

Introduction to Data Engineering on Databricks

Adastra Thailand Campus on-tour program

Stamford International University
28 May 2024



Meet our team



Wiparat P.

Head of Operations



Sirakorn L.

Practice Lead – AWS, Data Engineering,
and Development



Atiwat P.

Data Engineer



Thanyaporn S.

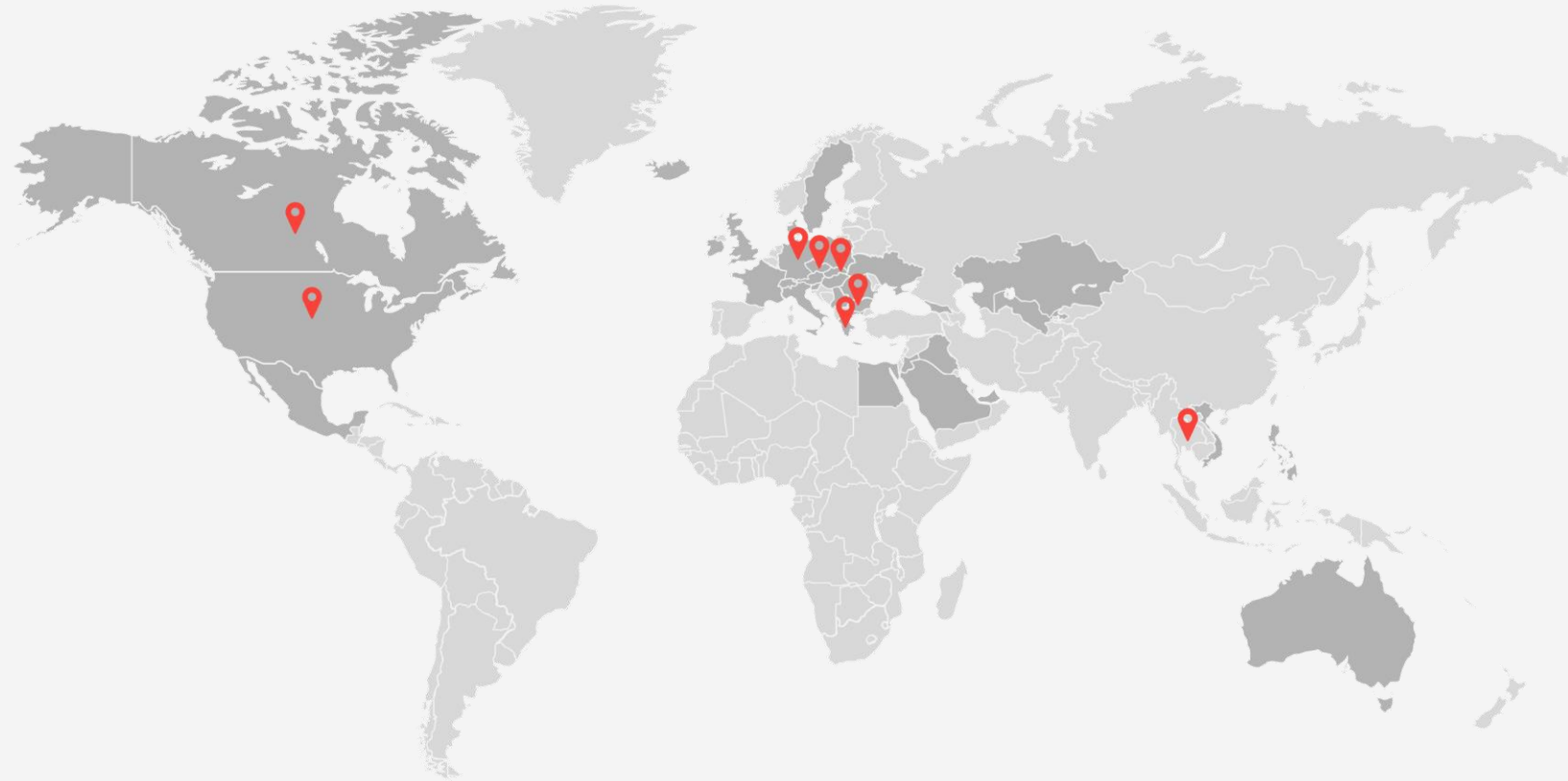
Recruitment Manager




Manassaphorn W.

HR Manager

Adastra's Global Presence



 Adastra Offices

 **CANADA**

Toronto
Calgary
Ottawa
Vancouver
Montreal

 **UNITED STATES**

Los Angeles
Miami
New York

 **AUSTRIA/
SLOVAKIA**

Bratislava

 **CZECH REPUBLIC**

Prague

 **GERMANY**

Frankfurt
Wolfsburg
Hanover
Munich
Magdeburg
Darmstadt

 **BULGARIA**


Sofia
Varna
Plovdiv

 **GREECE**

Thessaloniki

 **THAILAND**

Bangkok
Chiang Mai

 Countries where we have delivered projects



8

Countries



22

Offices



500+

Customers



2,200+

Professionals



40+

Countries where we
have delivered projects



20+

Languages
supported

Realize Your Data-Driven Destiny

For 20+ years, customers have trusted Adastra to design and deliver comprehensive data-driven solutions that fuel efficiency, innovation and long-term success. Our diverse set of Superpowers transform the way organizations utilize their data, unlocking its full potential.





Our Partners





Before we start

- Sign up for Databricks Community Edition at community.cloud.databricks.com
- Grab the copy of this slide with this short URL: <https://bit.ly/ath-2024-stiu>
- Or with the following QR code:



./ADASTRA

Data and Data Engineering



Big Data: how can it be massive?



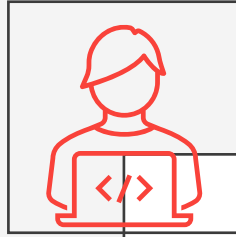
- Cheaper device makes it possible to generate massive data.
- Cheaper storage unit makes it possible to store data first without thinking whether to use it or not.
- Internet makes it capable for users to distribute massive amounts of data.
- How can we process them?
- What are the aspects of processing them?
 - Make predictions and forecasts
 - Deliver insights in understandable format
 - Productionize the process

Data Careers



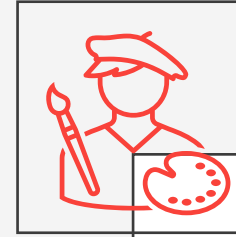
Data Scientists

- Use statistics, machine learning, mathematics to make predictions and forecasts



Data Engineers

- Build data systems that allow data scientists and data analysts to perform their work



Data Analysts/BI Developer

- Deliver data in an understandable format to help make business decisions

Data Engineering



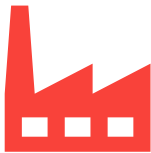
Get data to where it's needed



Get data into a usable condition



Manage all the data after the process



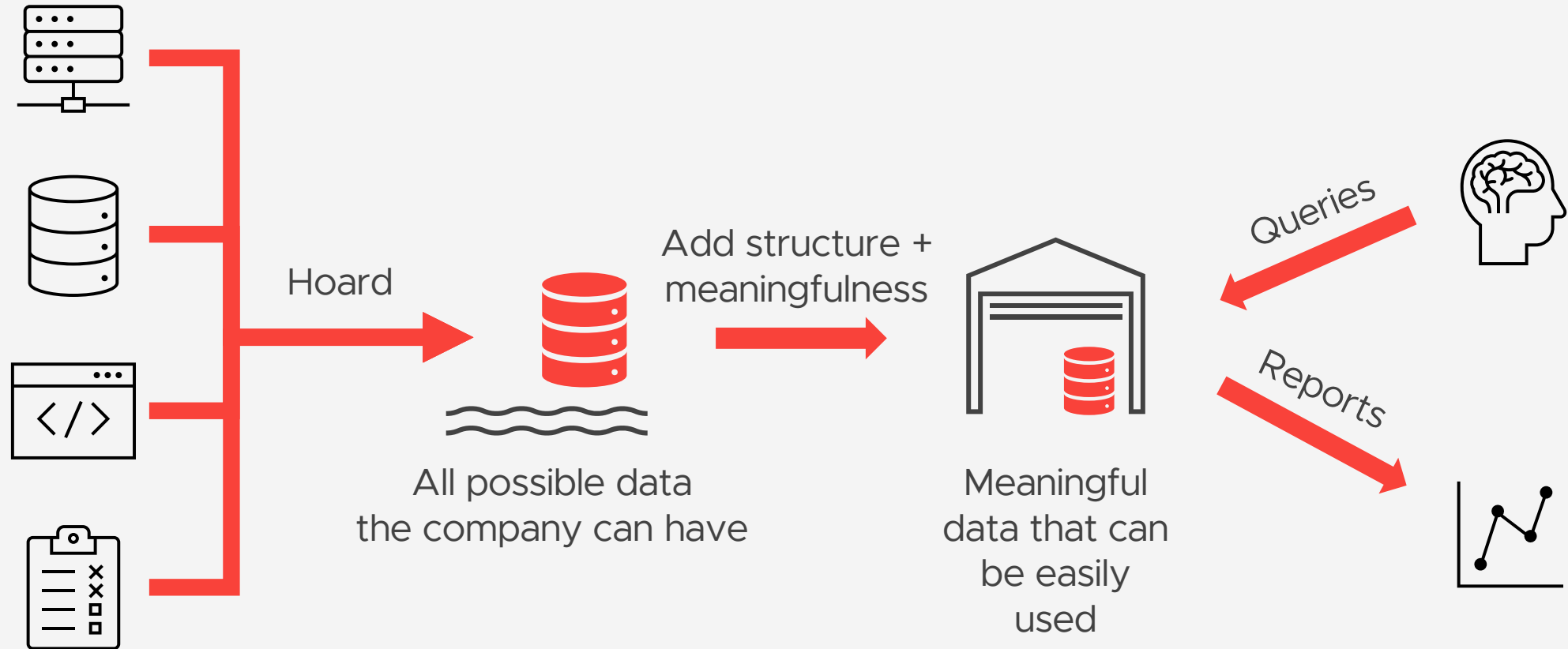
Productionize the process

./ADASTRA

Data Platforms



Data Pipeline





Transactional Database

- Transactional: fast retrieval, fast updates
- Structured
- Silo-ed for specific departments or function
- **Online Transactional Processing (OLTP)**



Data Lake

- Giant reservoir of data in any forms
- Can be in unprocessed format and unstructured data. Excel files, voice, images, anything.
- Flexibility for exploration
- **Focus on volume over usability**

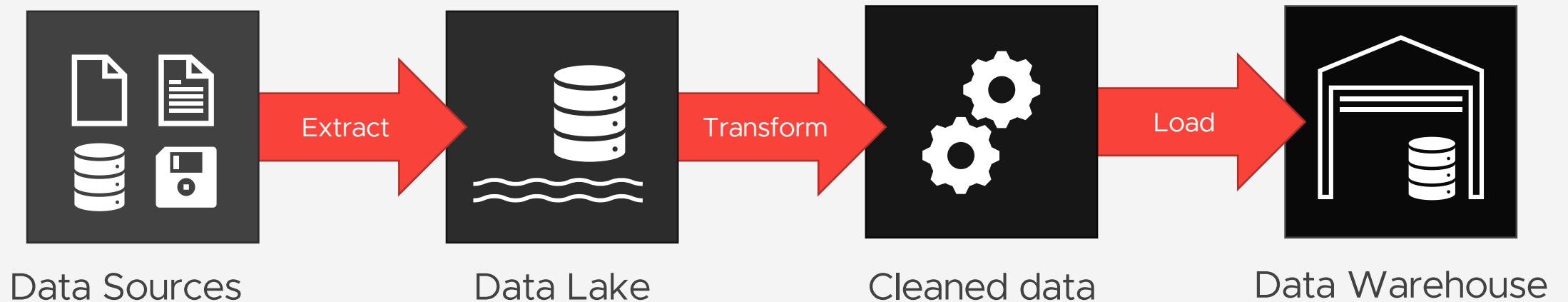


Data Warehouse

- Central repository for processed and managed historical data
- **Ideally not silo-ed**
- **Designed and structured for large scale analytical purpose**
- Prioritize complex queries and analysis over speedy updates
- Allow answering of specific questions
- **Powerful: need its power for the “add meaningfulness” part and data retrieval**
- **Online Analytical Processing (OLAP)**

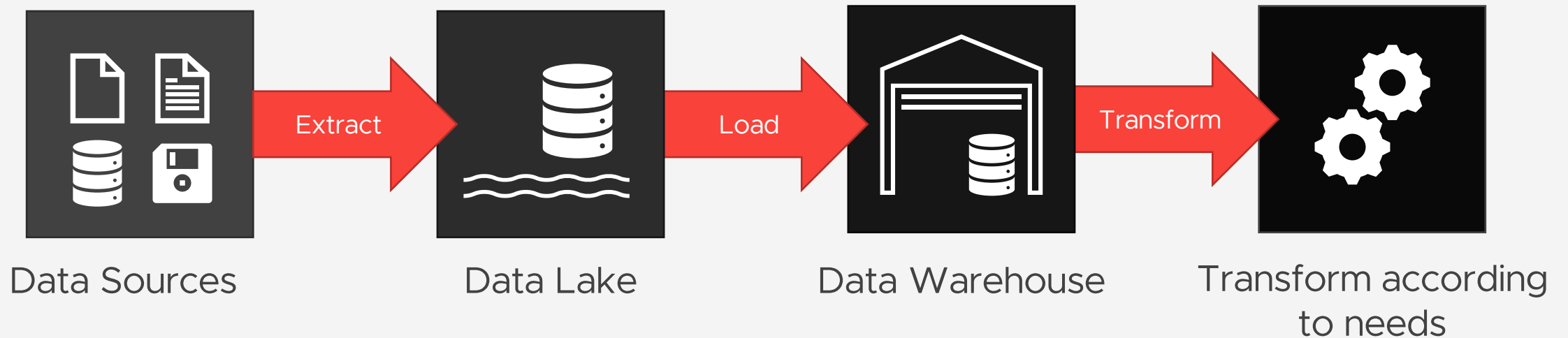
Extract-Transform-Load (ETL)

- Extract, Transform, and Load (ETL) is the traditional approach for data warehousing processes.
- Clean the data to answer business questions first.
 - Example: source data is daily, but business wants nothing more than monthly data - then sum it up
- Data in warehouse adheres to a structure per business requirement.



Extract-Load-Transform (ELT)

- Modern data warehousing approach do ELT (Extract, Load, and Transform) instead of ETL.
- **Transform** after **Load** so that we can transform per different requirements.
- Capable by advantages of scalability and flexibility in Cloud Computing.

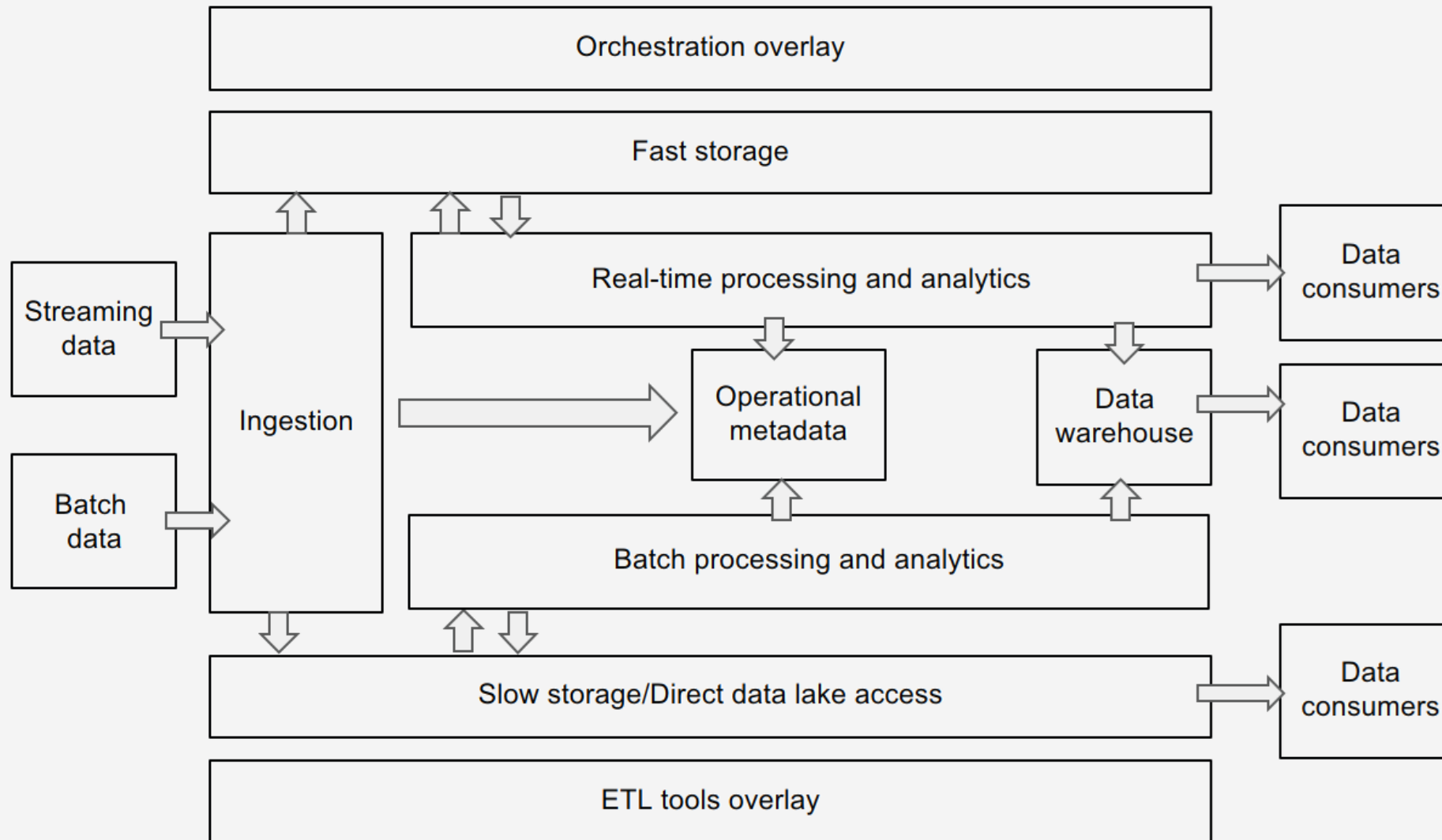




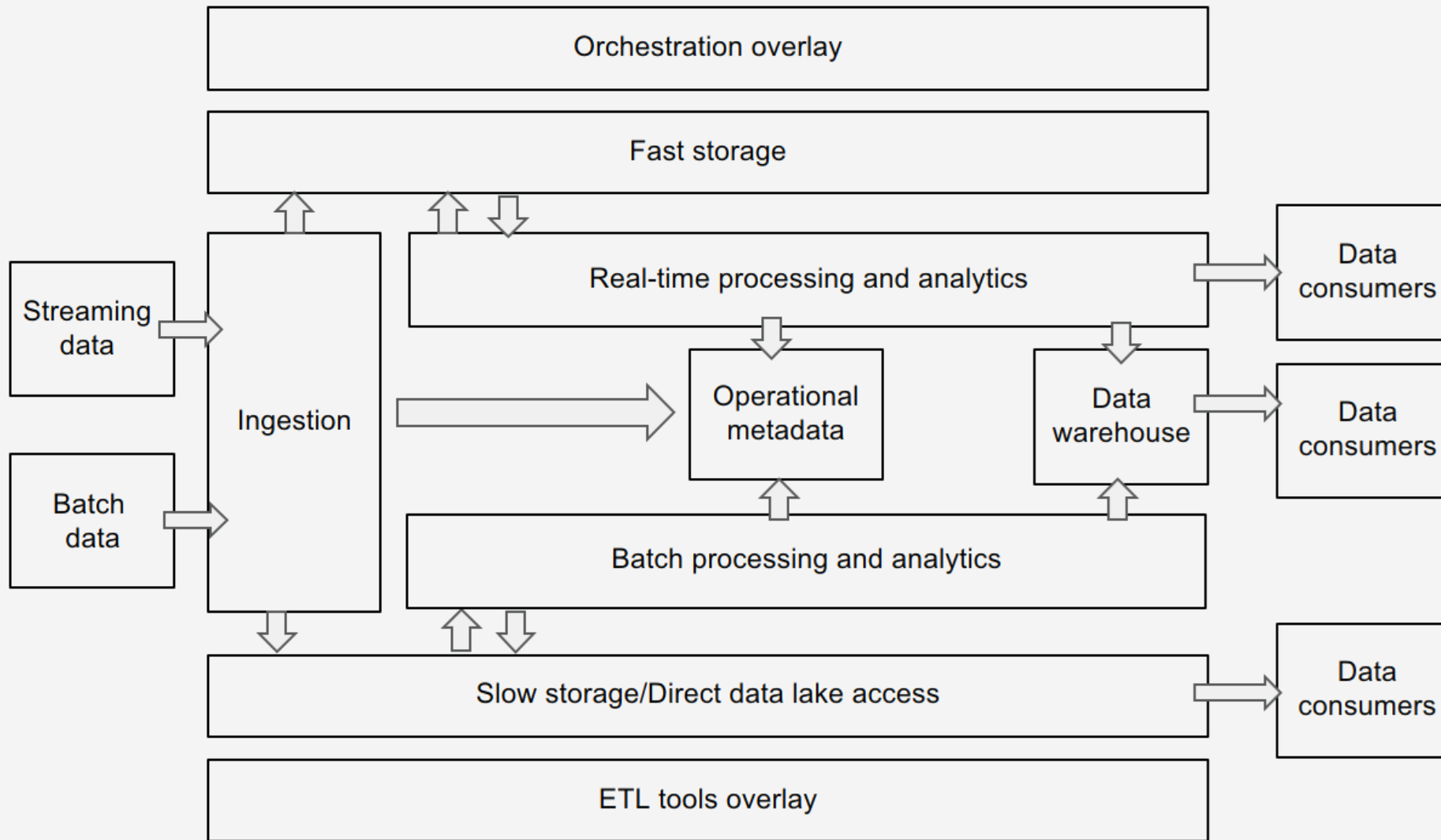
Data Lake House

- **Flexibility of Data Lake + Rigidity of transformed data ready to answer business questions of Data Warehouse**
- Storage in Lake
- Compute unit somewhere else
- Write results back to Lake
- Query from Lake!

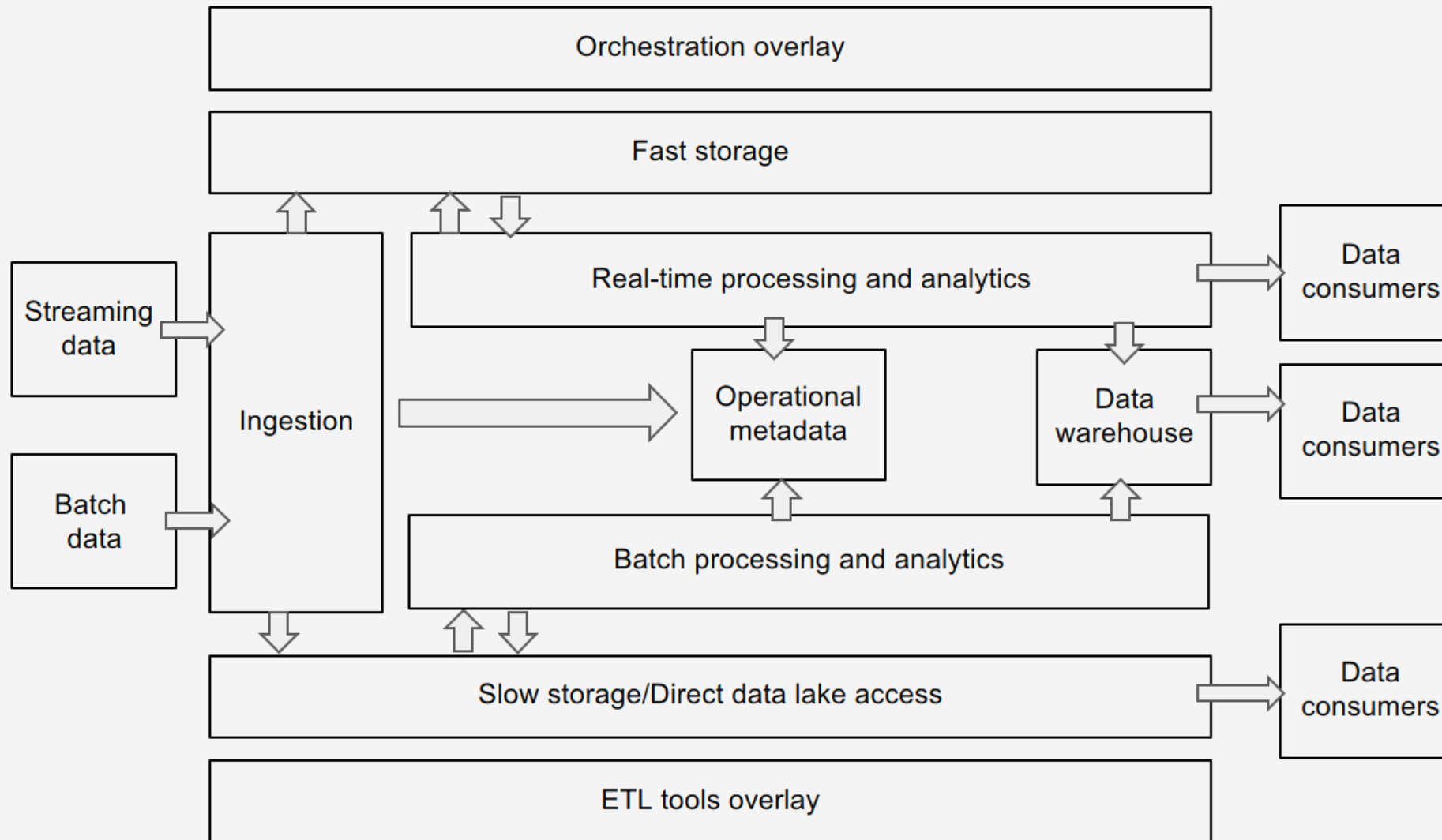
Components of Data Pipelines



Components of Data Pipelines (continued 1)



Components of Data Pipelines (continued 2)



Medallion Layers of Data Lake House

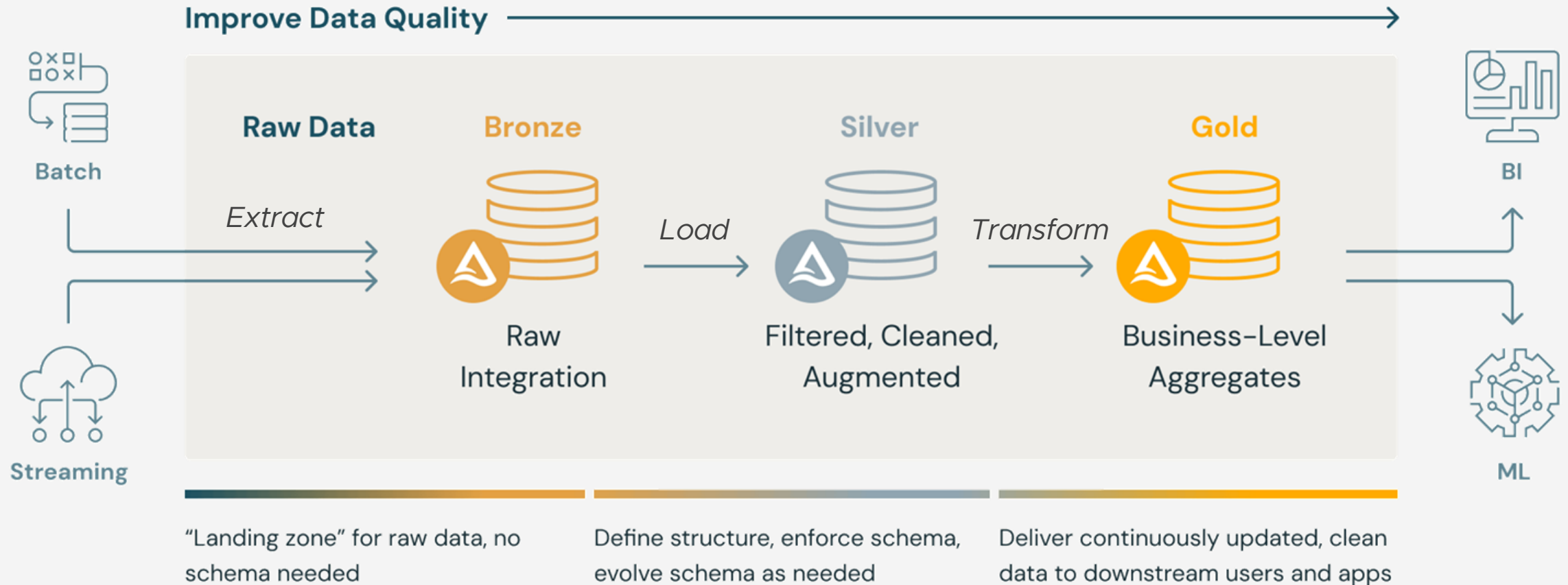


Image courtesy: Databricks



**Massive computation
= Massive computers needed**

A large, fluffy white cloud is centered in the frame, set against a solid dark blue background. The cloud has a soft, billowy texture with varying shades of white and light grey. Overlaid on the center of the cloud is the text 'Cloud Data Platforms' in a large, bold, white sans-serif font. Below this, the text 'Meaning: someone else's computer' is written in a smaller, regular white sans-serif font.

Cloud Data Platforms

Meaning: someone else's computer

Apache Spark

- Open-source unified analytics engine built for large-scale data processing.
- Single machine or across clusters of computers.
- Speed + ease of use -> popularity
- Java/Scala/Python



Databricks

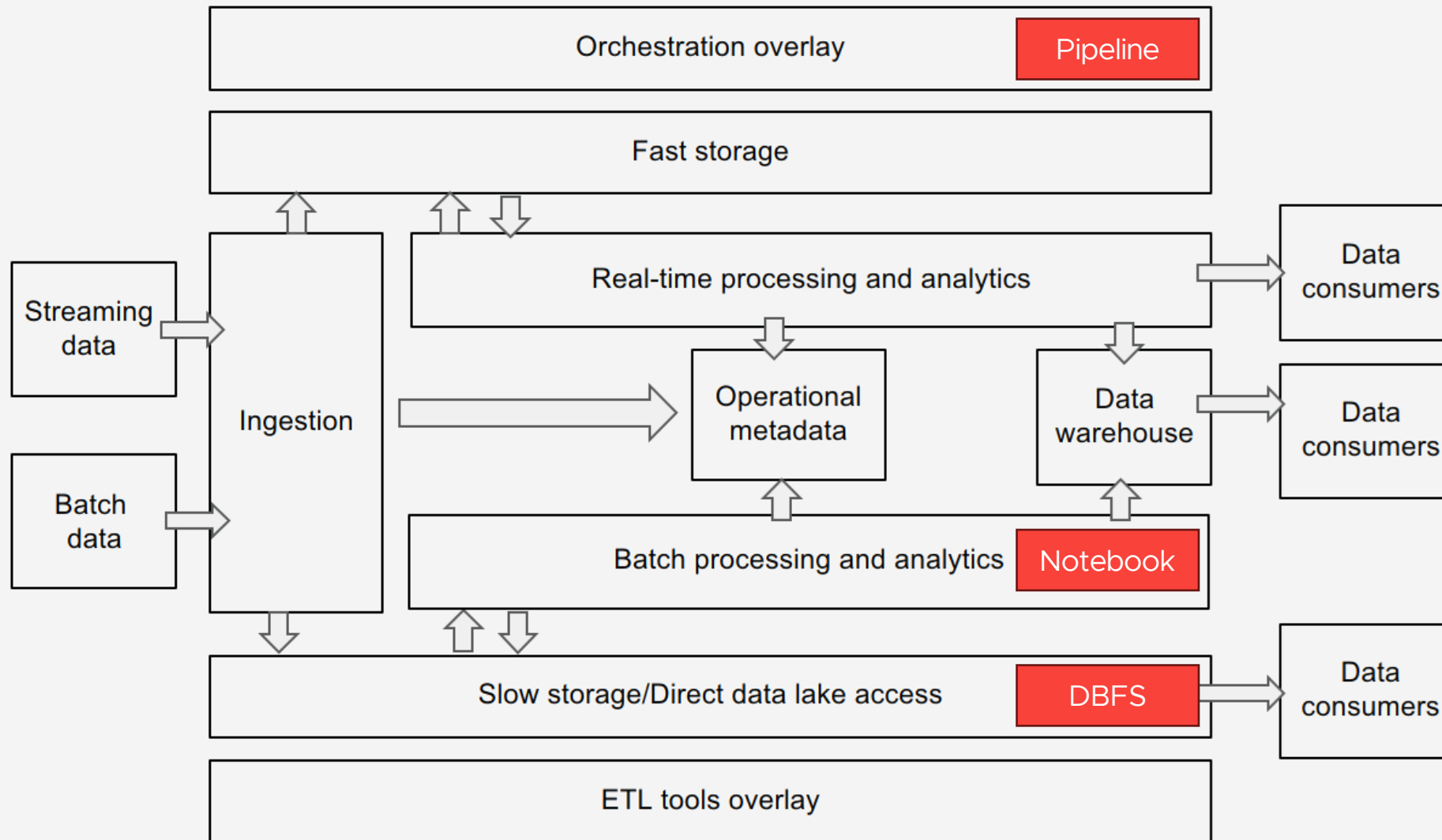
- Spark on the cloud
- Less hassle managing Spark cluster
- Provides useful features rather than computing engine
 - GUI for development
 - Data catalog
 - Orchestration*

** non-free plan only*



databricks

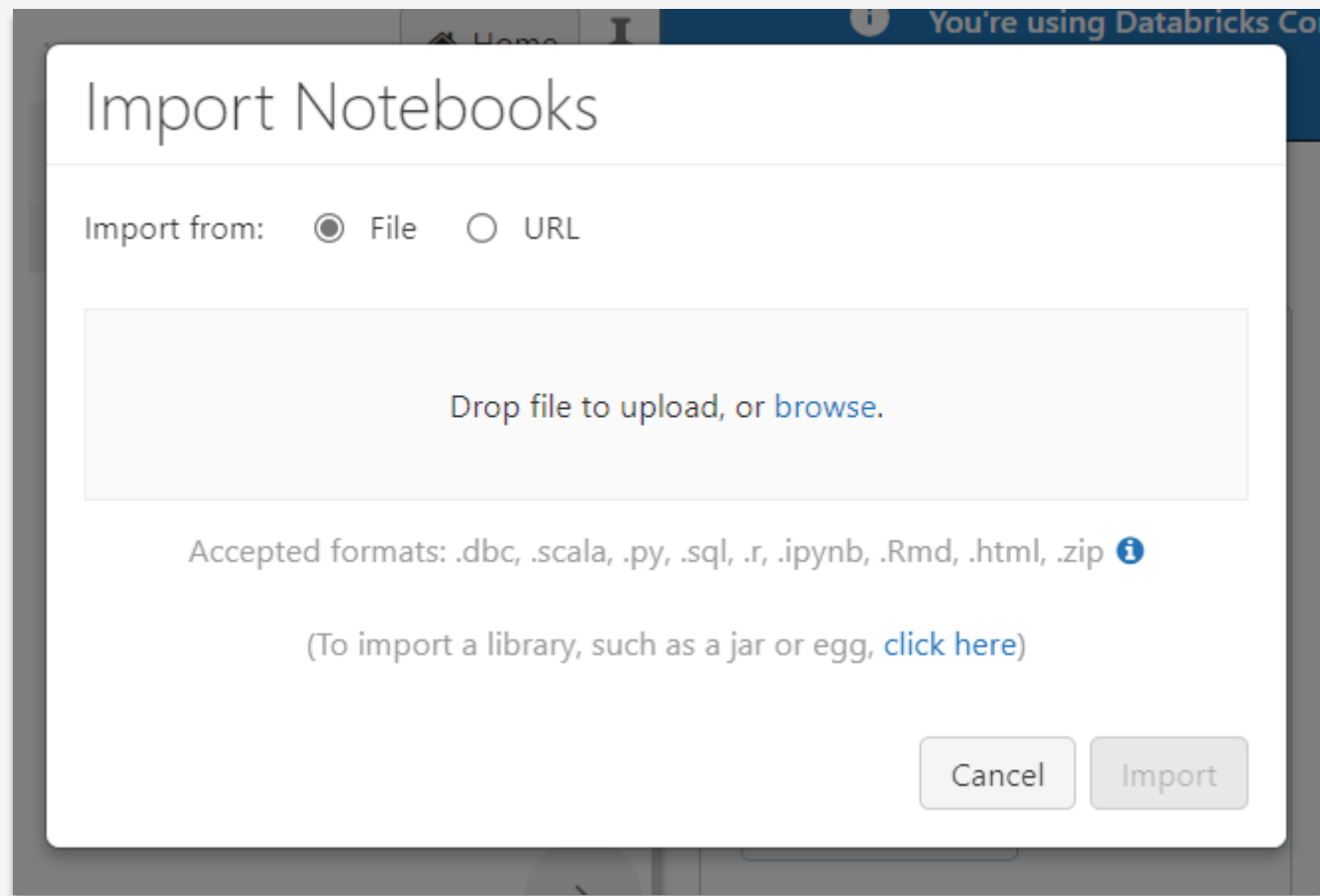
Data Pipelines on Databricks



//ADASTRA

Databricks Lab





<https://github.com/AdastraTH/2024-univ-workshop/raw/main/notebooks/ATH%20Workshop%202024.dbc>