

# Introduction to Data Engineering on Databricks

Adastra Thailand Campus on-tour program

Chiang Mai University, 3 July 2024



# Before we start...

*(take photos first, we will advance real quick!)*

**github.com/AdastraTH/2024-univ-workshop**



- Check notes down below file list which will share with you how to...
  - Grab the slides
  - Sign up for Databricks Community Edition  
**(which you should do during the lecture)**
  - Download Power BI Desktop  
(or find ways to work with Power BI web app if you are a Linux/Mac user)

# About the Speakers



**Sirakorn Lamyai (Tan)**

Practice Lead + AWS Lead



**Nat Rattanarom (Art)**

HR Recruiter

# Meet our team



**Apichart Hortiangtham**

Data Science Lead



**Methasit Pengmatchaya**

Google Cloud Platform Lead



**Chatchadaporn Saradet**

Data Engineer



**Jitdawan Pawanna**

Data Engineer



**Siwat Tansiri**

Data Engineer

Before we start...





CHIANG MAI UNIVERSITY



First teaching: 2503  
Establishment: 2507



First teaching: 2447  
Establishment: 2486



First teaching: 2477  
Establishment: 2539

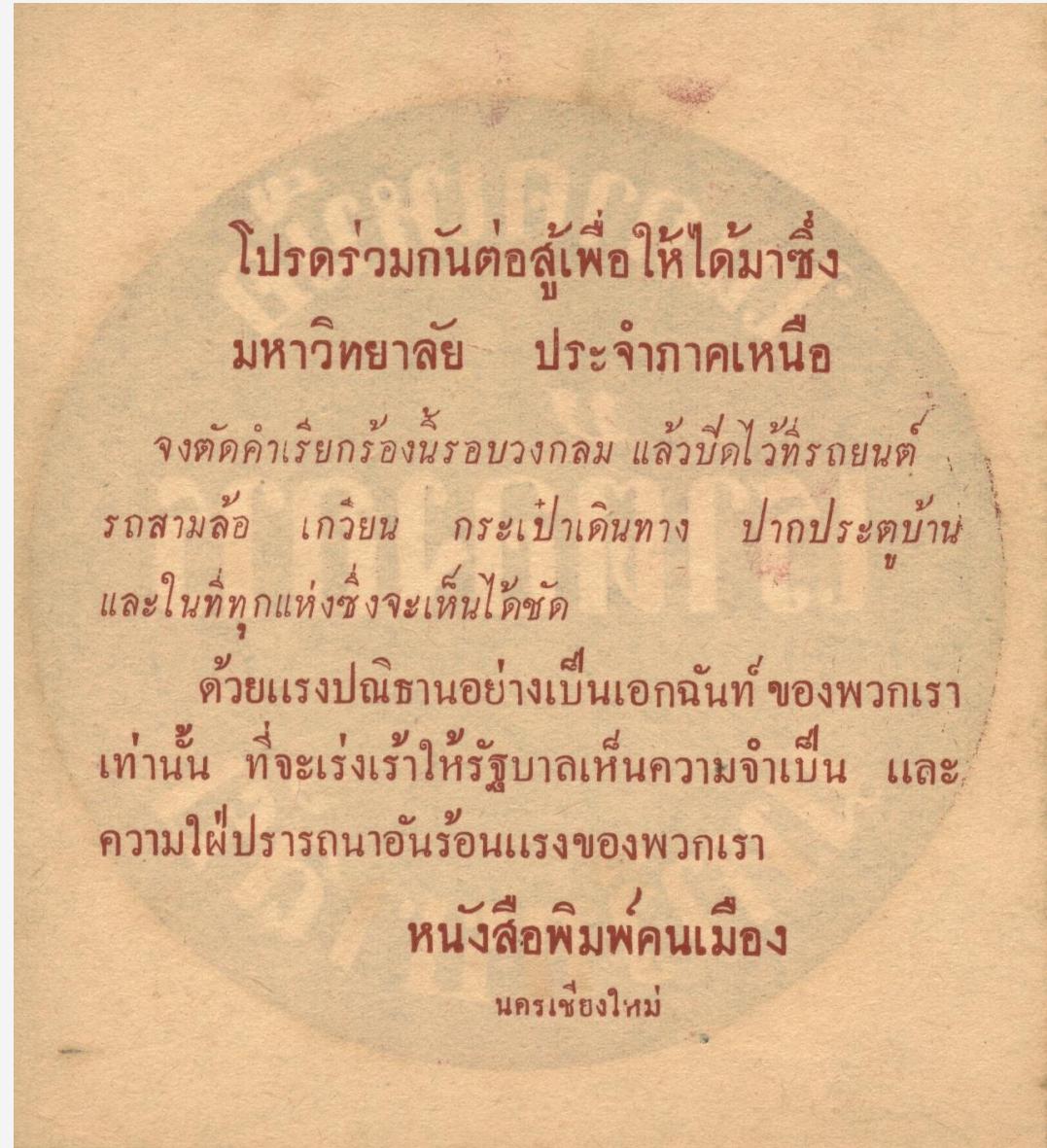


Image courtesy: Chiang Mai University's Historical Development Hall (Pin Mala Art Hall)

Image courtesy:  
Facebook page  
“History of  
Lampang”, fair  
use (low  
resolution)



Image courtesy:  
“เทคนิคการผลิต  
รายการโทรทัศน์  
ของสถานีโทรทัศน์  
ช่อง 8 ลำปาง” by  
ลัดดา เบญจเตชะ,  
a Thammasat  
University  
dissertation

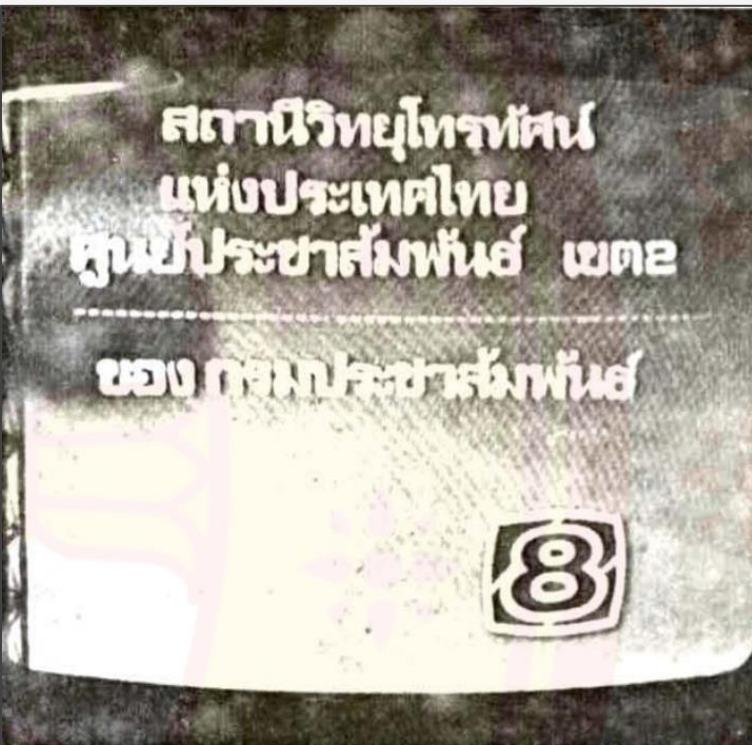




Image courtesy: MAIIAM Contemporary Art Museum



Image courtesy: KAD Theater

**ADASTRA (THAILAND) CO., LTD.**

บริษัท อดาстра (ไทยแลนด์) จำกัด

# 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

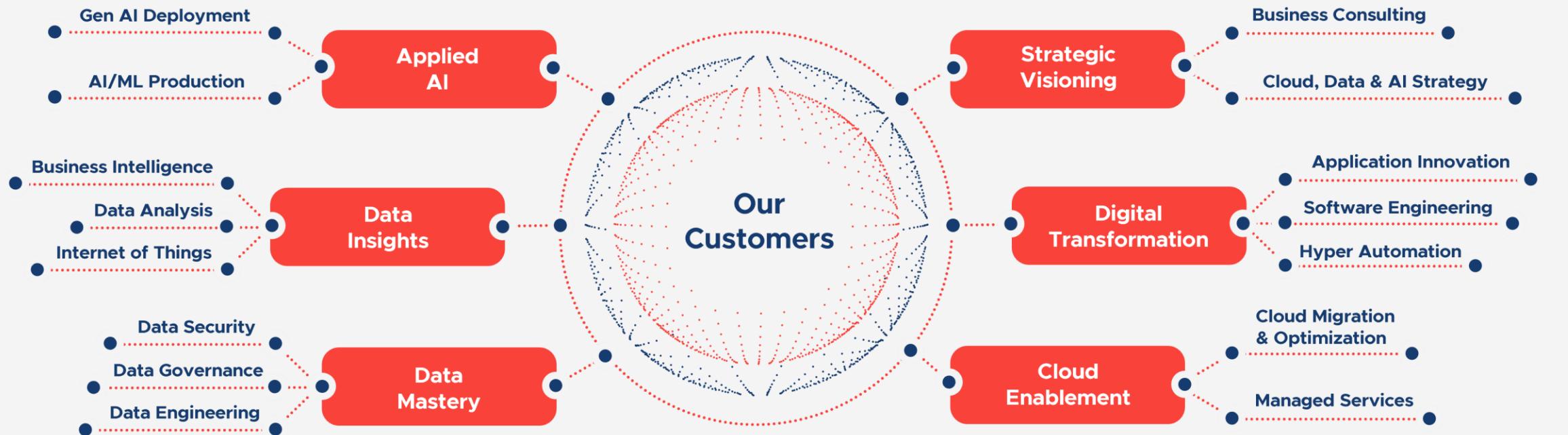
**2,200+**  
Professionals

**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

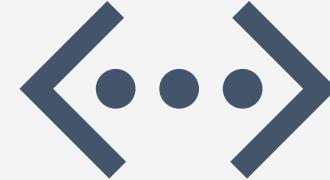
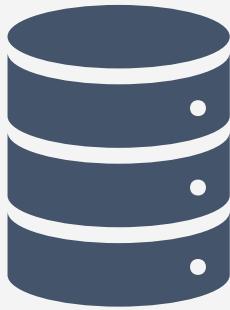


**„ADASTRA**

## 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

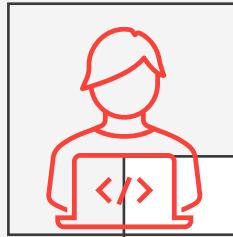
# Data Careers

## 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



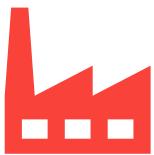
Get data to where it's needed



Get data into a usable condition



Manage all the data after the process



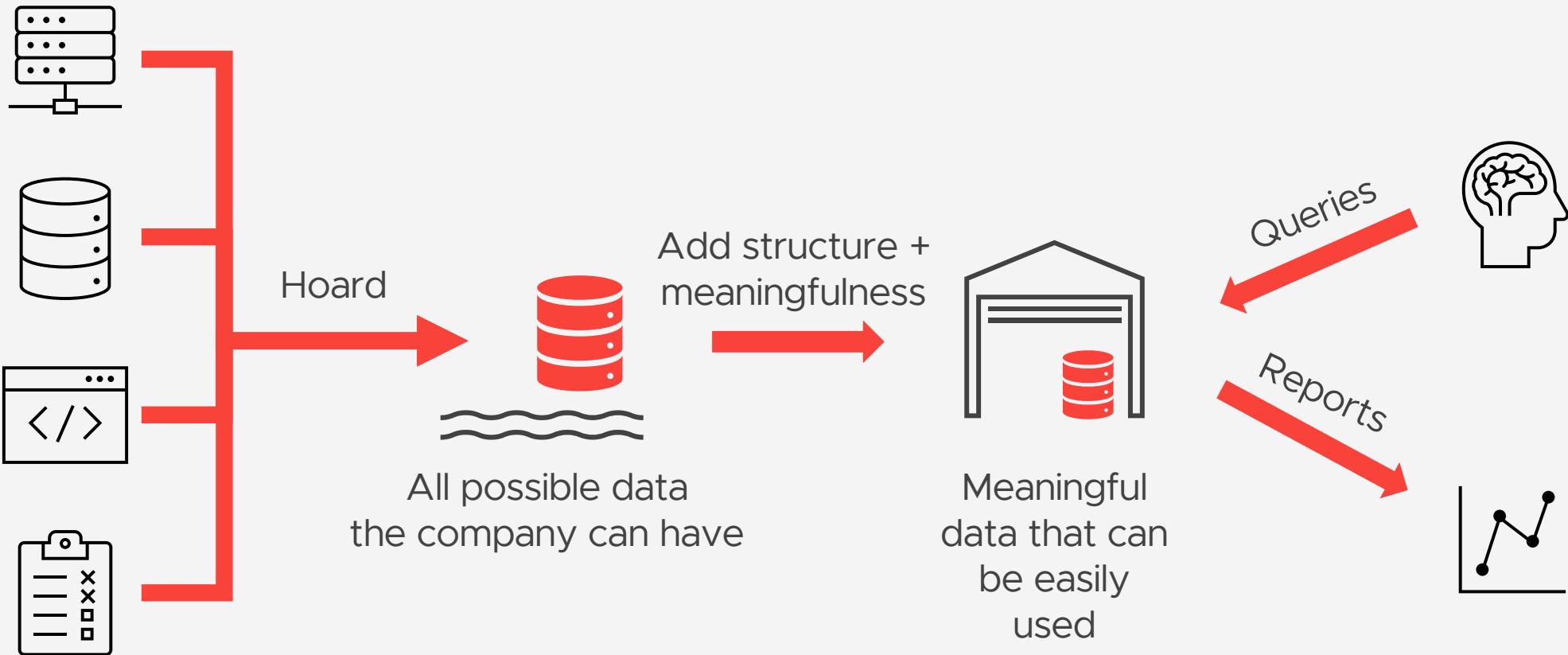
Productionize the process

**„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**

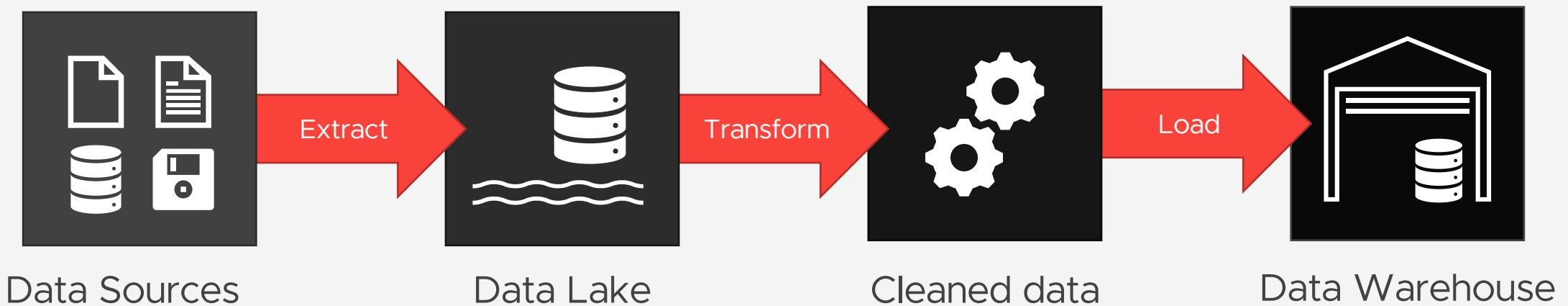


# Data Warehouse

- 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)**

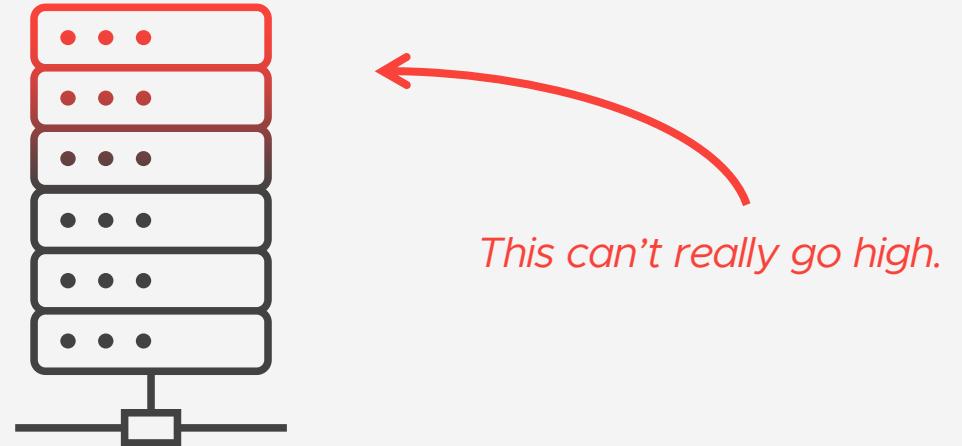
# Extract-Transform-Load (ETL)

- 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.

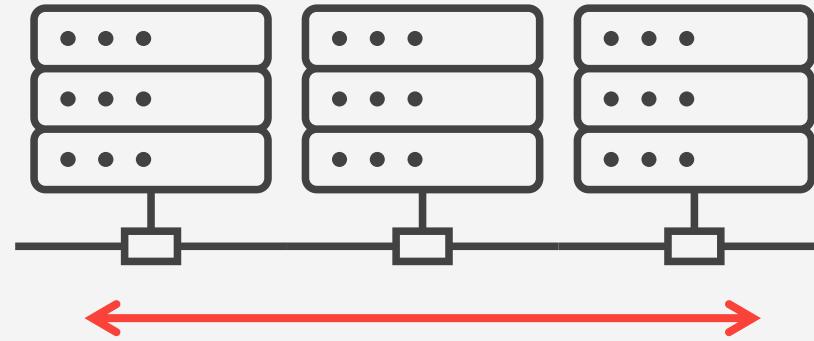


# Computation Scaling

- We can scale up our system by adding more resources to a single computational unit.
  - Exists limitations such as bottlenecks.
- We can scale out our system by connecting many smaller systems, therefore creating a distributed system.
  - Achieved Distributed Computing



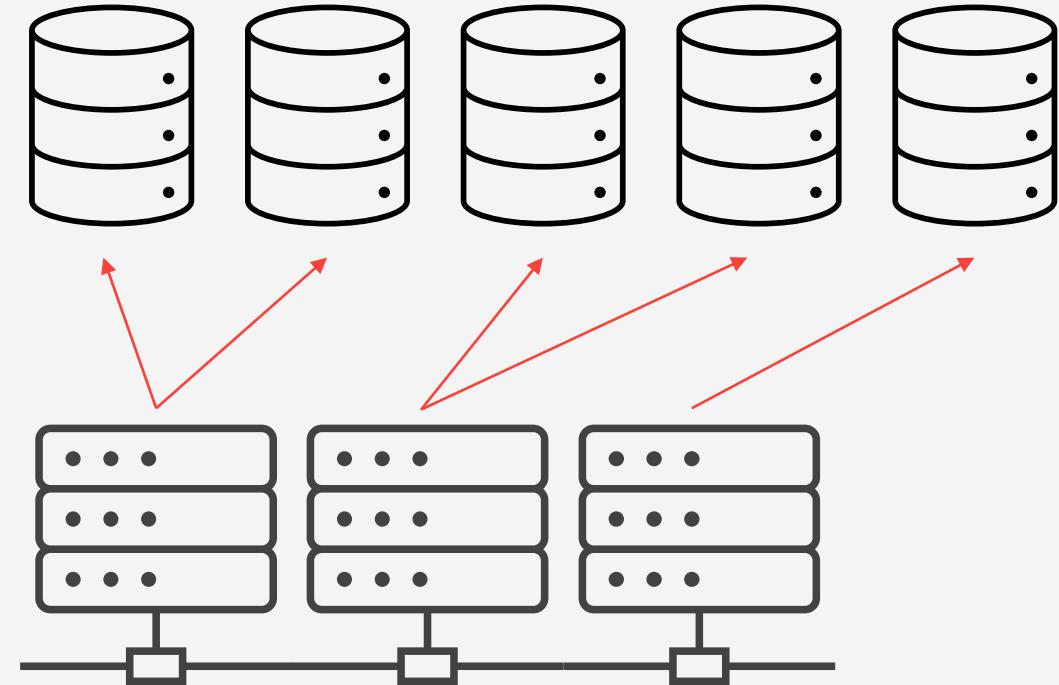
*This can't really go high.*



*This requires distributing and “talking” between devices.*

# Storage-Compute Decoupling

- Storage and compute demand does not scale proportionally!
- We eventually managed to decouple them and create a flexible solution.
- Still, some analytics workload are harder than others.



# Distributed Computing!

Calculate summation  
of these numbers



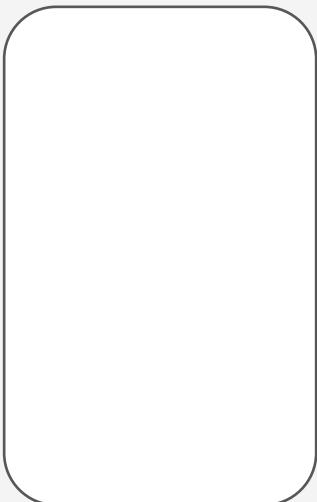
**Storage**

Workers, here is the plan: grab four each, sum  
them up, and let me know...

Master

Worker 1

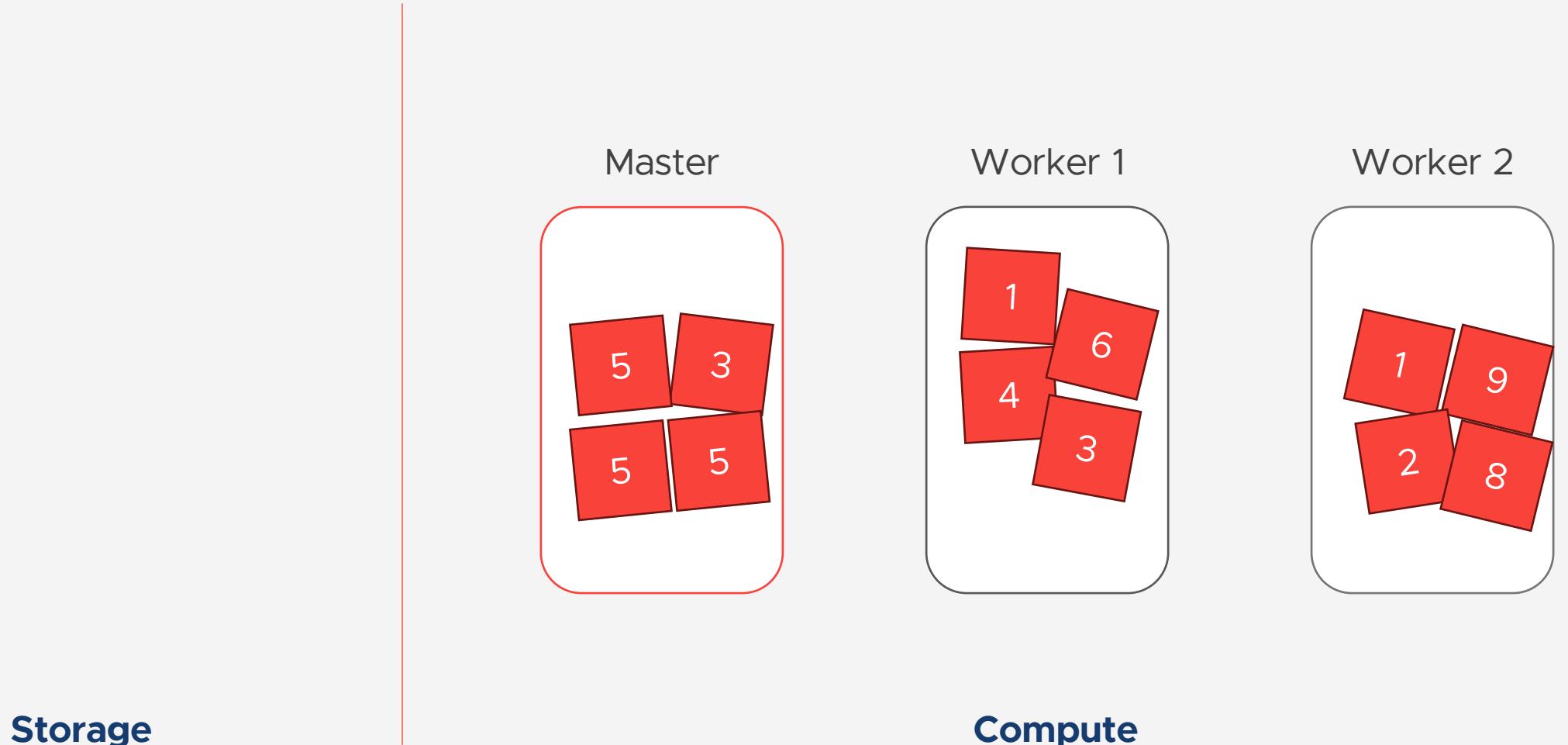
Worker 2



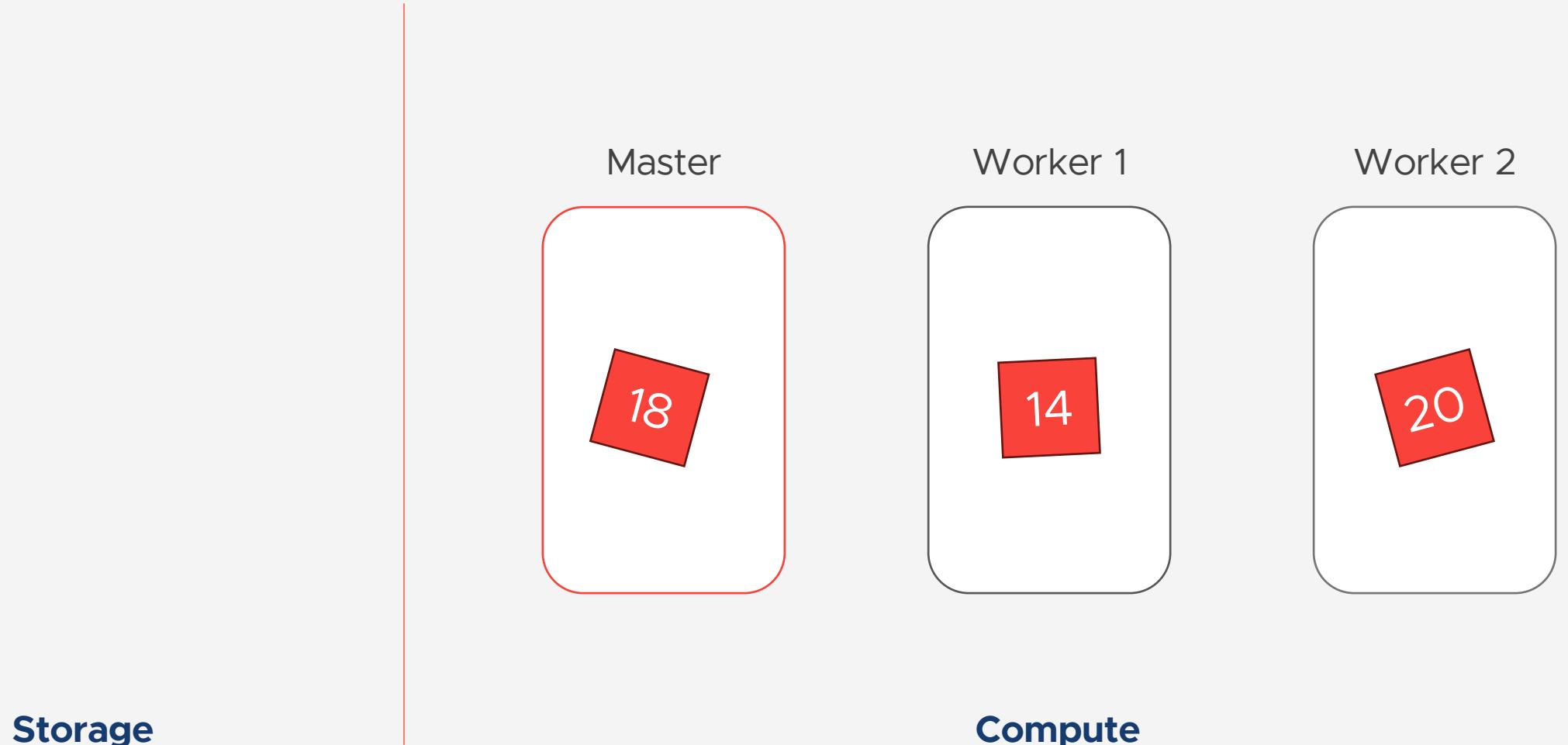
**Compute**



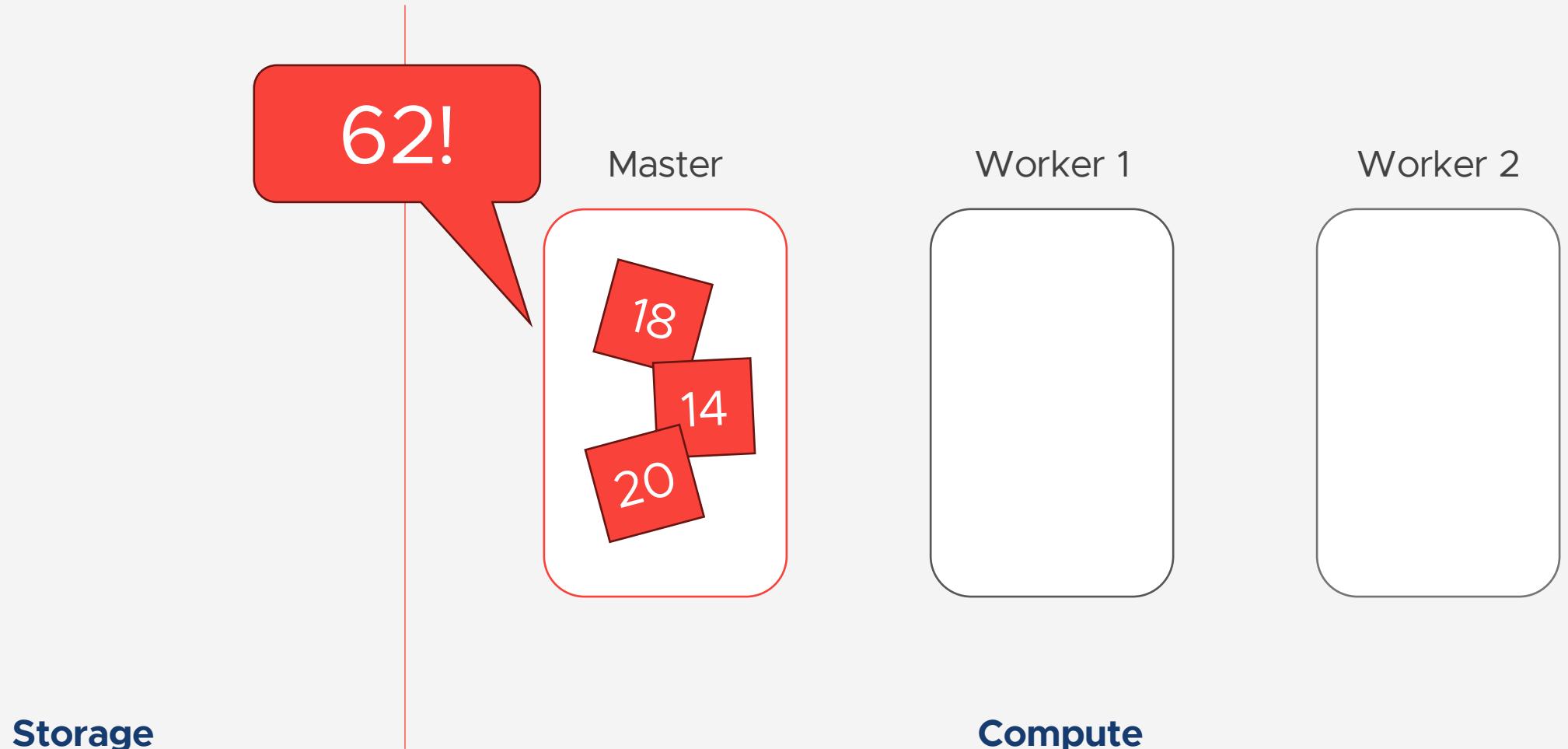
# Distributed Computing!



# Distributed Computing!



# Distributed Computing!

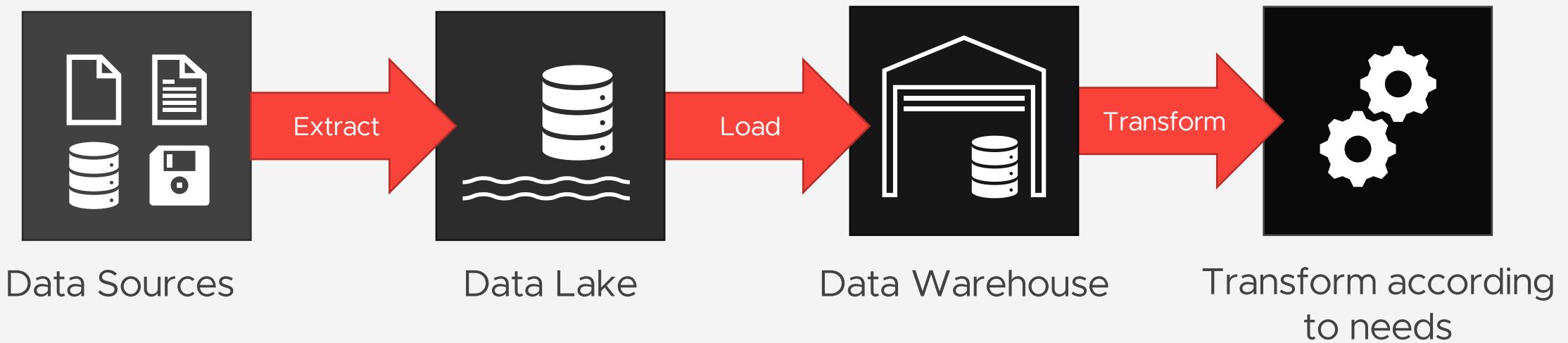


\* That is an exclamation mark, not a factorial sign.



# 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.

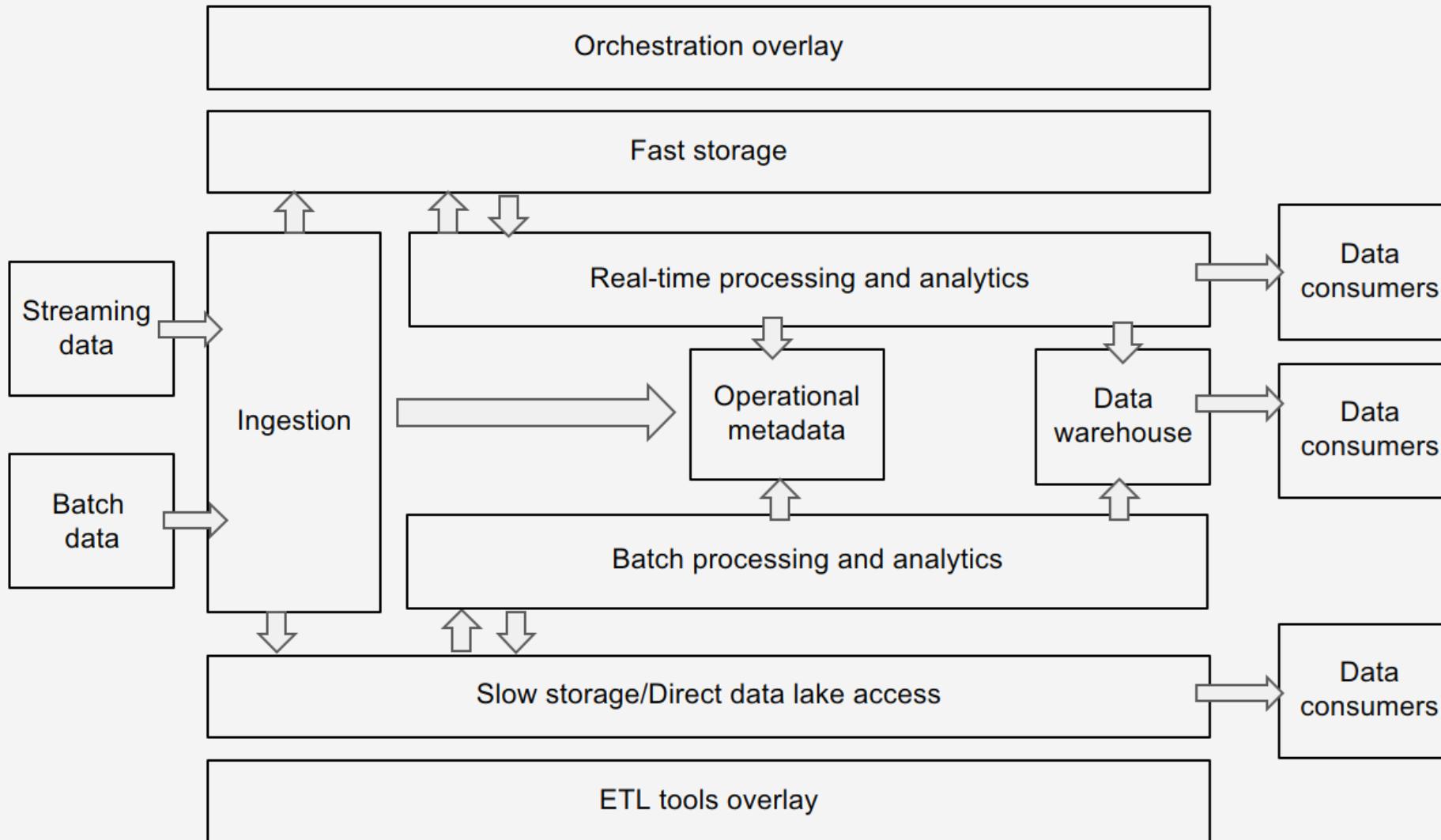




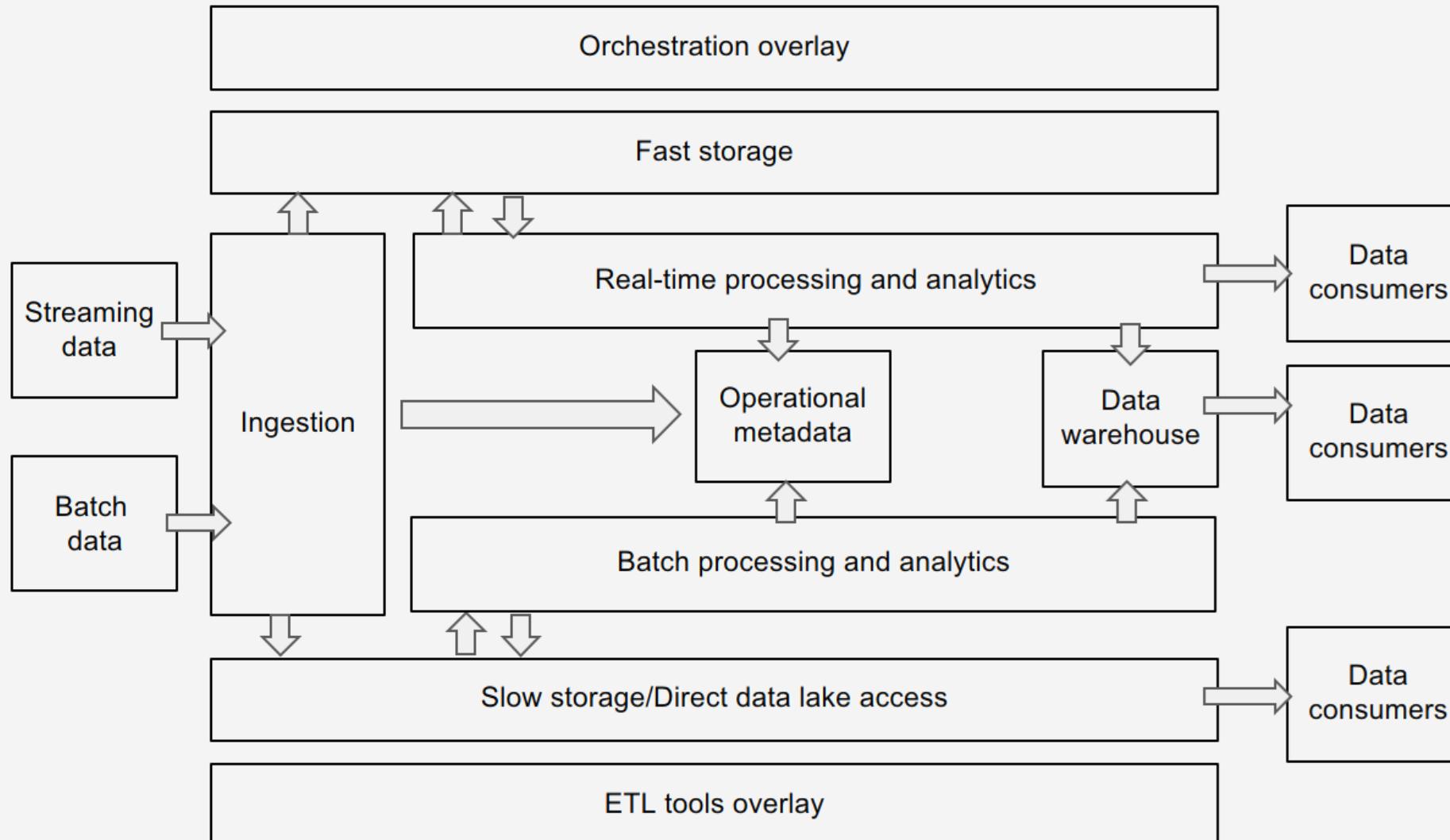
# Data Lake House

- **Flexibility of Data Lake + Rigidness 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!

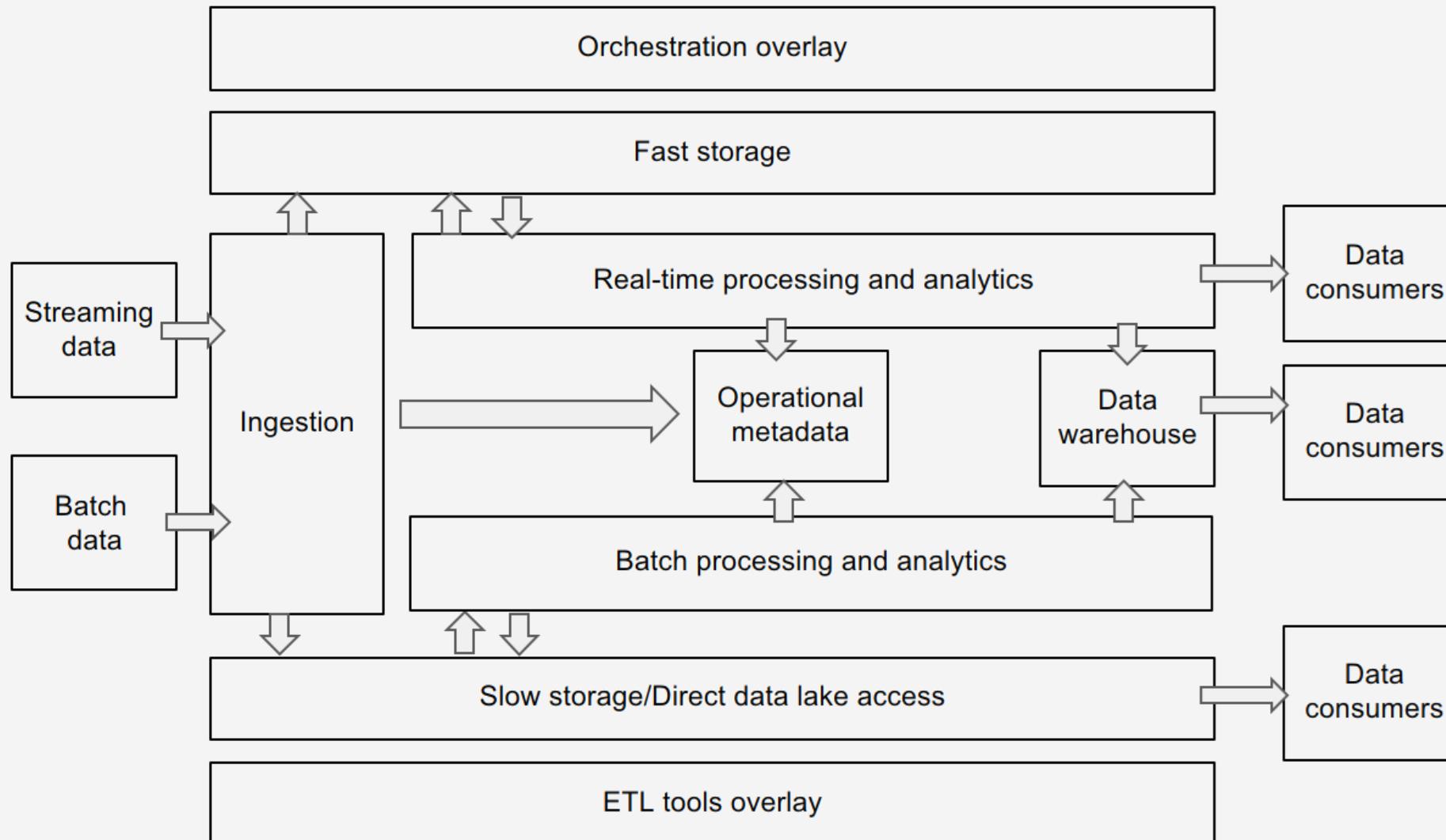
# Components of Data Pipelines



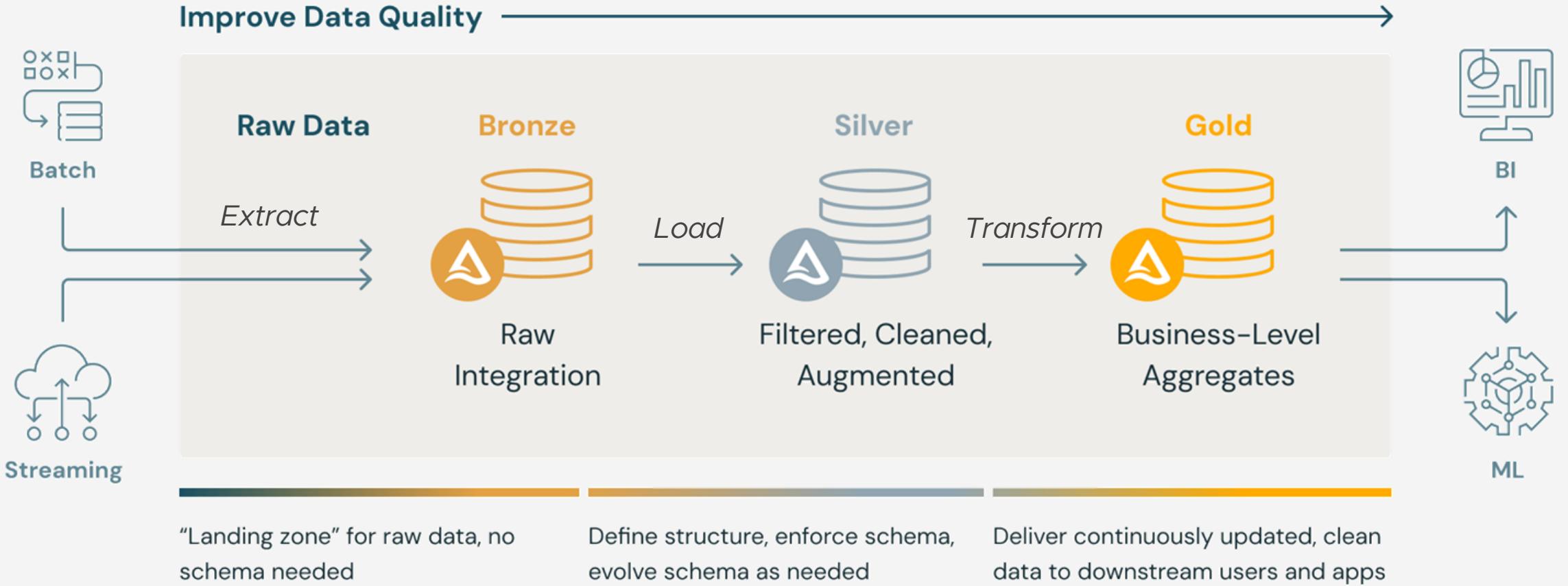
# Components of Data Pipelines (continued 1)



# Components of Data Pipelines (continued 2)



# Medallion Layers of Data Lake House



*Image courtesy: Databricks*



**Massive computation**  
**= Massive computers needed**

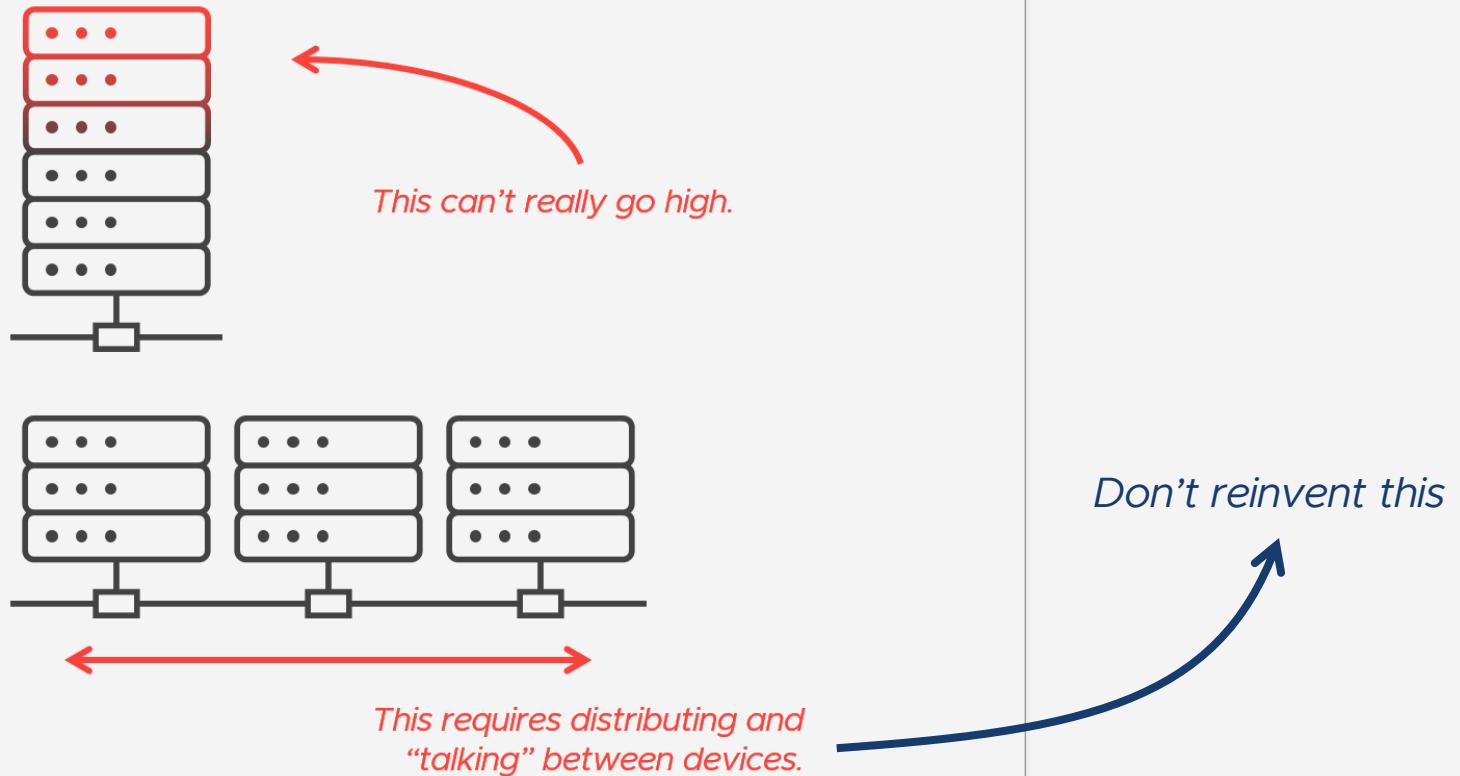
The background of the image is a dark, deep teal or slate blue color. In the center, there is a large, puffy white cloud. The cloud has a dense, textured center with some darker grey areas, and it tapers off into smaller, wispy clouds on the left and right sides.

# **Cloud** Data Platforms

Meaning: someone else's computer

# Computation Scaling

- We can scale up our system by adding more resources to a single computational unit.
  - Exists limitations such as bottlenecks.
- We can scale out our system by connecting many smaller systems, therefore creating a distributed system.
  - Achieved Distributed Computing



# 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



# Spark Core

Spark SQL +  
DataFrame

Streaming

MLLib

GraphX

Spark Core APIs

R

SQL

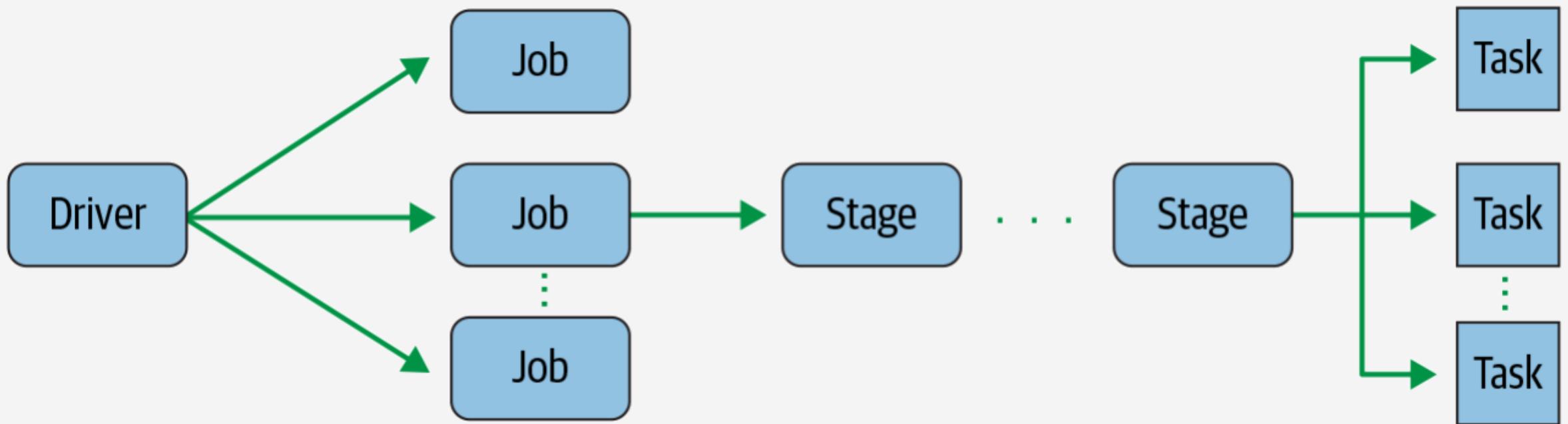
Python

Scala

Java



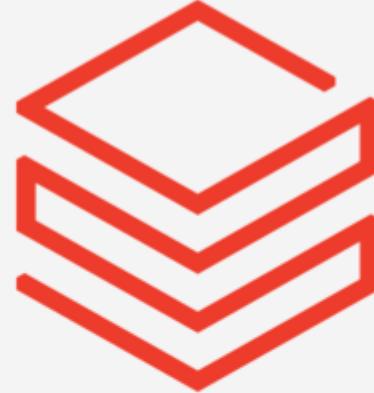
# Spark Execution



# Databricks

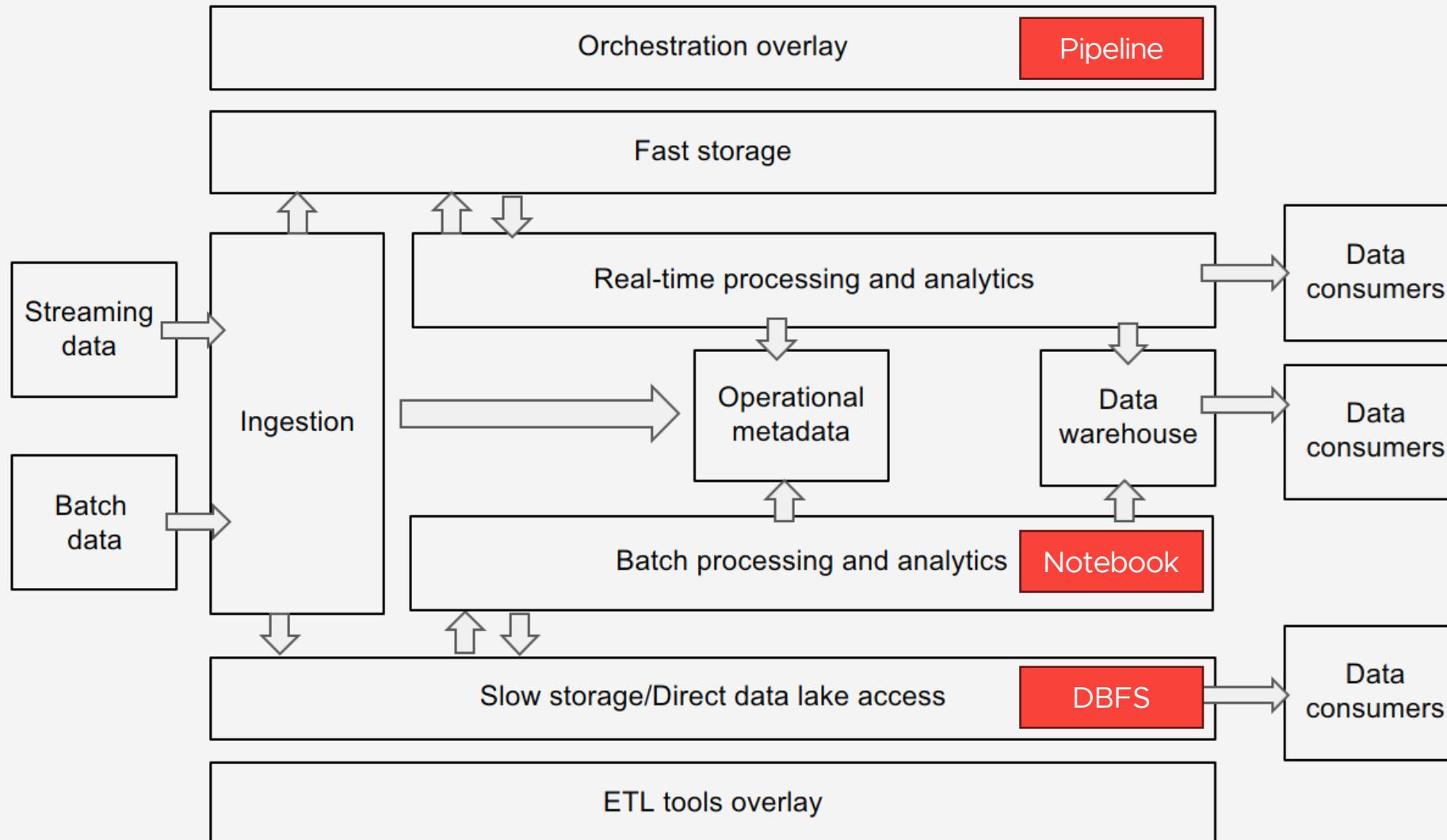
- Spark on the cloud
- Less hassle managing Spark cluster
- Provides useful features rather than computing engine
  - GUI for development
  - Data catalog
  - Orchestration\*

\* non-free plan only



# databricks

# Data Pipelines on Databricks



**„ADASTRA**

## Databricks Lab



The screenshot shows the Databricks Community Edition workspace interface. The top navigation bar includes a user profile icon, a 'Work' button, and several other tabs like 'Databricks Community Edition' and '2024-univ-workshop/notebooks/'. Below the header is a red bar with back, forward, and search icons, and the URL <https://community.cloud.databricks.com/?o=5292693329990630#>. The main area has a dark header with the 'databricks' logo and a 'New' button. The left sidebar lists 'Workspace' (selected), 'Recents', 'Search', 'Catalog', 'Workflows', 'Compute', 'Machine Learning', and 'Experiments'. The central workspace is titled 'Workspace' and shows 'Users' dropdowns for 'sirakorn.lamyai@adastragrp.com' and 'ATH Workshop 2024'. A 'Create' menu is open in the bottom right corner, listing 'Create', 'Import', and 'Permissions'.

Work

Databricks Community Edition

2024-univ-workshop/notebooks/

← → 🔍 https://community.cloud.databricks.com/?o=5292693329990630#

databricks

+ New

Workspace

Recents

Search

Catalog

Workflows

Compute

Machine Learning

Experiments

Workspace

Users

sirakorn.lamyai@adastragrp.com

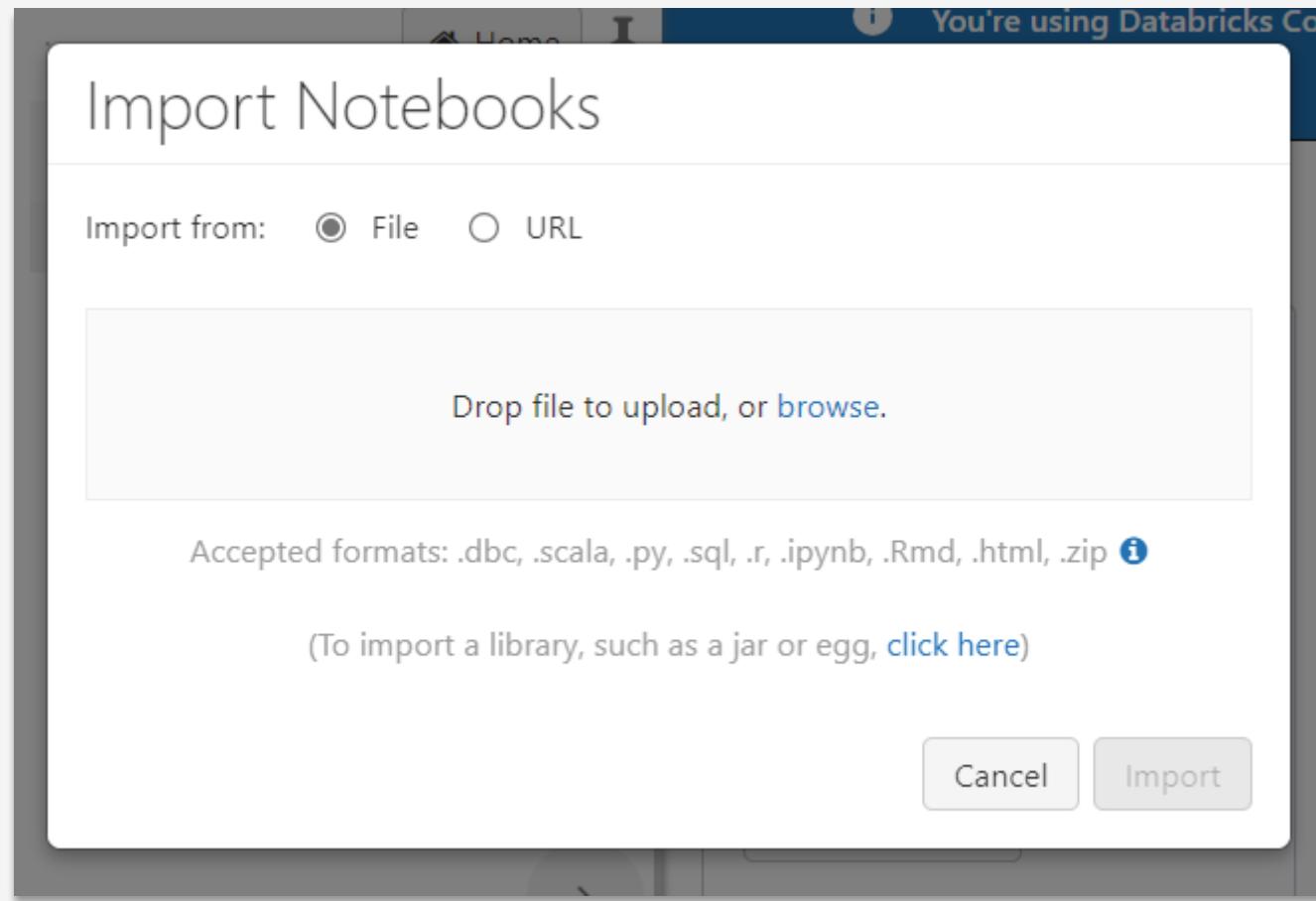
ATH Workshop 2024

Home

Create

Import

Permissions



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