

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



Thanyaporn S.

Recruitment Manager




Manassaphorn W.

HR Manager

Adastra's Global Presence



 **Adastra Offices**



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

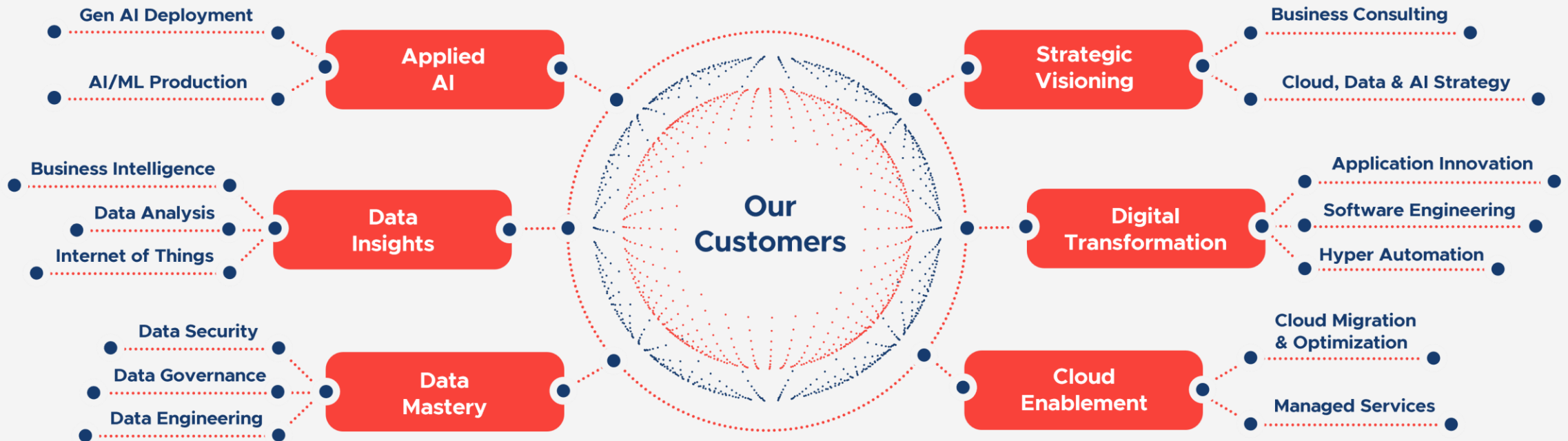


20+

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:



Sound check!

Ever heard of these terms?



Database



SQL



Data Lake



Data Warehouse



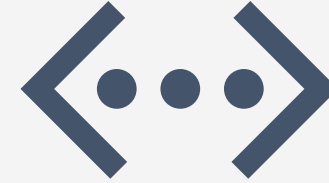
Business Intelligence

./ADASTRA

Data and Data Engineering



Big Data: how can it be massive?



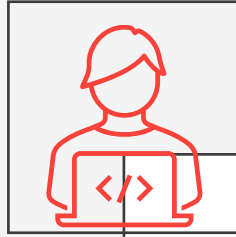
- Cheaper device makes it possible to generate massive data.
- Cheaper storage unit makes it possible to store data first without thinking whether to use it or not.
- Internet makes it capable for users to distribute massive amounts of data.
- How can we process them?
- What are the aspects of processing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

Data Careers



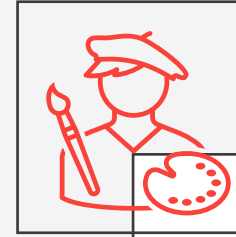
Data Scientists

- 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



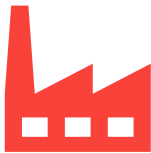
Get Data to where it's needed



Get data into a usable condition



Manage data



Productionize the process

./ADASTRA

Data Platforms





Database

- For data collection
- Silo-ed for specific departments or function
- Mostly transactional
- Fast retrieval, fast updates
- **Online Transactional Processing (OLTP)**

How can we make the most of these data?



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
- **Online Analytical Processing (OLAP)**

How can we store even more types of data?

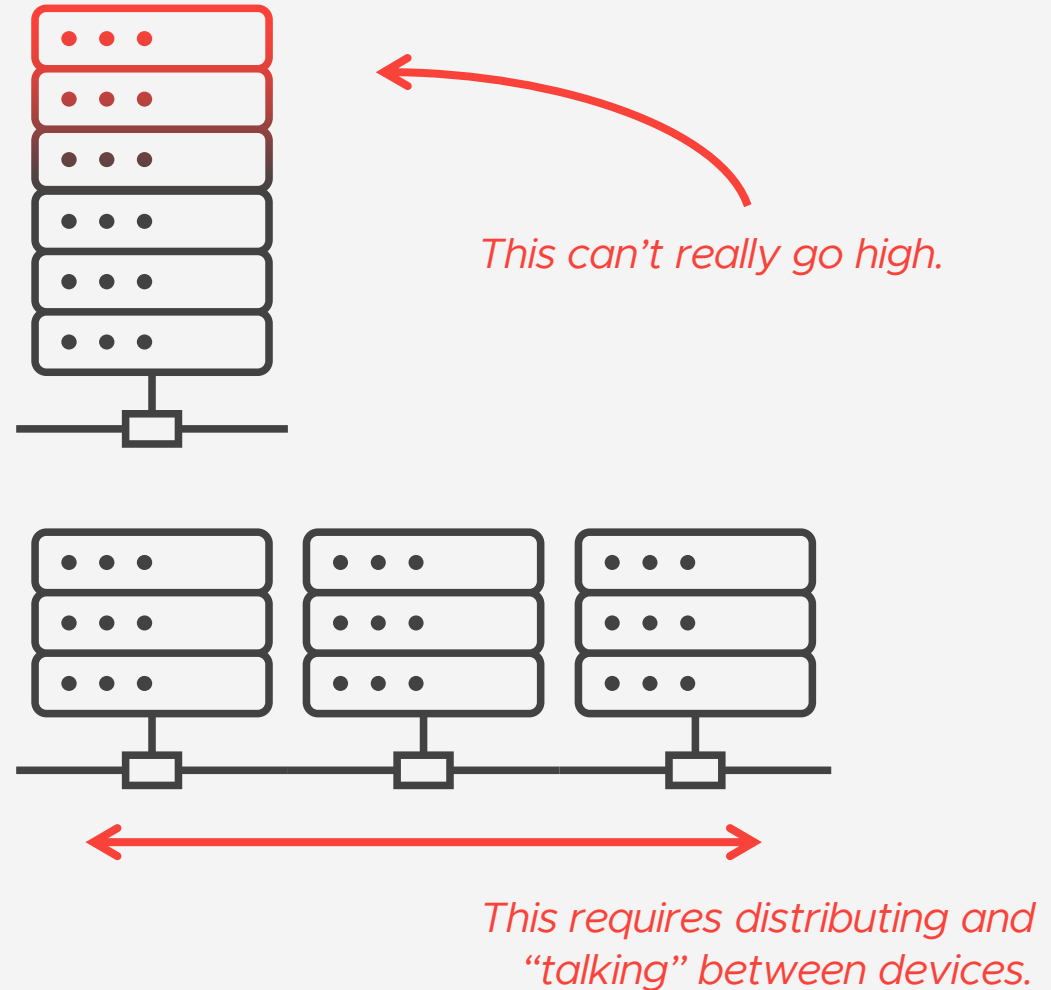


Data Lake

- Giant Reservoir of data in any forms, including unprocessed format and unstructured data.
- Can be literally anything from Excel files to images
- Flexibility for exploration
- Focus on volume over usability

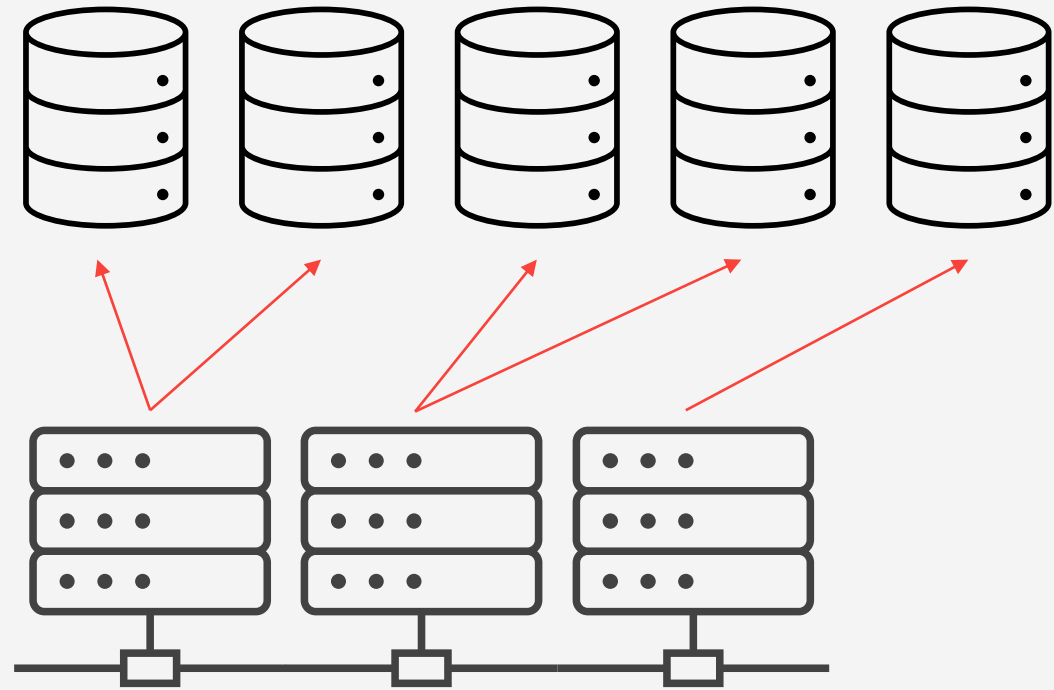
Computation Scaling

- 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



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...

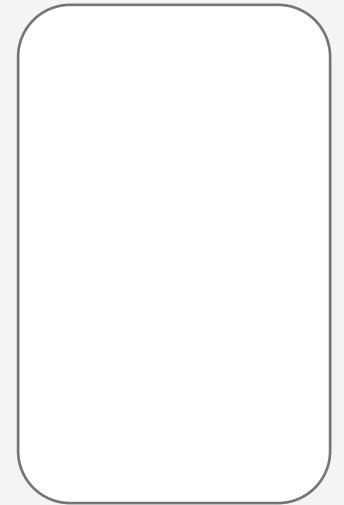
Master



Worker 1



Worker 2

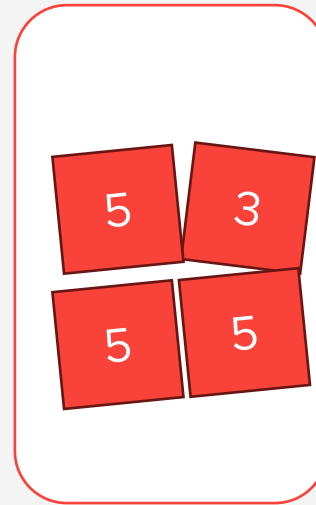


Compute

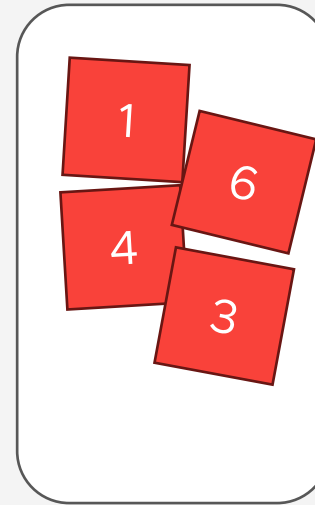
Distributed Computing!

Storage

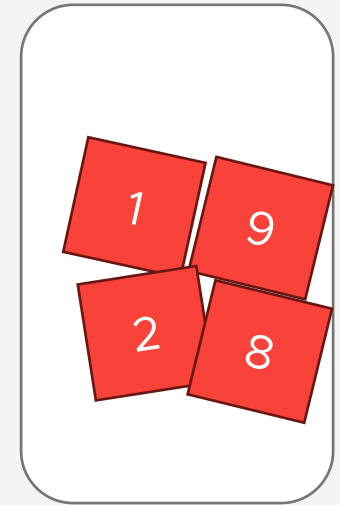
Master



Worker 1

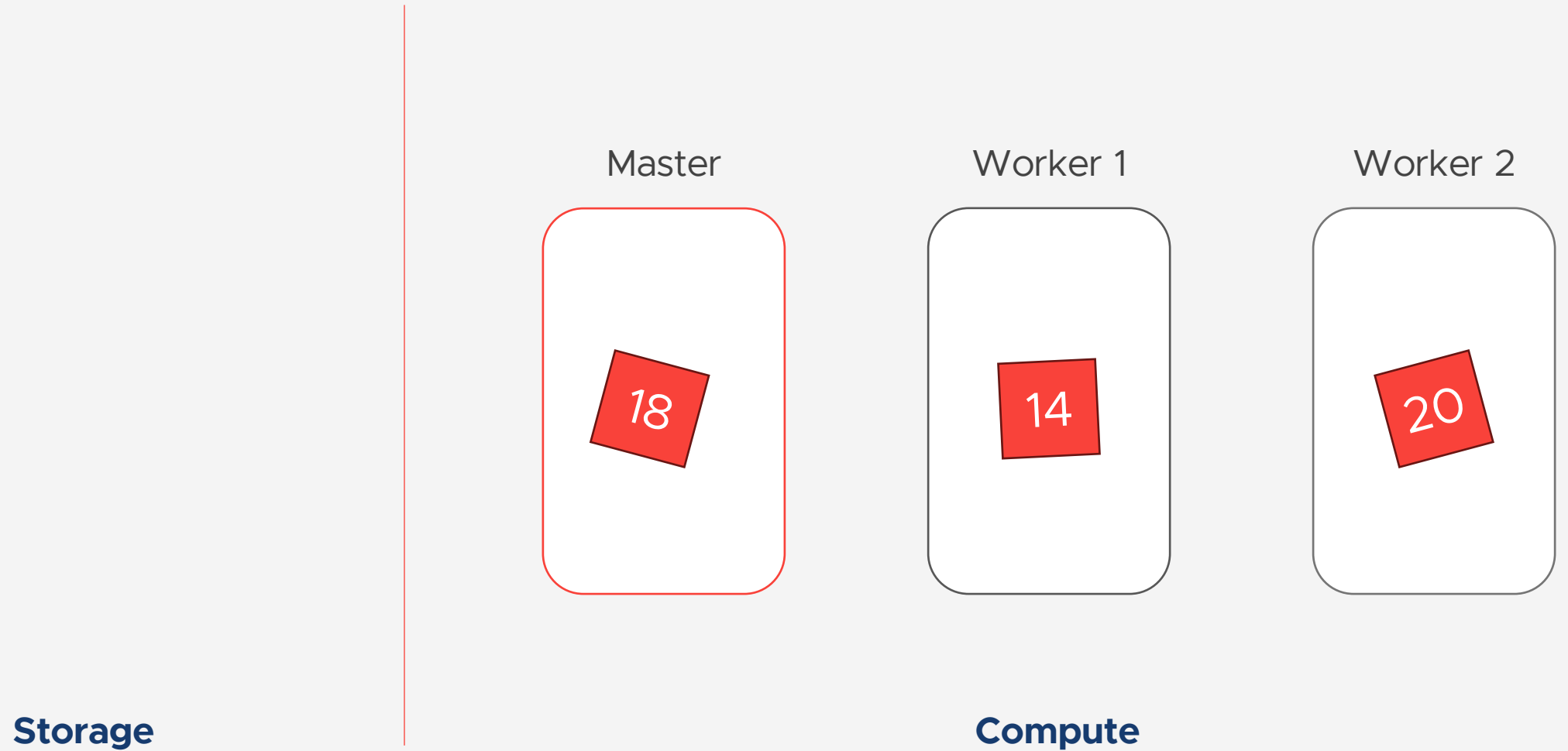


Worker 2

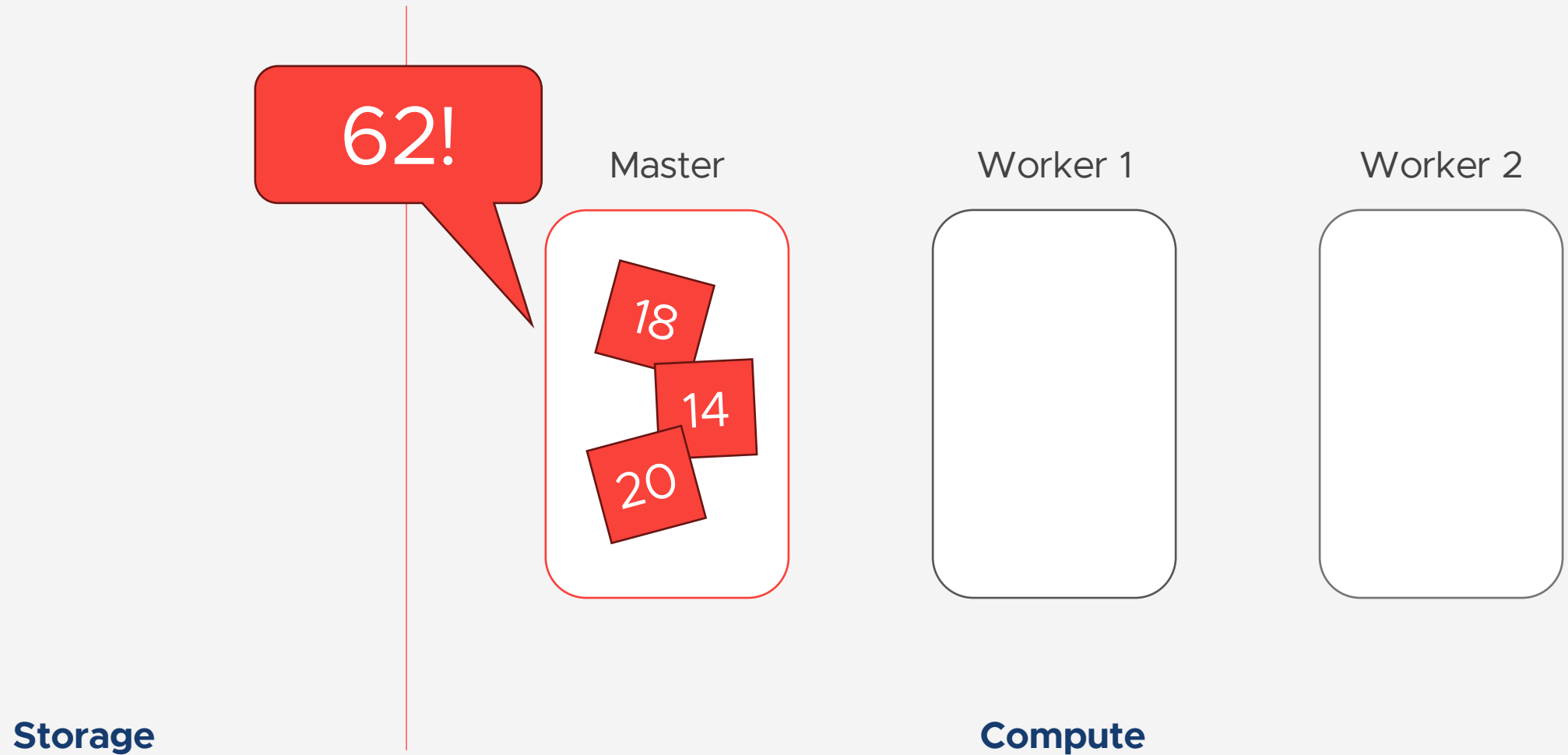


Compute

Distributed Computing!



Distributed Computing!

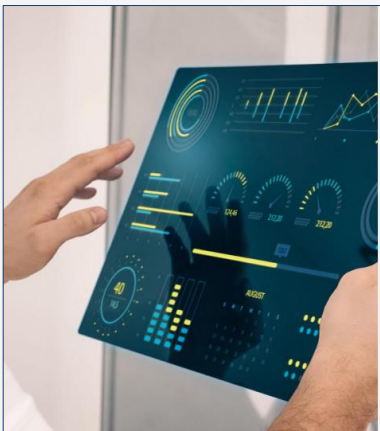


* That is an exclamation mark, not a factorial sign.




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!



Database

- For data collection
- Silo-ed for specific departments or function
- Mostly transactional
- Fast retrieval, fast updates
- **Online Transactional Processing (OLTP)**



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
- **Online Analytical Processing (OLAP)**



Data Lake

- Giant Reservoir of data in any forms, including unprocessed format and unstructured data.
- Can be literally anything from Excel files to images
- Flexibility for exploration
- Focus on volume over usability



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!



Medallion Layers of Data Lake House

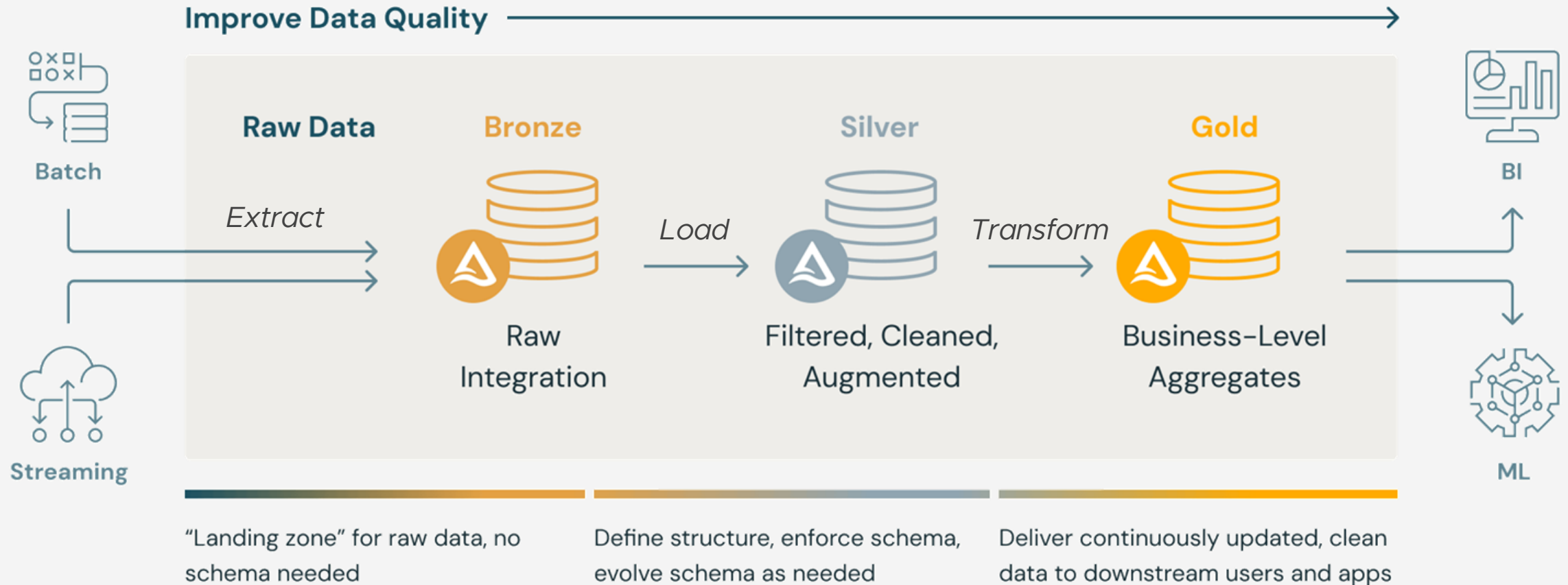
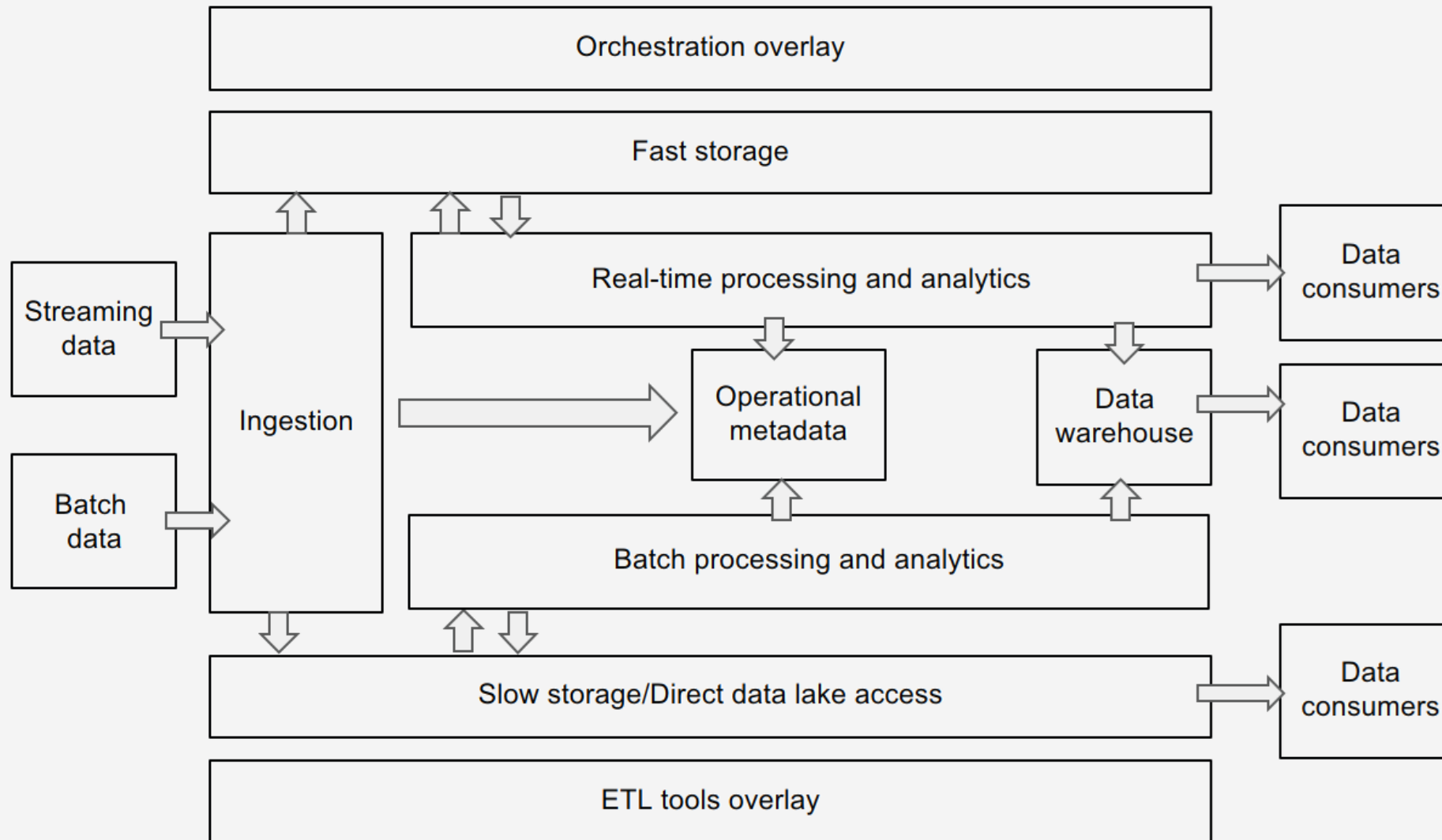


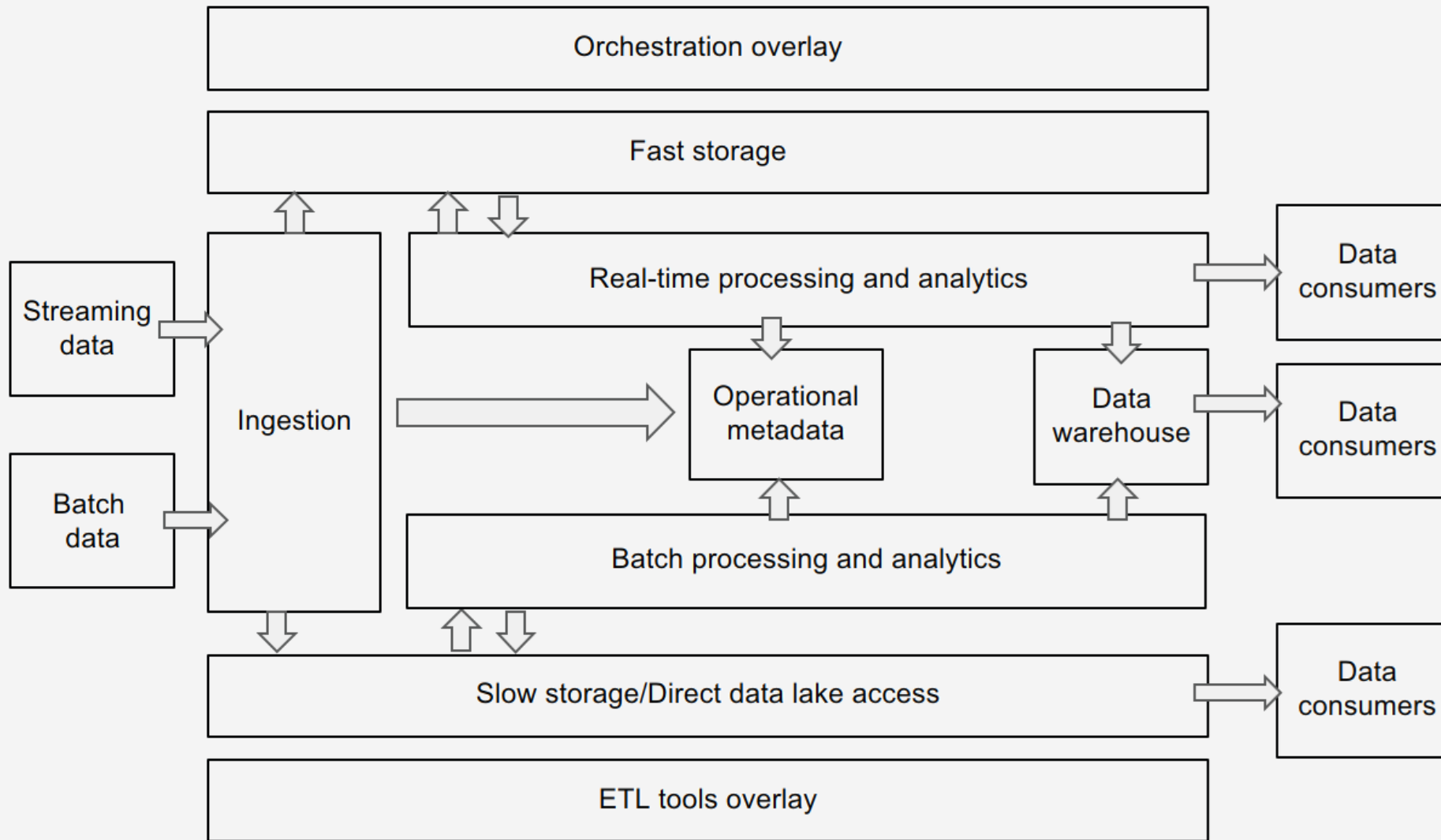
Image courtesy: Databricks



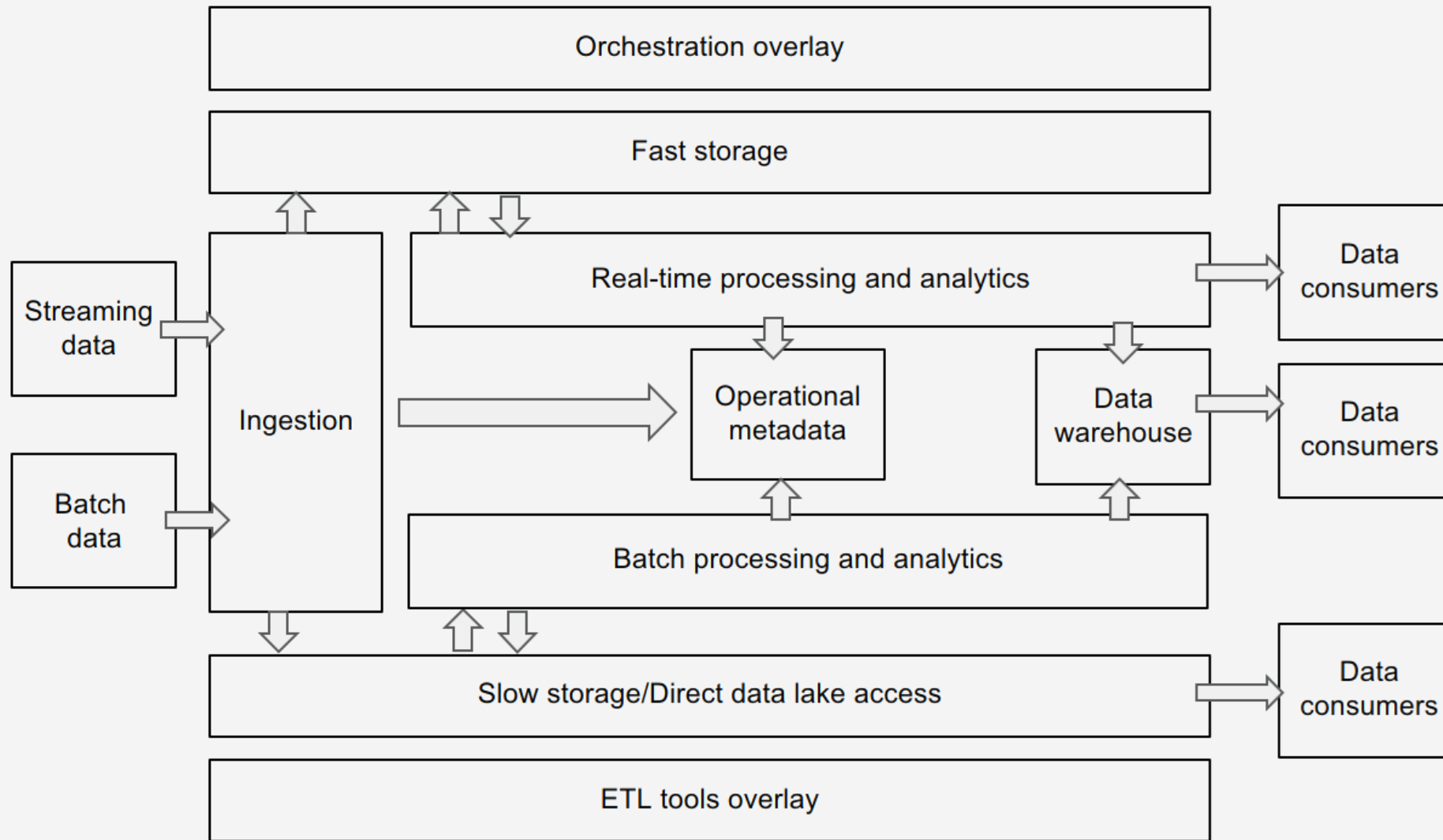
Components of Data Pipelines



Components of Data Pipelines (continued 1)



Components of Data Pipelines (continued 2)



**Massive computation
= Massive computers needed**

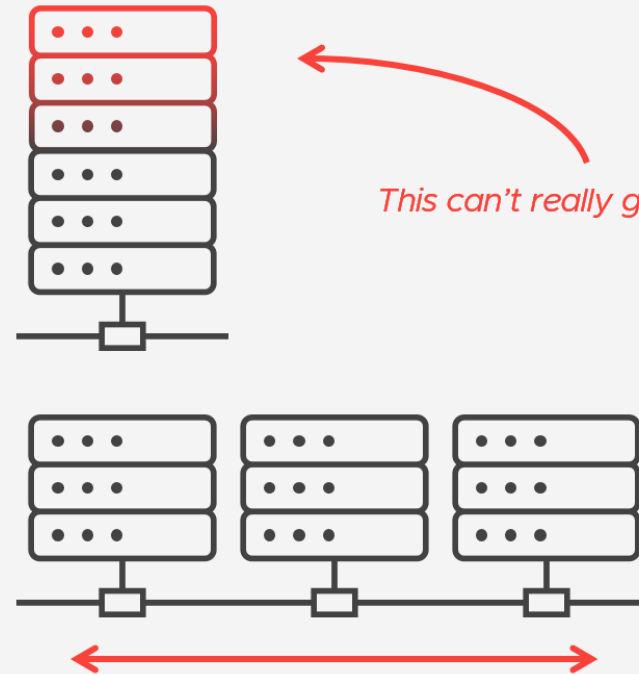
A large, fluffy white cloud is centered in the frame, set against a solid dark blue background. The cloud has a soft, billowy texture with varying shades of white and light grey. Overlaid on the center of the cloud is the text 'Cloud Data Platforms' in a large, bold, white sans-serif font. Below this,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

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.

Don't reinvent this

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

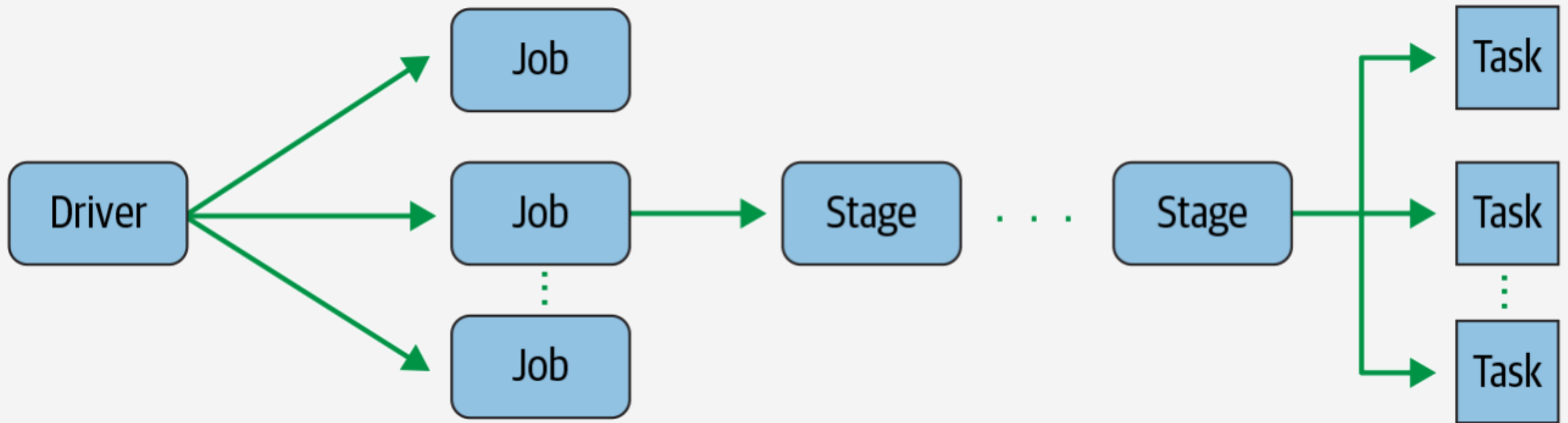
SQL

Python

Scala

Java

Spark Execution



Databricks

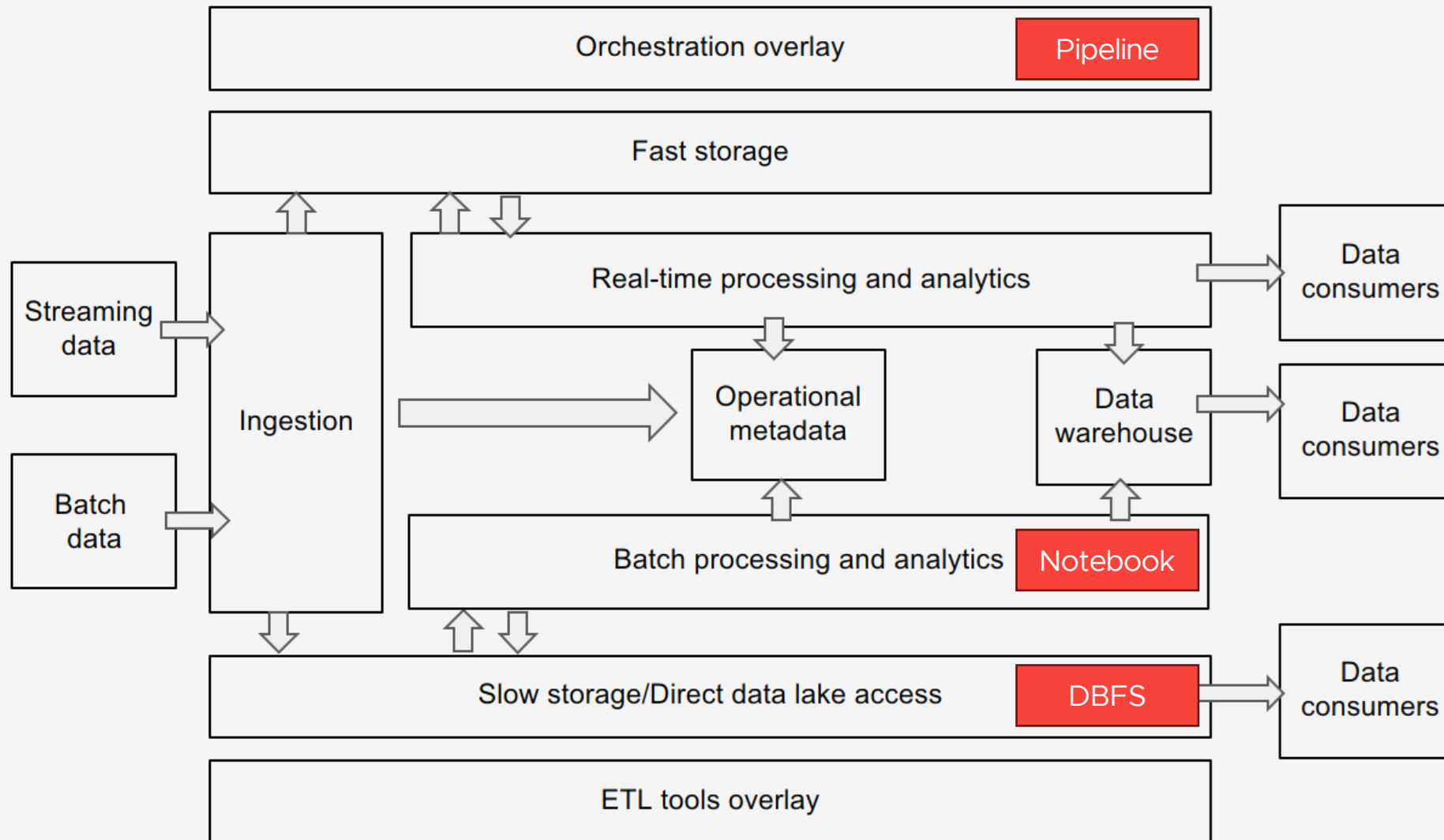
- Spark on the cloud
- Less hassle managing Spark cluster
- Provides useful features rather than computing engine
 - GUI for development
 - Data catalog
 - Orchestration*

** non-free plan only*



databricks

Data Pipelines on Databricks

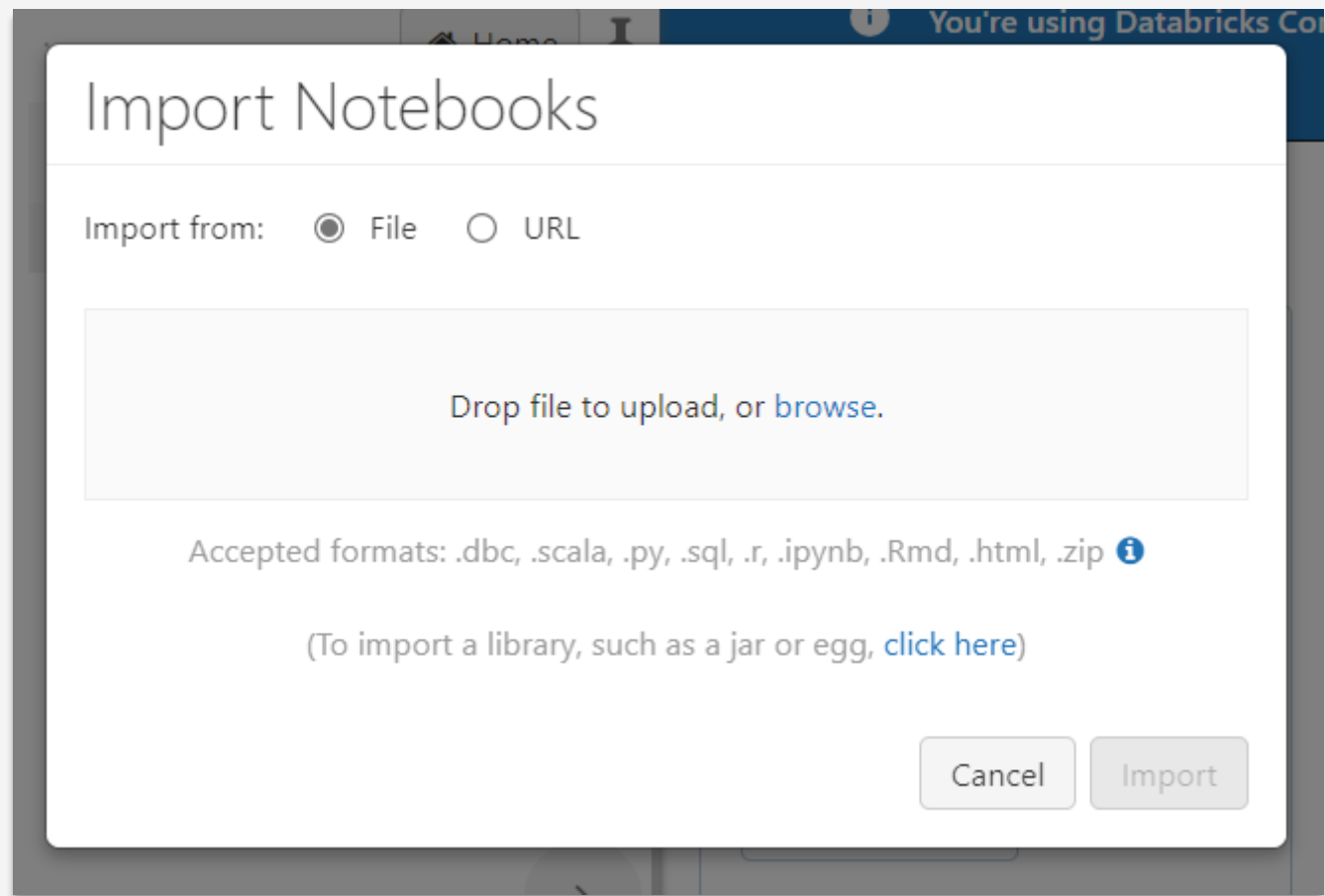


./ADASTRA

Databricks Lab



The screenshot displays the Databricks Workspace interface in a web browser. The browser's address bar shows the URL `https://community.cloud.databricks.com/?o=5292693329990630#`. The Databricks logo is visible in the top left corner. A sidebar on the left contains navigation links: **New**, **Workspace** (selected), **Recents**, **Search**, **Catalog**, **Workflows**, **Compute**, **Machine Learning**, and **Experiments**. The main area is titled **Workspace** and shows a breadcrumb path: **Home** > **sirakorn.lamyai@adastragrp.com** > **ATH Workshop 2024**. A context menu is open over the **ATH Workshop 2024** folder, showing options: **Create**, **Import** (highlighted), and **Permissions**.



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