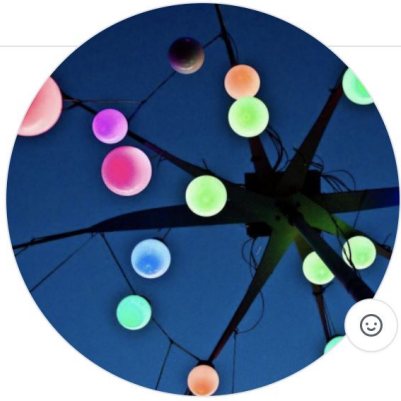




Navigating the Landscape of a **Fully Open Source Data Stack**

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



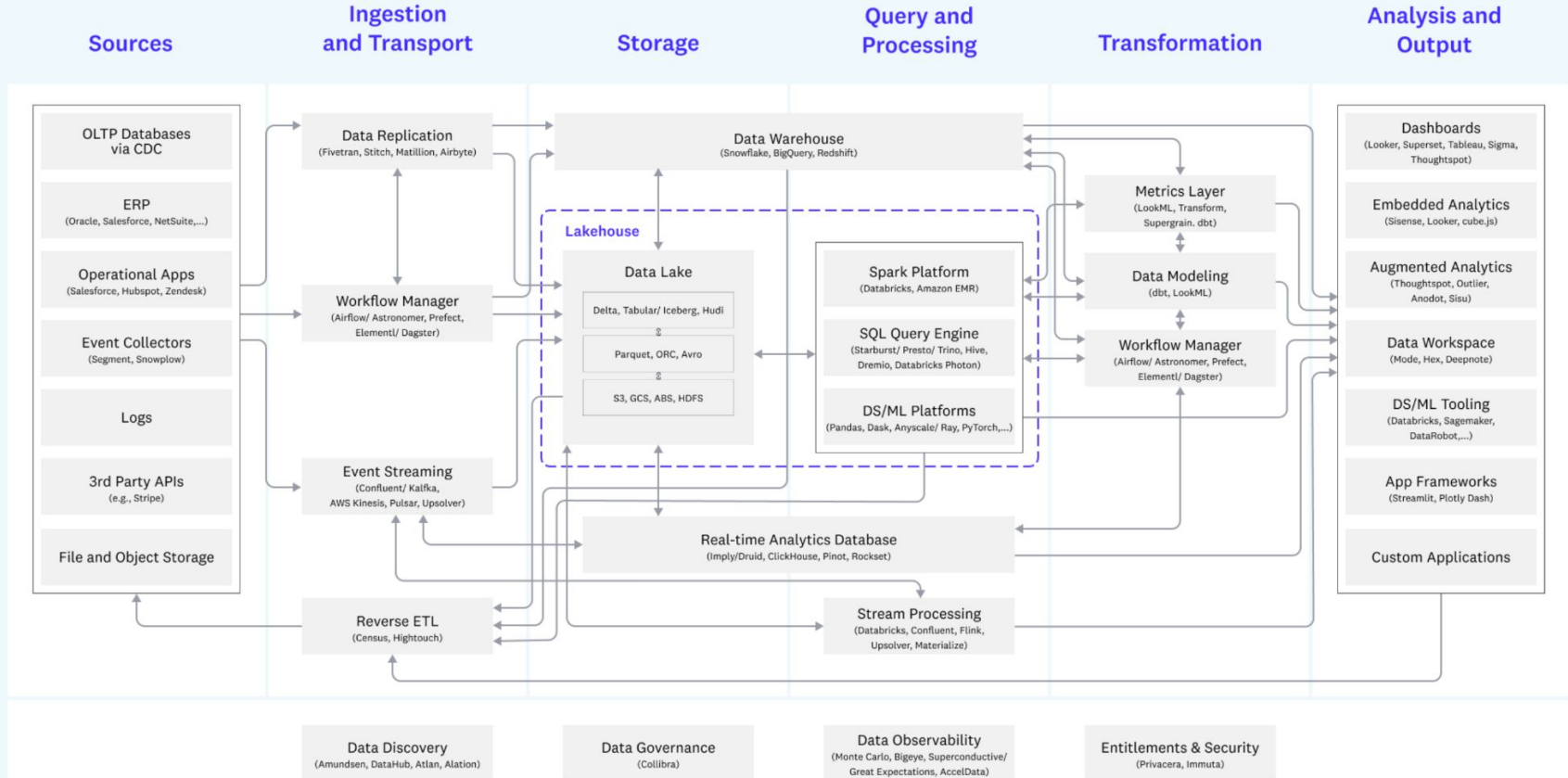
- Passionate about building data tools!
- Started Apache **Airflow** at Airbnb in 2014
- Started Apache **Superset** at Airbnb in 2015
- Started **Preset** - The Apache Superset company in 2019



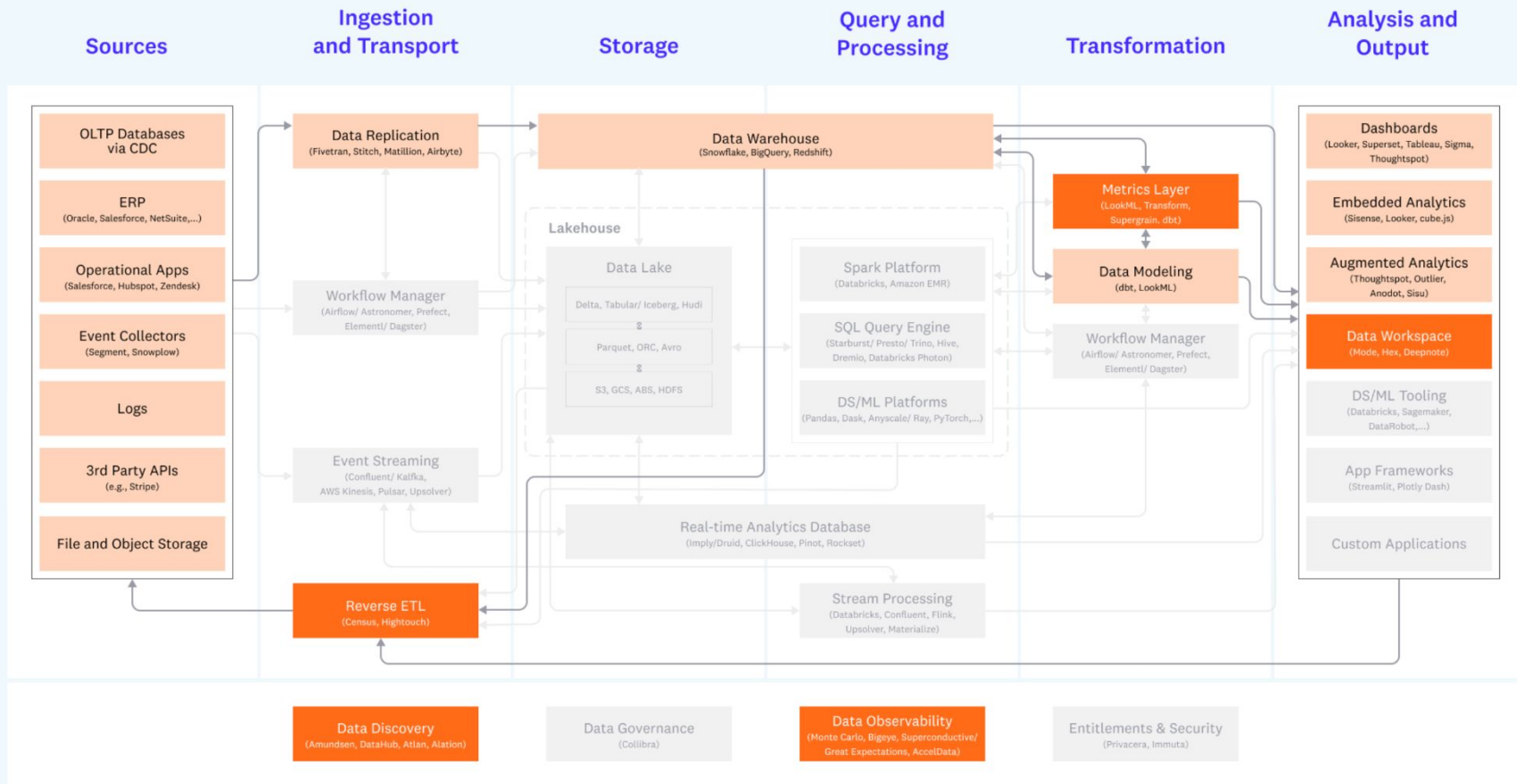
The Talk

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)

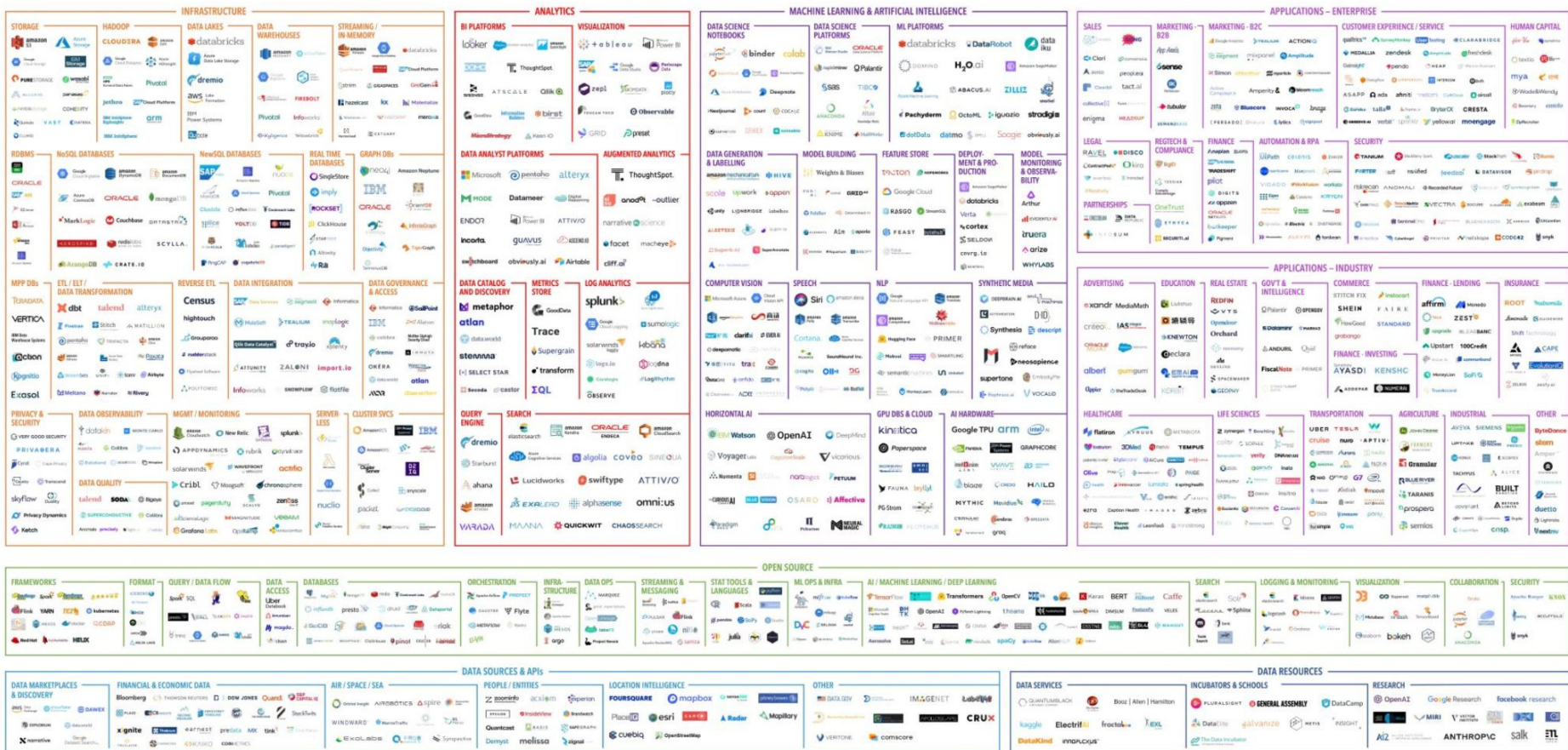


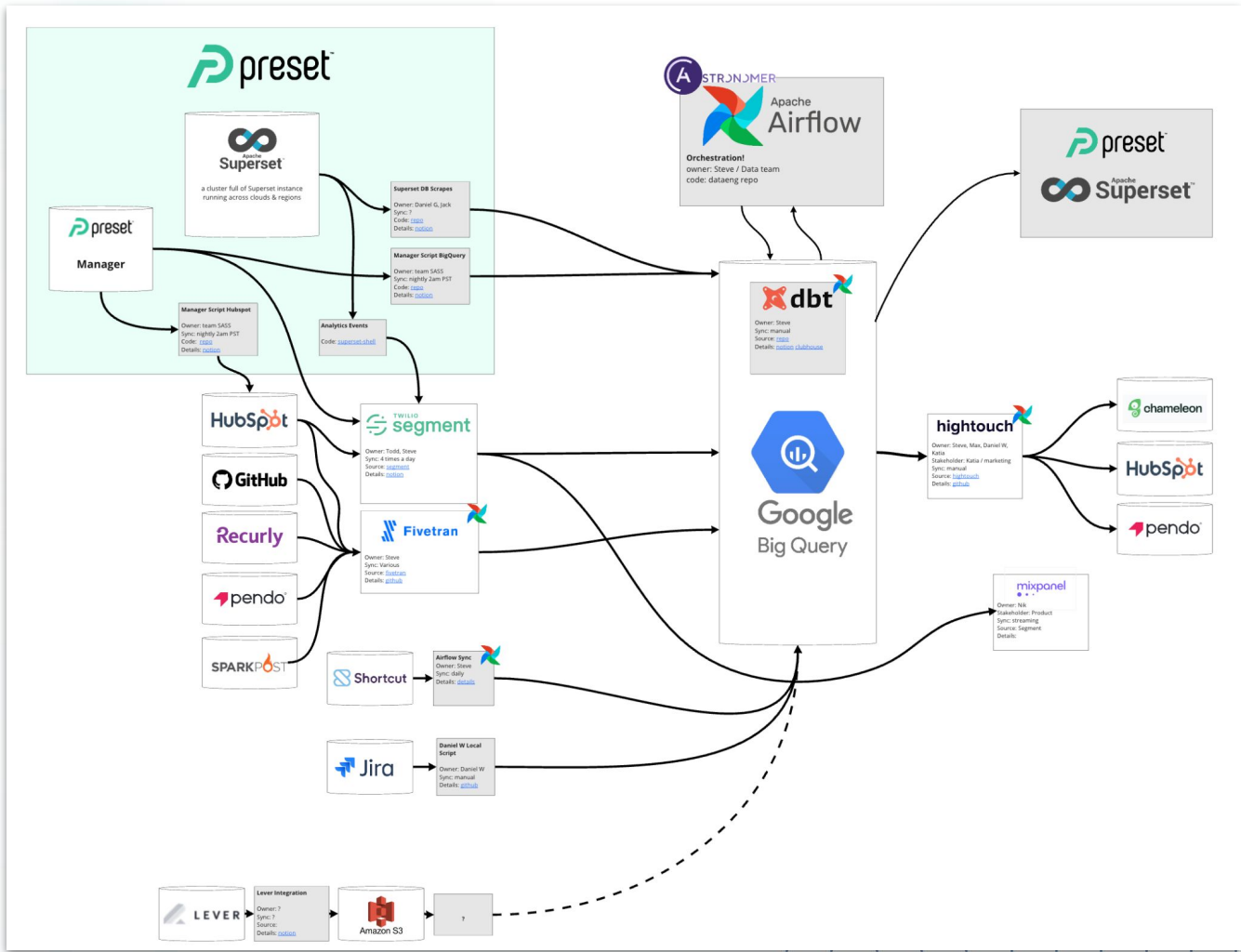
Blueprint 1: Modern Business Intelligence



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

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021





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



OLAP & RT



Why not?



COSS



Commercial



Data Replication - the EL in ELT

Project

COSS




Airbyte



SINGER



meltano



Stitch



rudderstack™

DiY !?




Apache Airflow




dagster




Commercial



Fivetran



MATILLION



TWILIO 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 - expediently 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



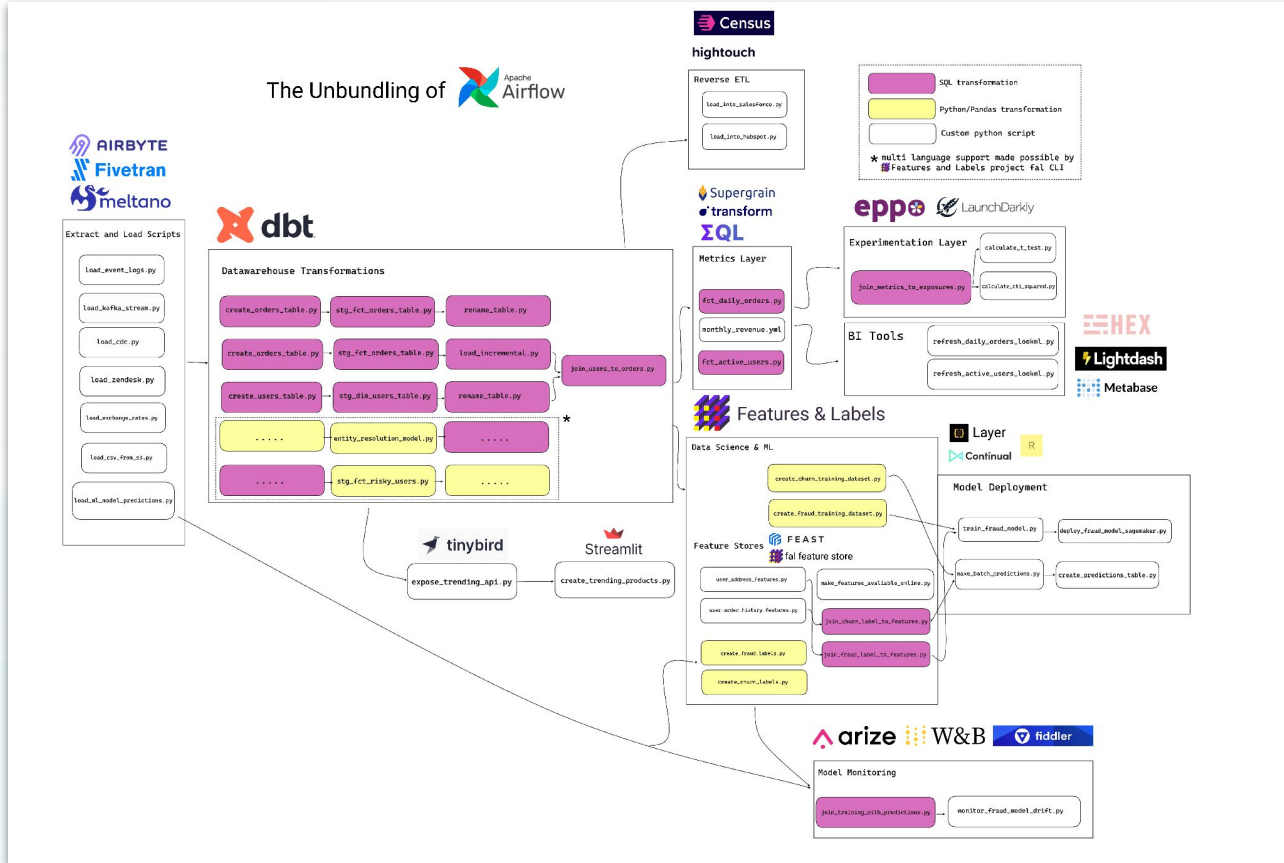
Data Transformation & Orchestration



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.”

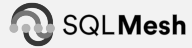
Data Transformation / Orchestration



Data Transformation / Orchestration - batch

Project

SQL-first



Orchestration-first



ASTRONOMER

dataflow + workflow



dataflow-first



COSS



Commercial



Informatica™

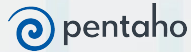
IBM DataStage®

BI / Dashboarding

Visualization, exploration and dashboarding

BIAS ALERT!!!

Project



COSS



Commercial



Power BI



+ 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

Data workspaces / notebooks

Project



COSS



Commercial



databricks



Amazon SageMaker



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 and hate
- Migrations are hard - select wisely!
- The sole act of selecting a stack is hard
- Managing a bunch of open source projects is hard - managed services have their advantages
- Gluing things is a fair amount of work too
- OSS FTW!

The image features a central graphic consisting of several concentric circles. The innermost circle is a solid dark blue. Surrounding it are several rings of varying shades of red, from a deep, dark red to a lighter, more vibrant red. The outermost ring is a solid black. Overlaid on this circular pattern is the text "That's all Folks!" written in a white, elegant cursive script. The text is positioned diagonally across the center of the graphic, starting from the lower-left and ending at the upper-right. The overall composition is balanced and visually striking due to the high contrast between the white text and the dark background elements.

That's all Folks!