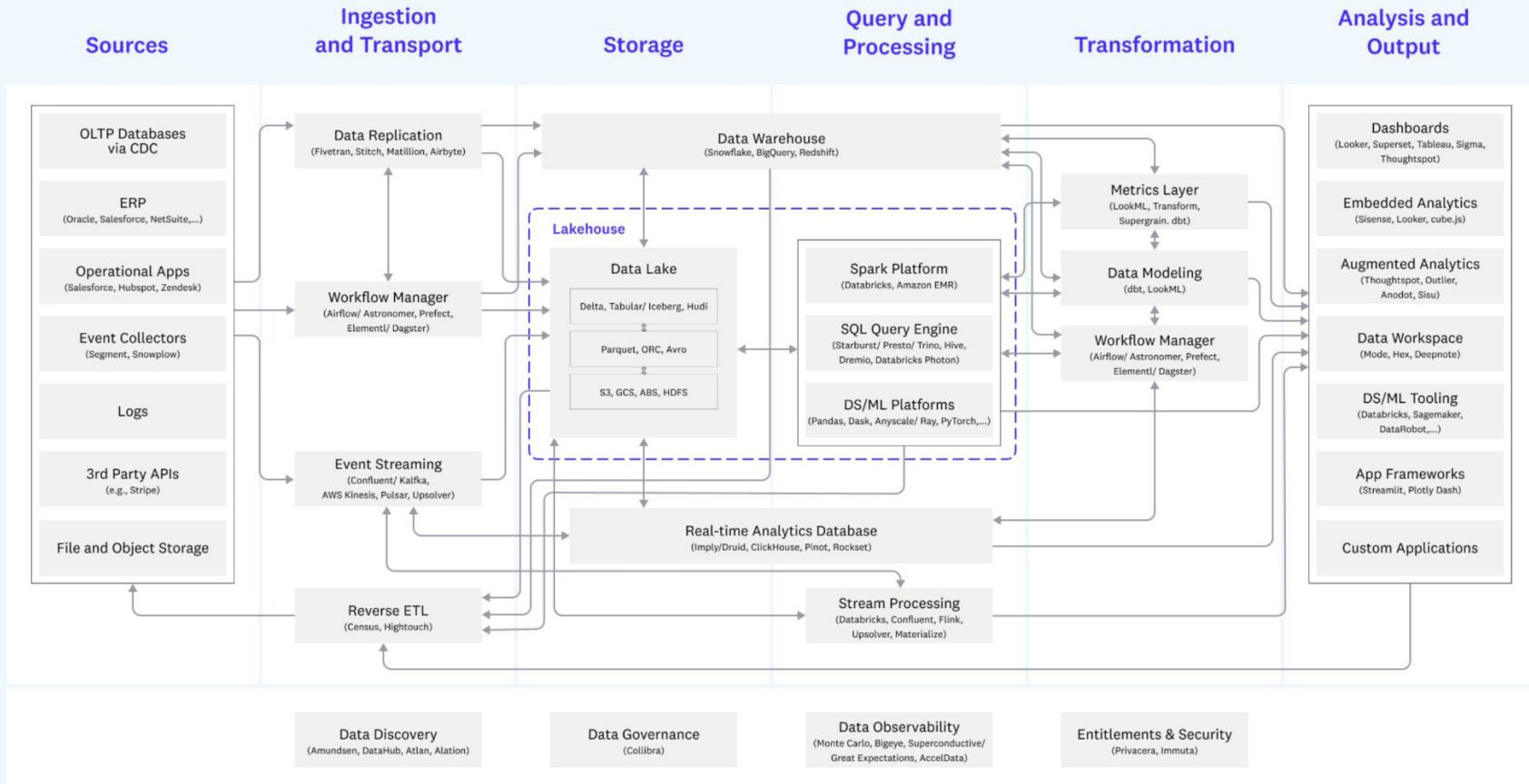


# The Talk

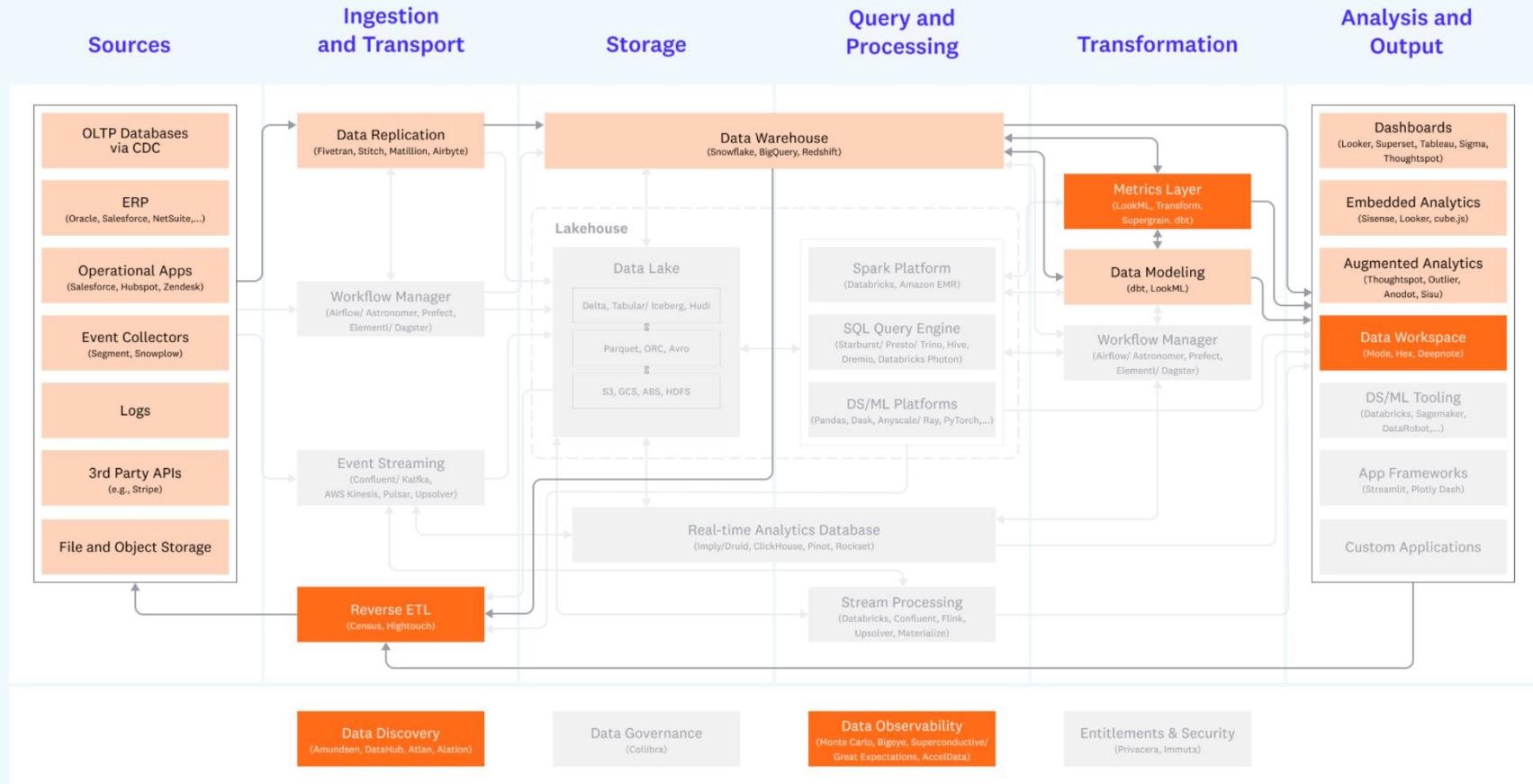
---

- 
1. Decompose the “modern data stack” into its key components
  2. Overview the OSS solutions and how they compare to one-another and top vendor solutions
  3. **[disclaimer]** this talk is incomplete, biased and dated!

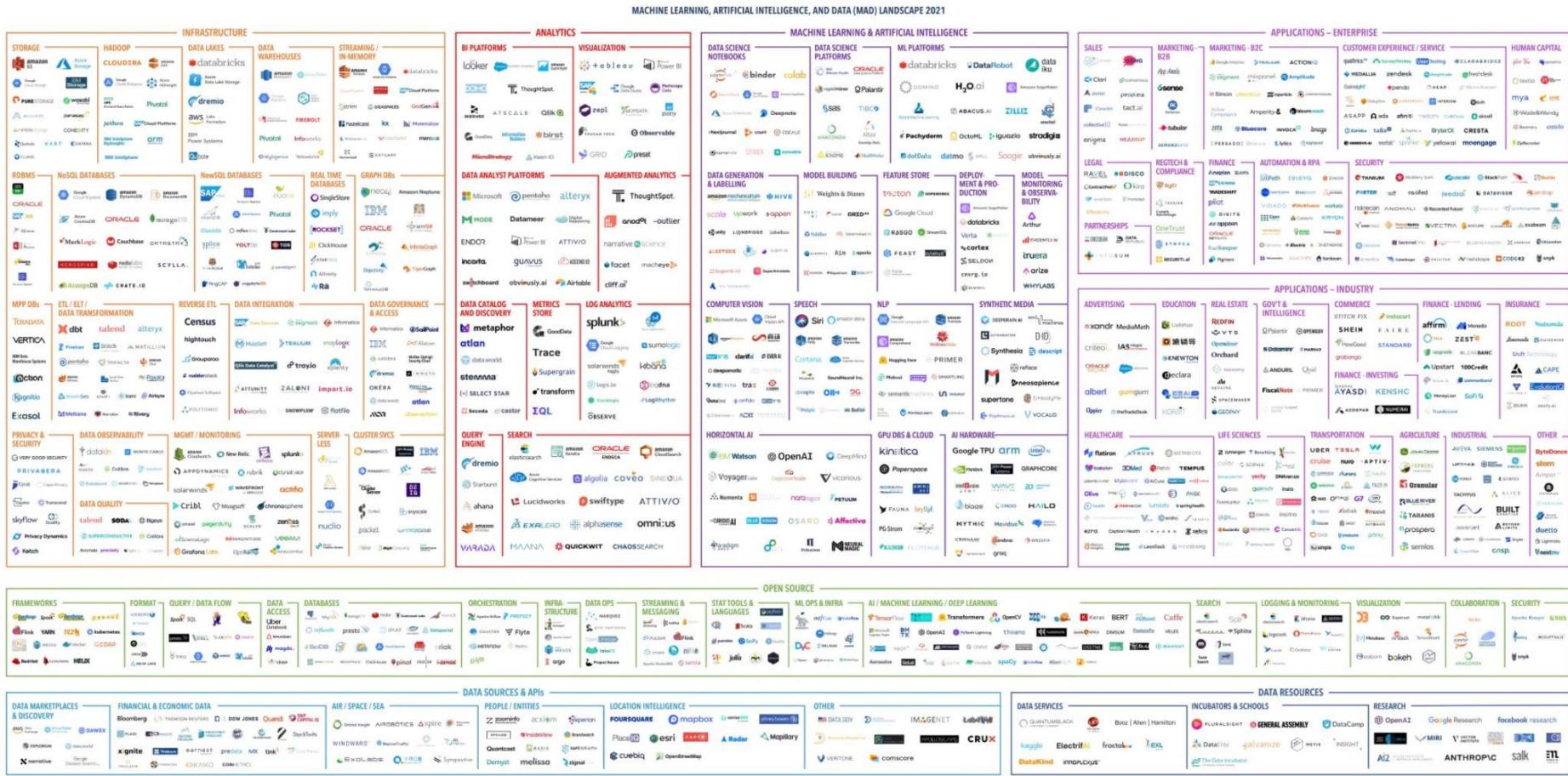
# Unified Data Infrastructure (2.0)

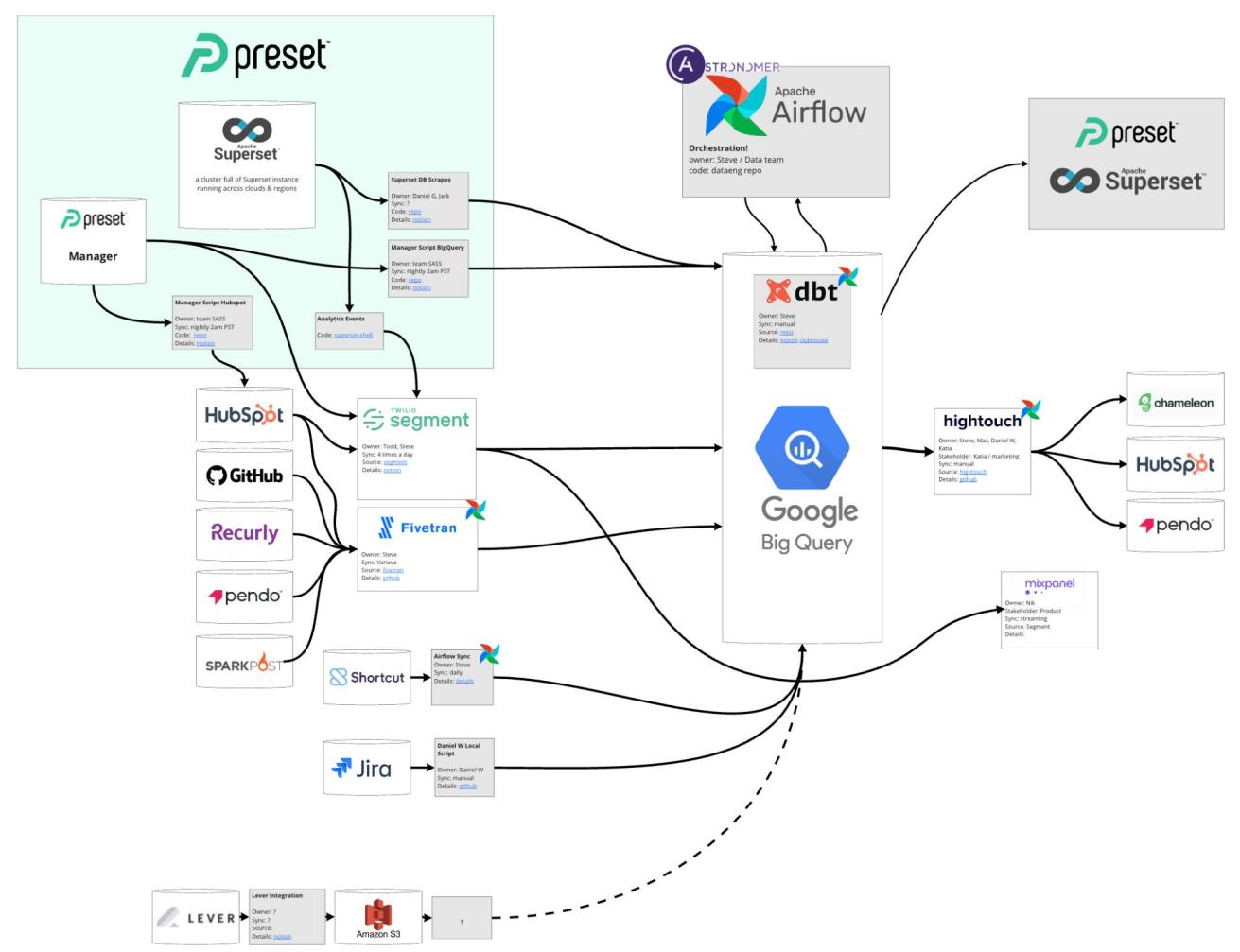


# Blueprint 1: Modern Business Intelligence



So. In this talk. We're just going to make sense of this ->





# Scoping by layer

## Vital components

- Data warehouse / data lake
- Data replication
- Data transformation & orchestration (ETL & ELT)
- Dashboards

## Secondary concerns

- Data catalogs
- Data workspace / notebooks

## Saving it for another talk

- Data observability
- Embedded analytics
- Steaming stuff.
- ...

# warehouse / lake / lakehouse / database

## Project

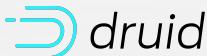
Lakehouses



presto



OLAP & RT



Why not?



PostgreSQL



## COSS

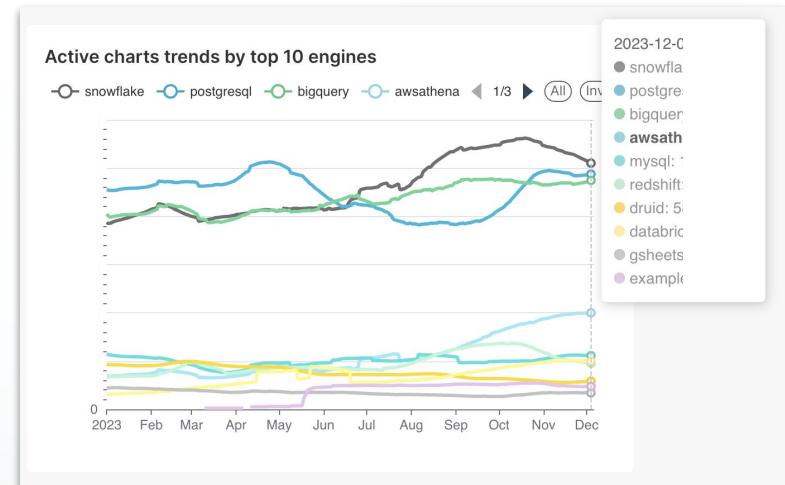


## Commercial



# warehouse / lake / lakehouse / database

- BigQuery and Snowflake, with the convenience of cloud/SaaS taking the data world by storm
- A surprising amount of Postgres! At least from the Preset vantage point
- COSS vendors offering similar convenience with their hosted, multi-tenant cloud offerings
- Trino/Presto firing on all cylinders
- Special purpose databases seem to be becoming more general purpose over time
- OLAP tier is amazing, tech is evolving quickly and is highly competitive on the COSS front



# Data Replication - the EL in ELT

Project

COSS

SINGER



Airbyte



Stitch



rudderstack™

DiY !?



Commercial



Fivetran



segment

# Data Replication

---

- Don't ducking DIY this!
- Feels like an area where OSS should dominate
- Fivetran taking the market by storm
- Competitive on the OSS front
- Give Airbyte a shot before committing to Fivetran

# Data Catalogs

## Project



Apache **Atlas**

## COSS



Acryl Data

## Commercial



**atlan**



**Collibra**

# About Data Catalogs

---

- Nice to have - exponentially important as data teams grow
- DataHub seem extremely dynamic and competitive with Accryl offering a hosted solution
- Highly competitive, fighting for a relatively small TAM

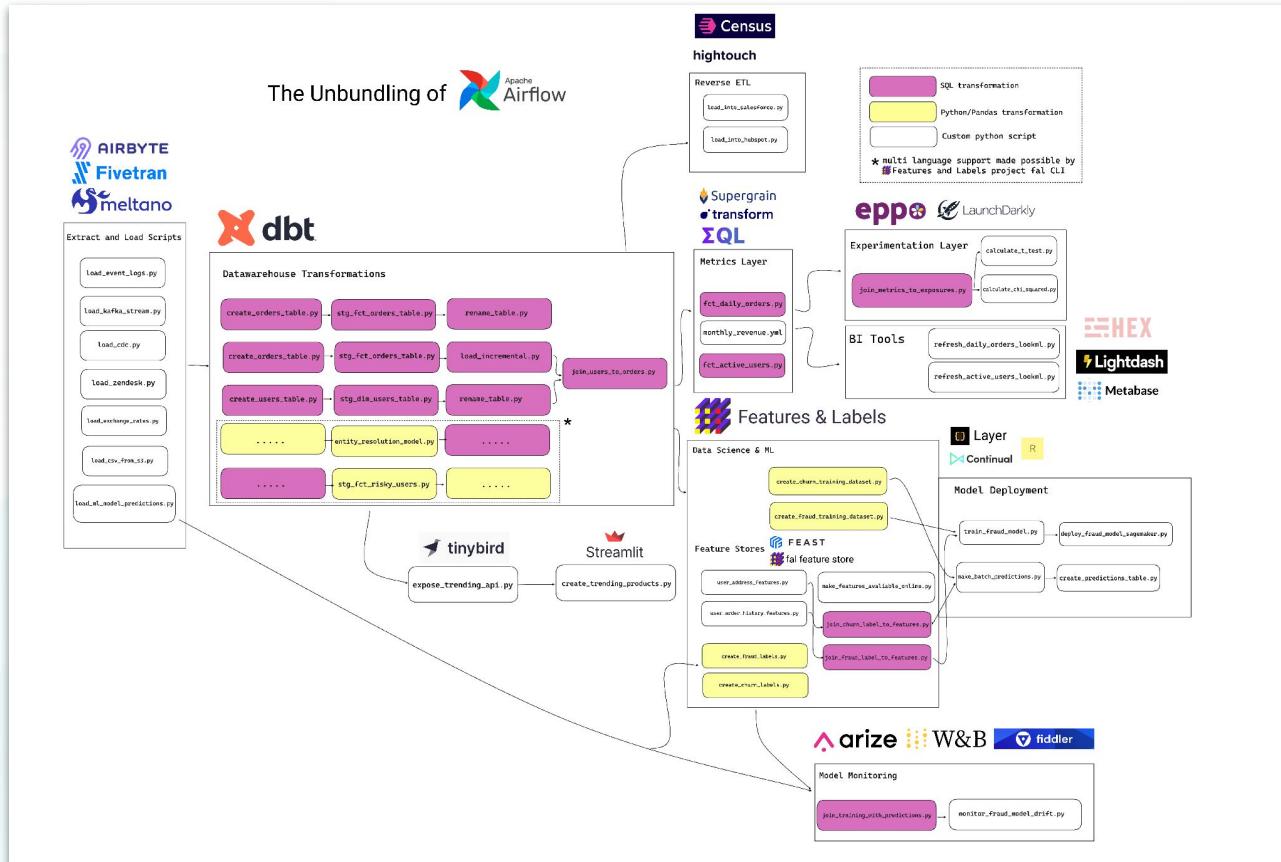
# Data Transformation & Orchestration



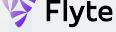
# SQL

**"Mountains of templated SQL and YAML -** The industry is doubling down on templated SQL and YAML as a way to manage the "T" in ELT. To be clear, this direction was already materializing when I created Airflow in 2014, but many in the industry thought we were going to move towards dataframe-type APIs (like Spark dataframe API, or the Apache Beam spec). From my visibility on the Airflow ecosystem, templated SQL is a large portion of what Airflow DAGs are operating on, and the prevalence of DBT is also a clear signal that a large portion of the ELT logic is implemented in that way."

# Data Transformation / Orchestration



# Data Transformation / Orchestration - batch

Project	COSS	Commercial
SQL-first		
 SQL Mesh		
Orchestration-first		
 Apache Airflow		
dataflow + workflow		Informatica
 Flyte		
dataflow-first		
 Apache Spark™		

# BI / Dashboarding

# Visualization, exploration and dashboarding

Project

 Apache Superset™

 Metabase

 pentaho

 JASPERSOFT™

COSS

 preset

**BIAS ALERT!!!**

Commercial

 + tableau®

 Power BI 

 Looker

+ hundreds! maybe thousands!?

# Visualization, exploration and dashboarding

BIAS ALERT!!!

- **Apache Superset** is first in adoption and community size
- Superset has wide diversity of contribution - the ASF governance model
- Metabase has a troubling copy-left license (AGPL), unclear governance model
- Other tools popping up - see other osacon talks for more detail

LICENSE-AGPL.txt

Move Metabase® Enterprise Edition™ code to core repo (#13590)

# Data workspaces / notebooks

## Project



## COSS



## Commercial



databricks



## Observable



Azure Notebooks

# About Notebooks

---

- On the commercial side, notebooks are often a flexible frontend in front of hosted infrastructure solutions
- On the open source front, a lot can be achieved with Jupyter, NBViewer and static website generators

Wrappin' up...

# Some parting words...

---

- Stacks are complicated!
- Talk to other practitioners about their stack, see what they love nad h
- Migrations are hard - select wisely!
- The sole act of selecting a stack is hard
- Managing a bunch of open source project is hard - mânâge services have their advantages
- Gluing things is a fair amount of work too
- OSS FTW!

*That's all Folks!*