ADASTRA

Introduction to Data Engineering on Databricks

Adastra Thailand Campus on-tour program

Stamford International University
24 May 2024

Meet our team

Wiparat P.

Head of Operations

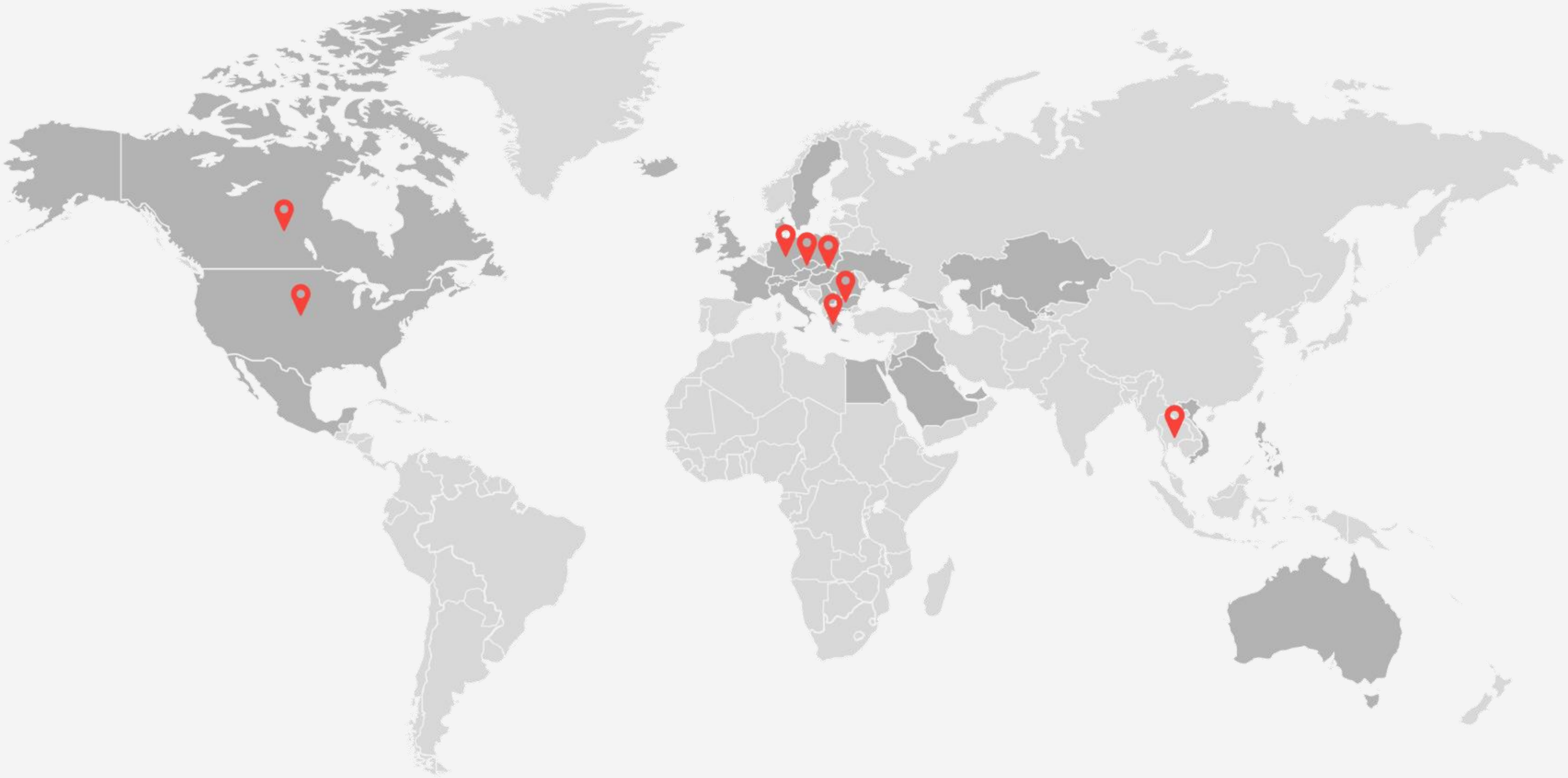
Sirakorn L.

Practice Lead – AWS, Data Engineering, and Development

Thanyaporn S.

Recruitment Lead

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

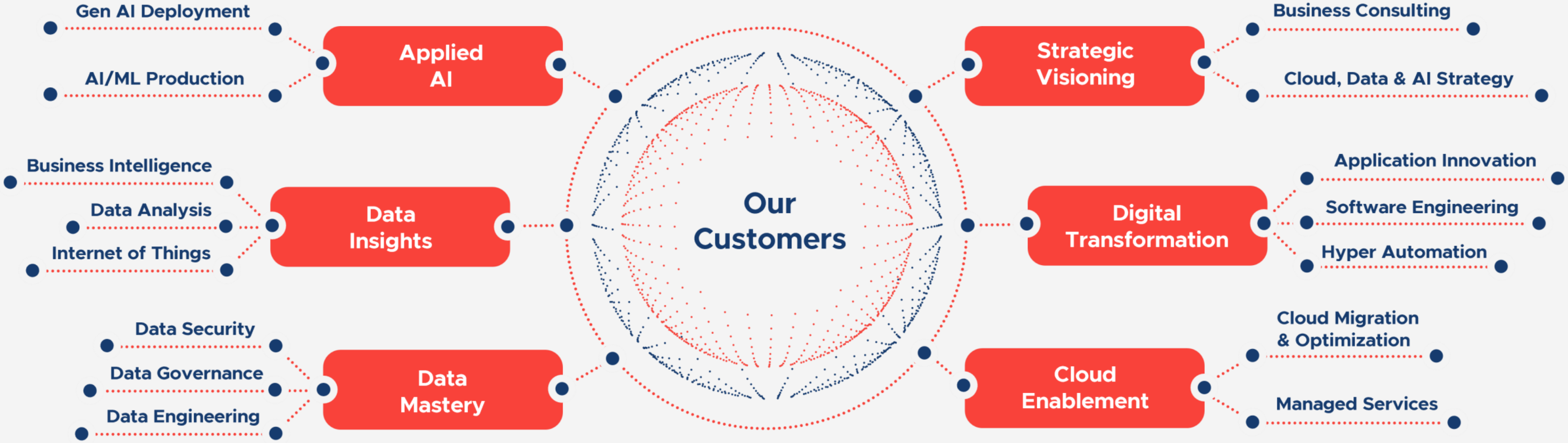
THAILAND
Bangkok
Chiang Mai

■ 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

Microsoft

aws

Google Cloud

ataccama

alteryx

databricks

Informatica

AUTOMATION ANYWHERE

snowflake

bloomreach

CLOUDERA

Attachtap

IBM

SAP

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)**

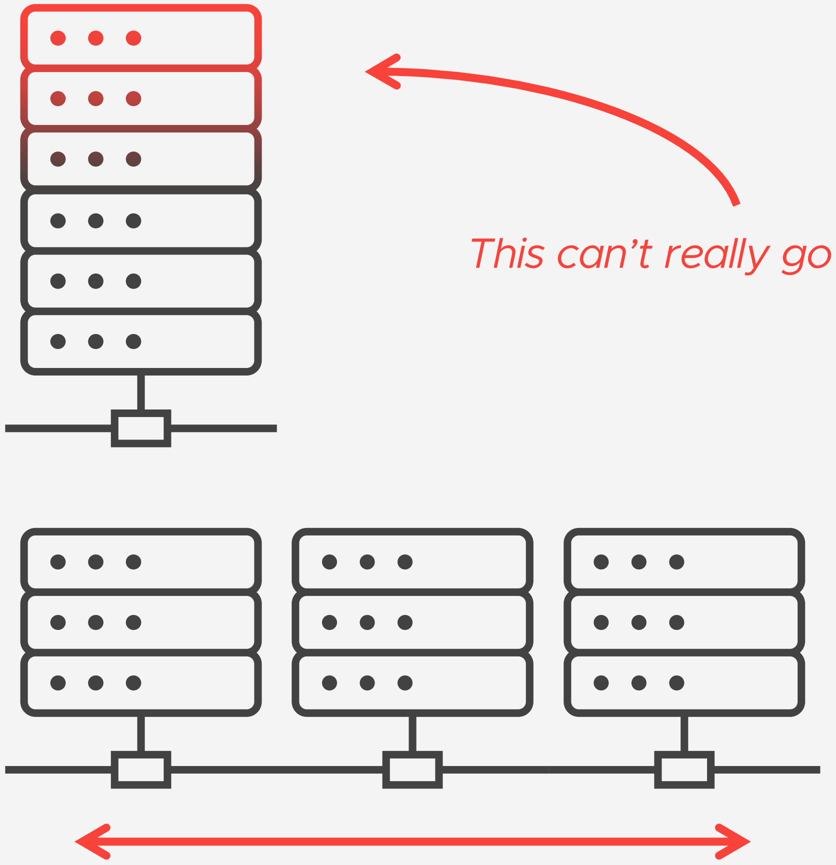
How can we store even more types of data?

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

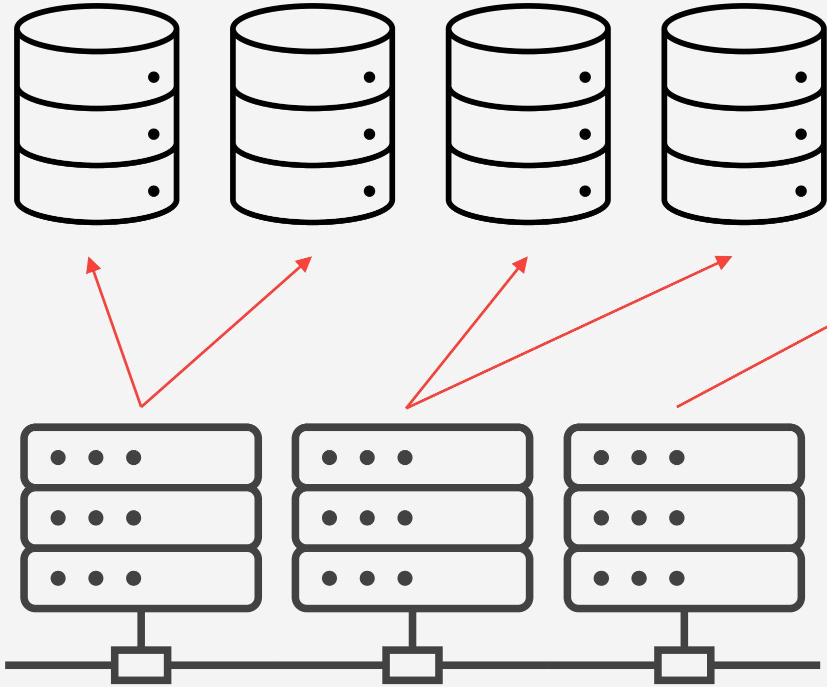


This can't really go high.

This requires distributing and "talking" between devices.

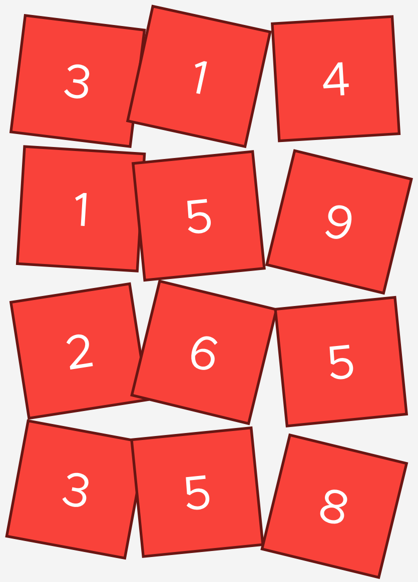
Storage-Compute Decoupling

- 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.



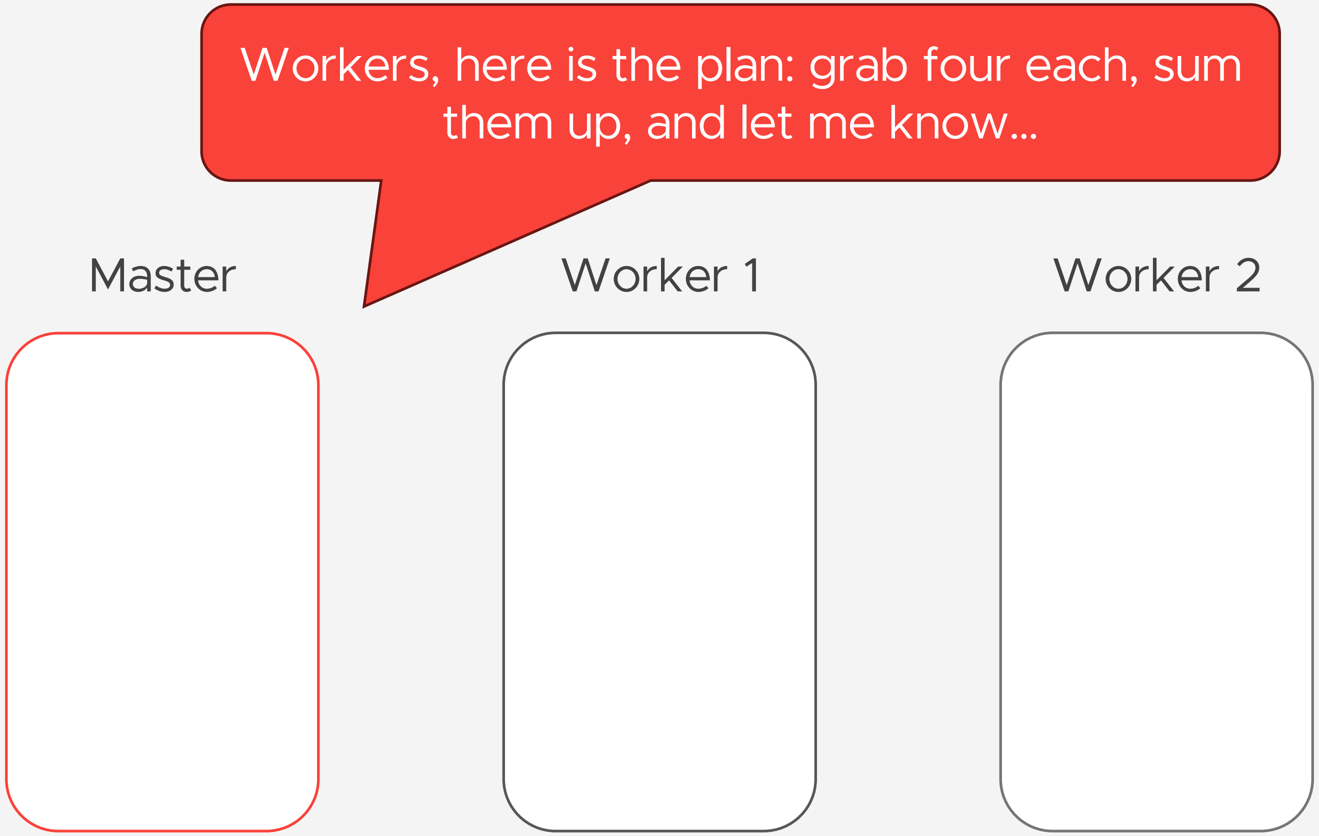
Distributed Computing!

Calculate summation of these numbers



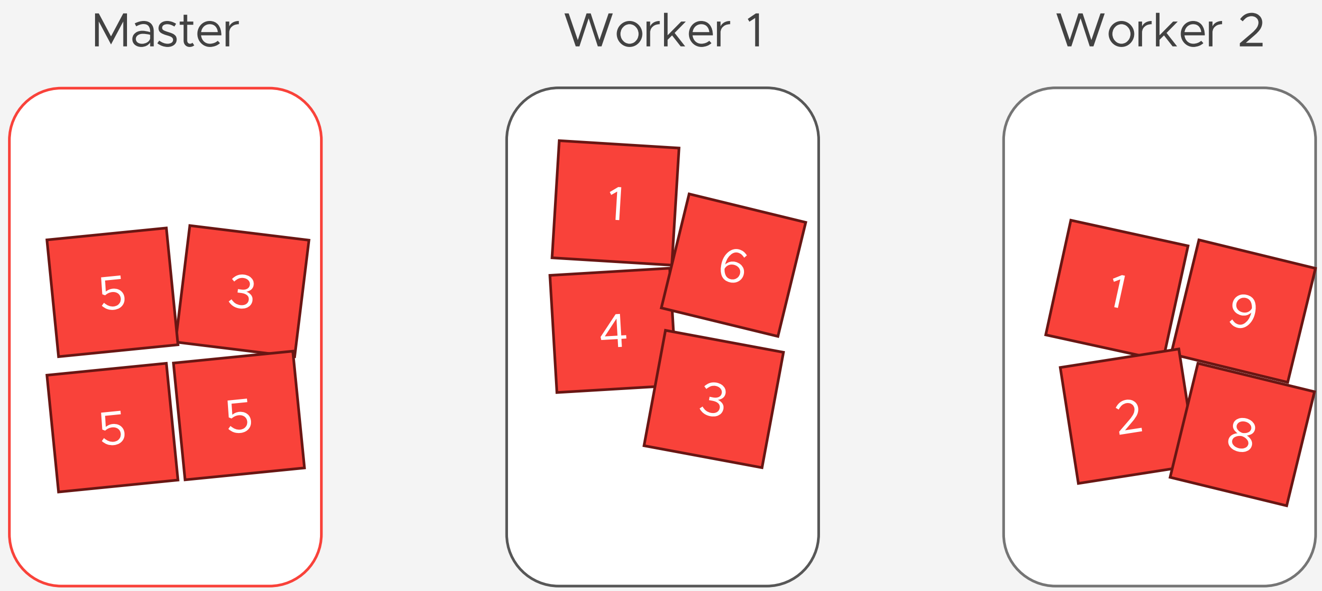
Storage

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



Compute

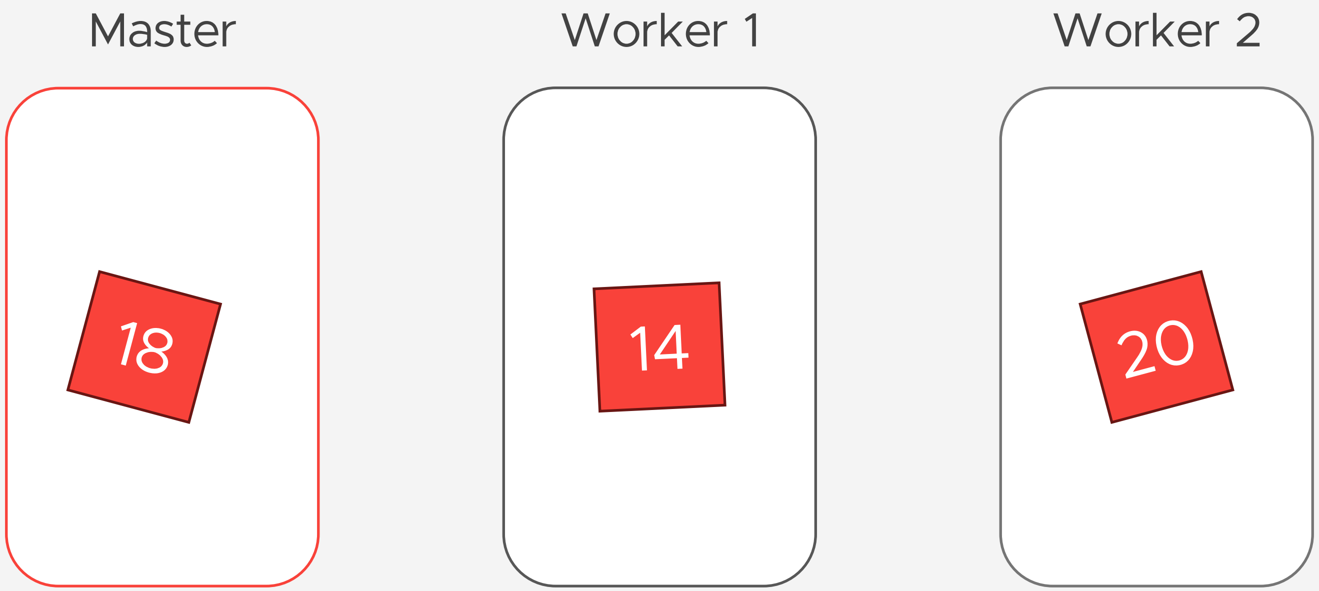
Distributed Computing!



Storage

Compute

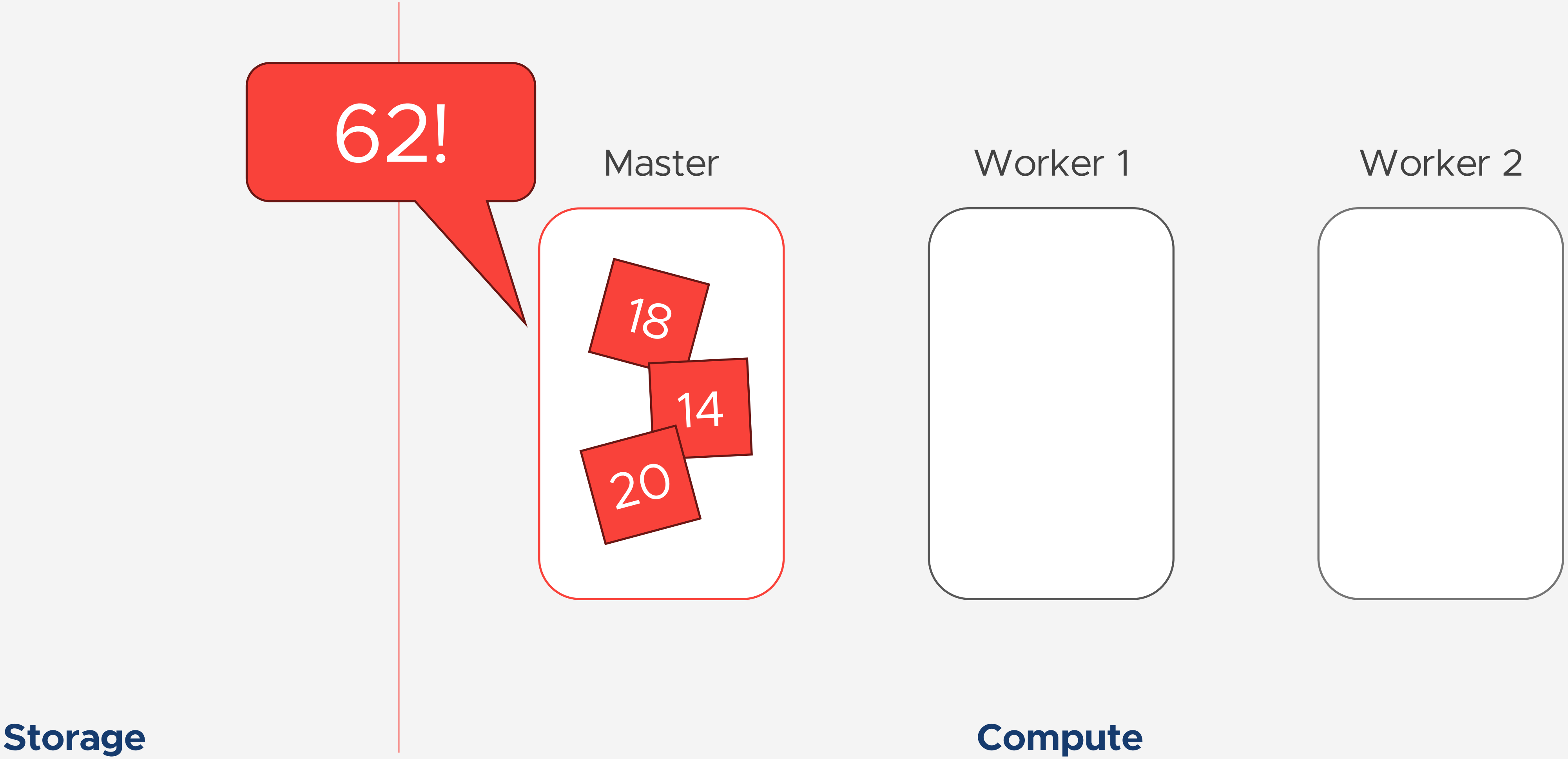
Distributed Computing!



Storage

Compute

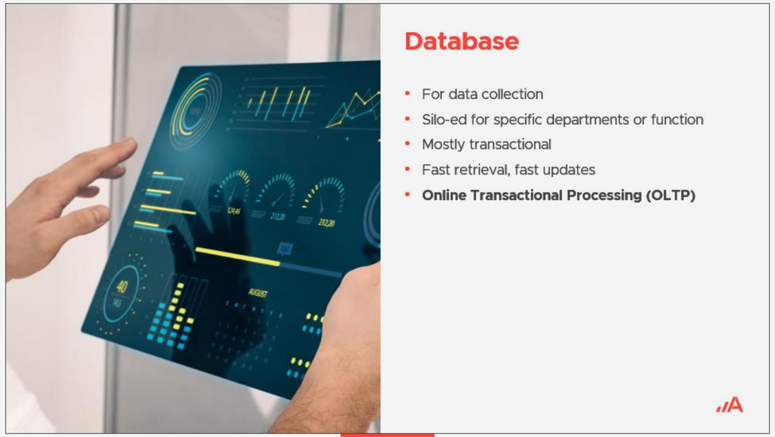
Distributed Computing!



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

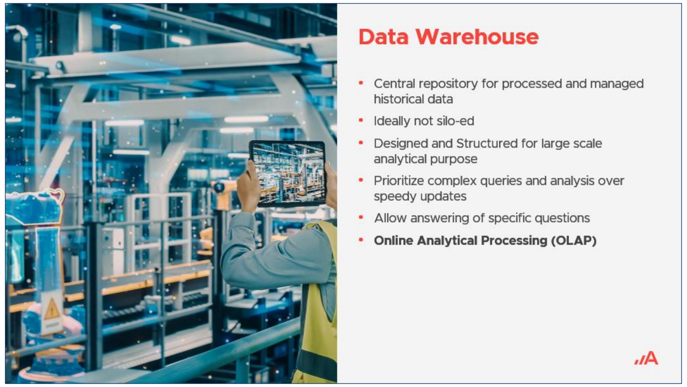
Data Lake House

- 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!



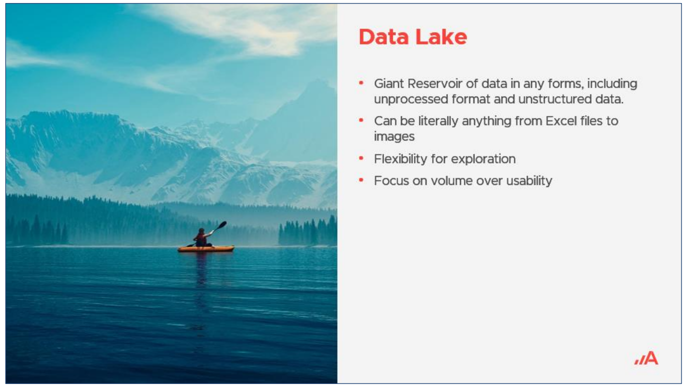
Database

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



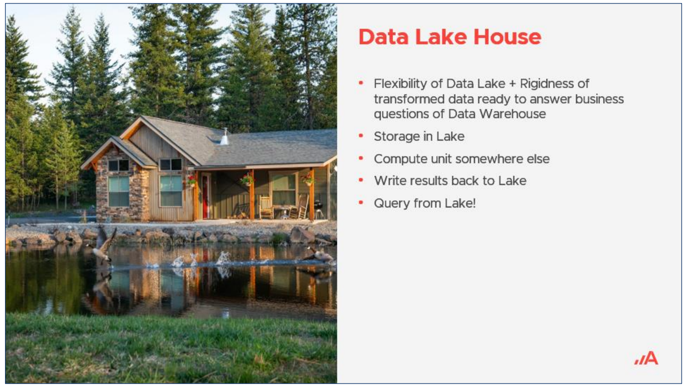
Data Warehouse

- 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

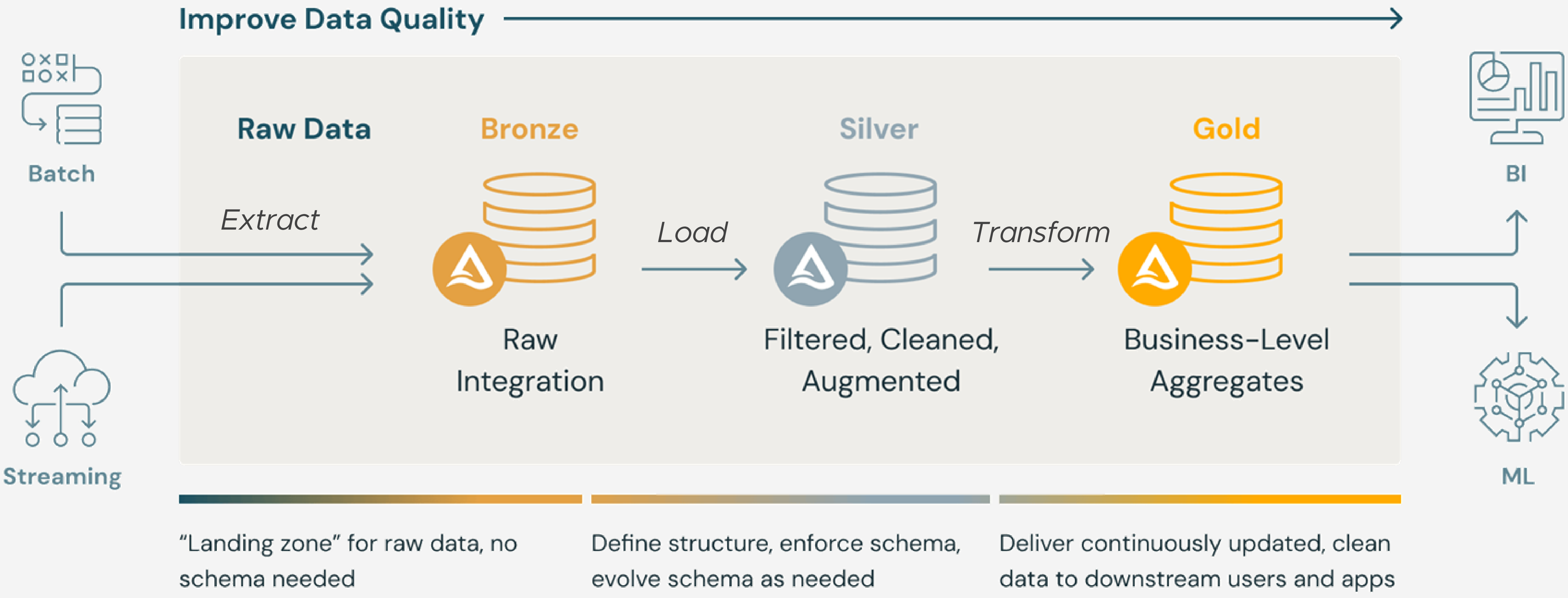
- Giant Reservoir of data in any forms, including unprocessed format and unstructured data.
- Can be literally anything from Excel files to images
- Flexibility for exploration
- Focus on volume over usability



Data Lake House

- Flexibility of Data Lake + 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!

Medallion Layers of Data Lake House



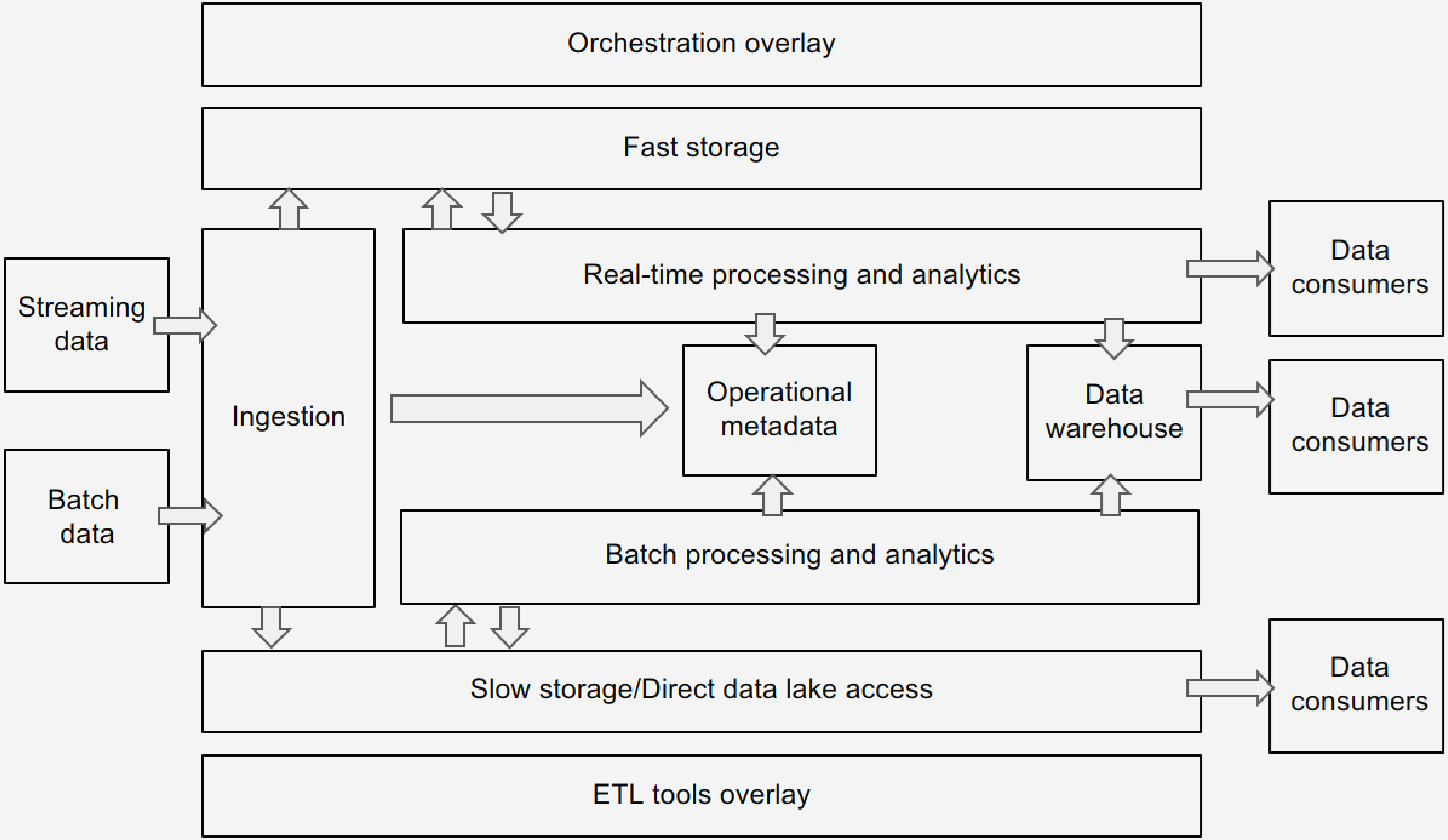
"Landing zone" for raw data, no schema needed

Define structure, enforce schema, evolve schema as needed

Deliver continuously updated, clean data to downstream users and apps

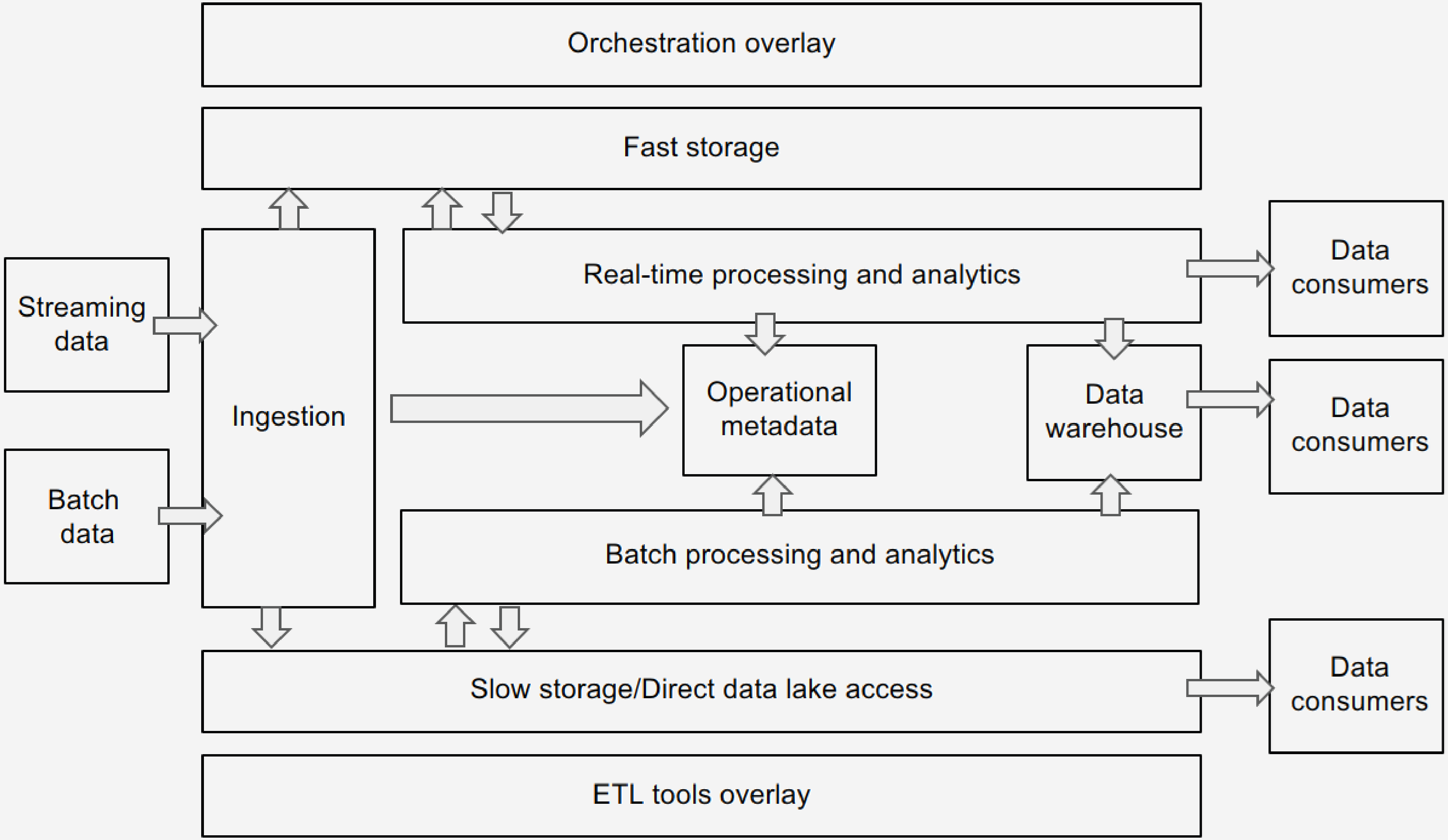
Image courtesy: Databricks

Components of Data Pipelines



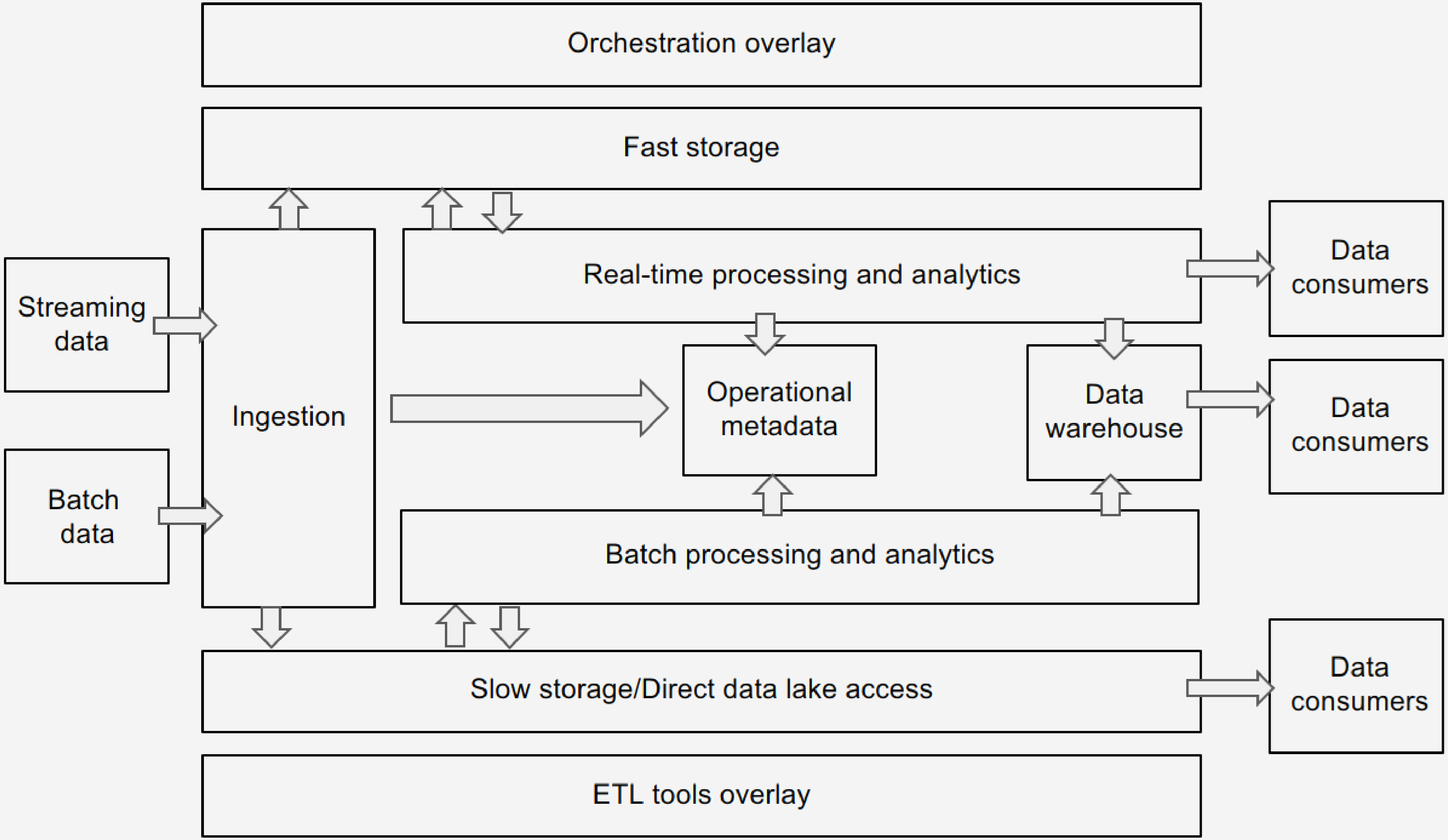
D. Zburivsky and L. Partner, *Designing Cloud Data Platforms*. Shelter Island, NY: Manning Publications Co., 2021.

Components of Data Pipelines (continued 1)



D. Zburivsky and L. Partner, *Designing Cloud Data Platforms*. Shelter Island, NY: Manning Publications Co., 2021.

Components of Data Pipelines (continued 2)

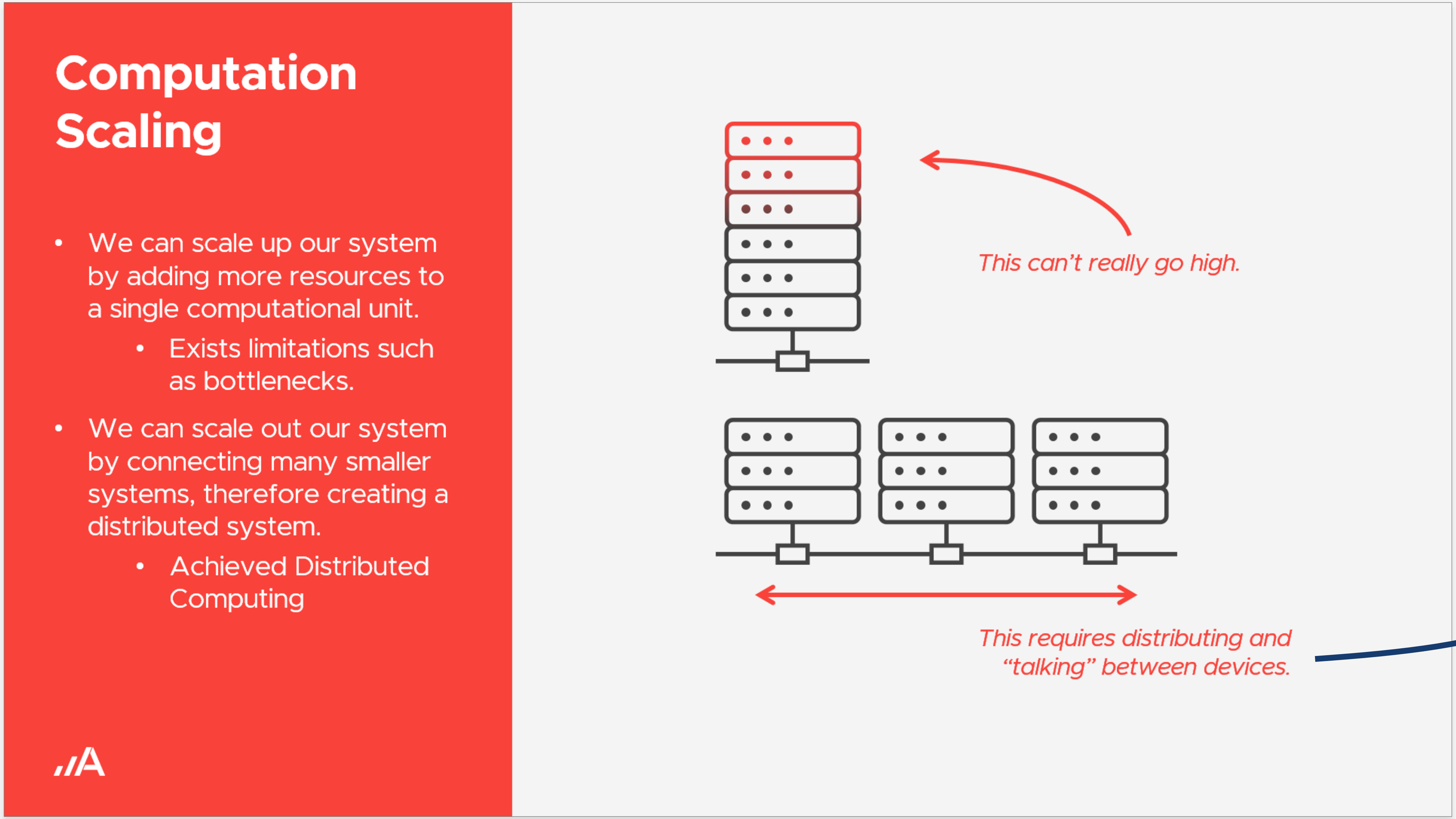


D. Zburivsky and L. Partner, *Designing Cloud Data Platforms*. Shelter Island, NY: Manning Publications Co., 2021.

Massive computation
= Massive computers needed

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.

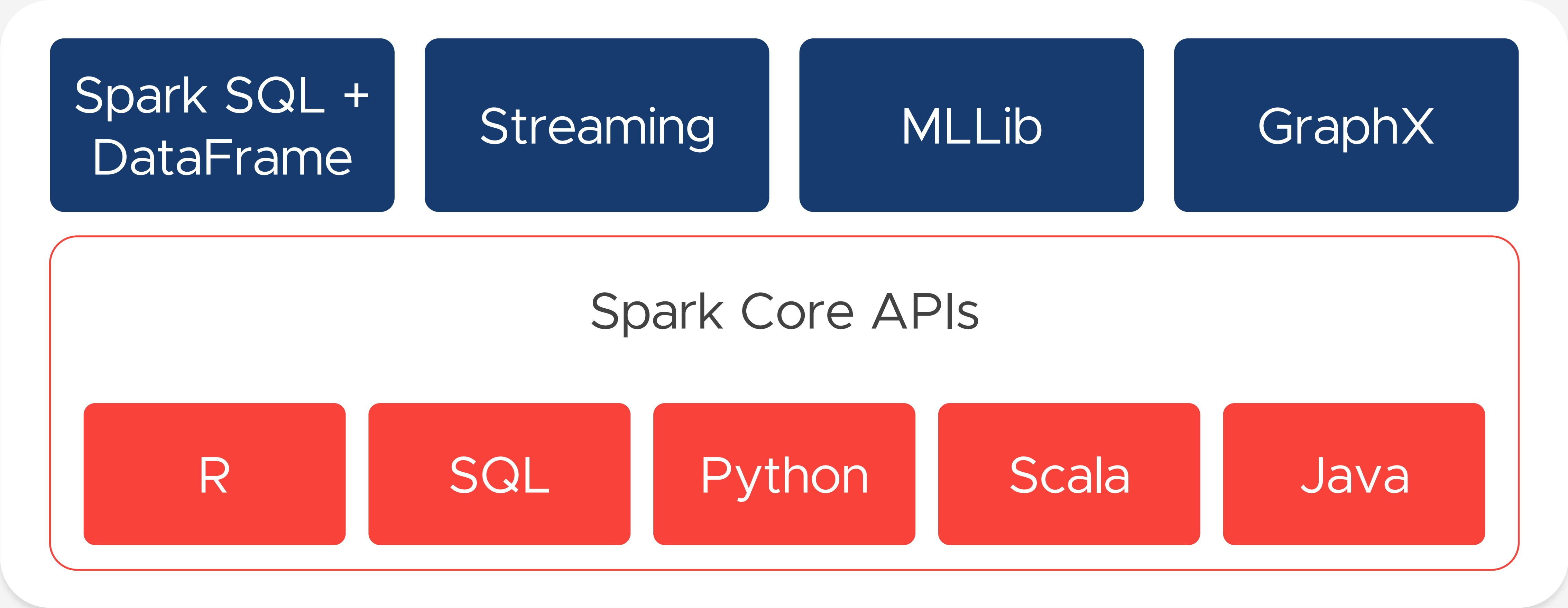
This requires distributing and "talking" between devices.

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

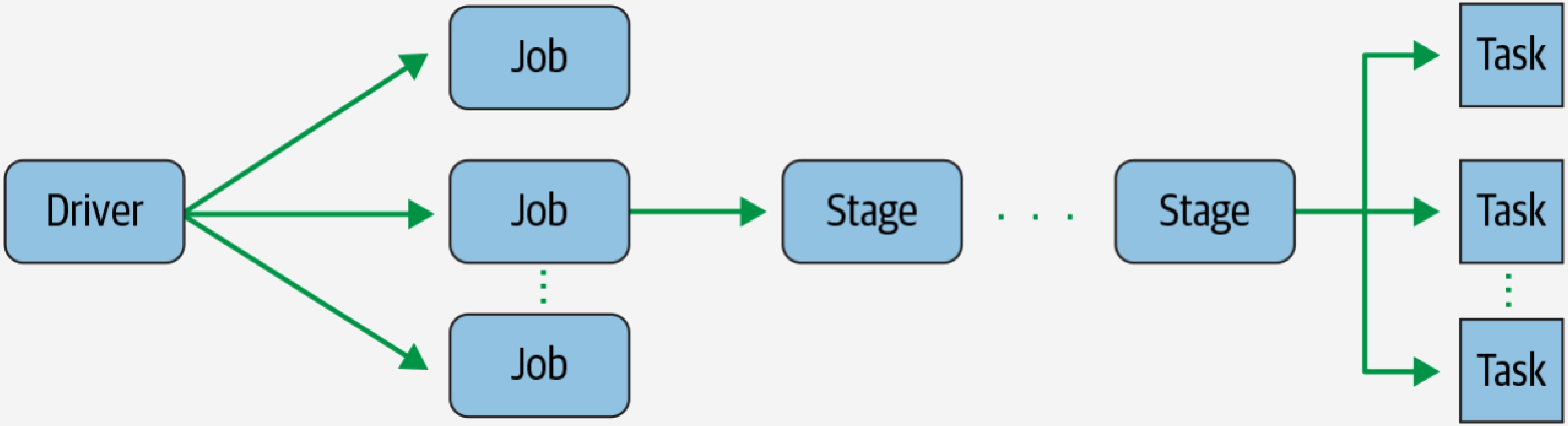
Python

Scala

Java

Adapted from Damji, J., Lee, D., Wenig, B., & Das, T. (2020). *Learning Spark, 2nd Edition*. California: O'Reilly Media, Inc., 2020

Spark Execution



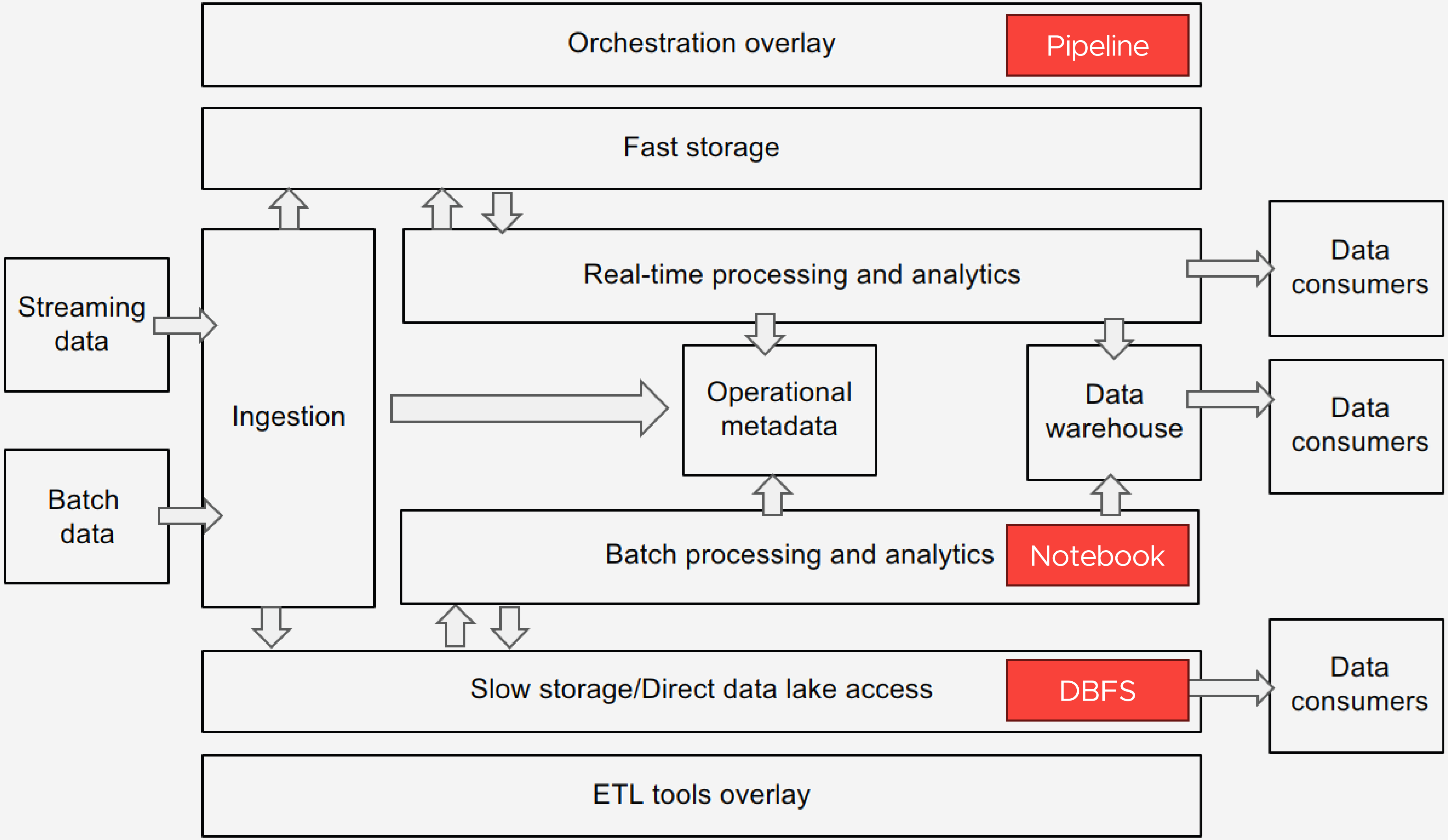
Replicated from Damji, J., Lee, D., Wenig, B., & Das, T. (2020). *Learning Spark, 2nd Edition*. California: O'Reilly Media, Inc., 2020

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*

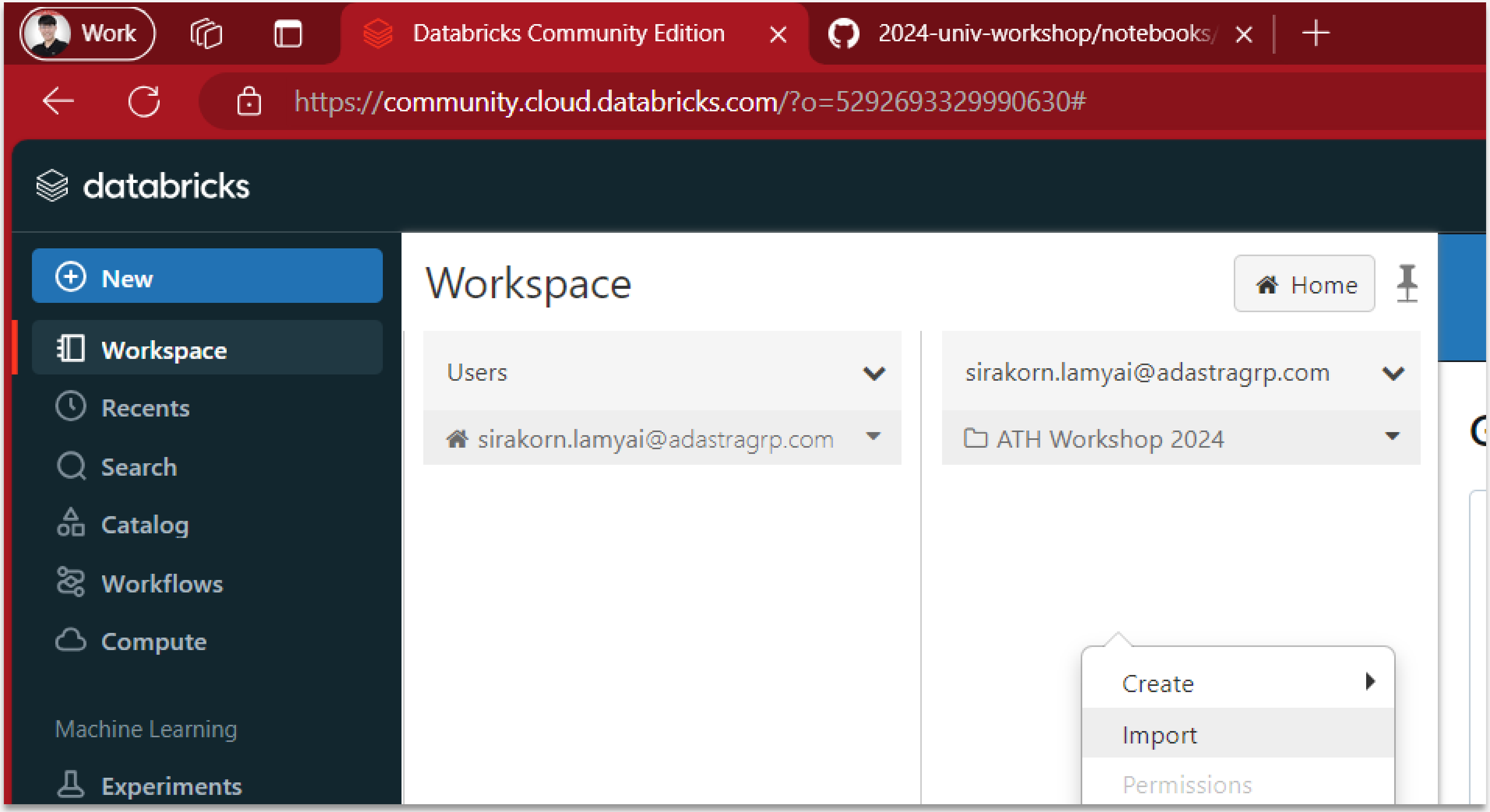
Data Pipelines on Databricks

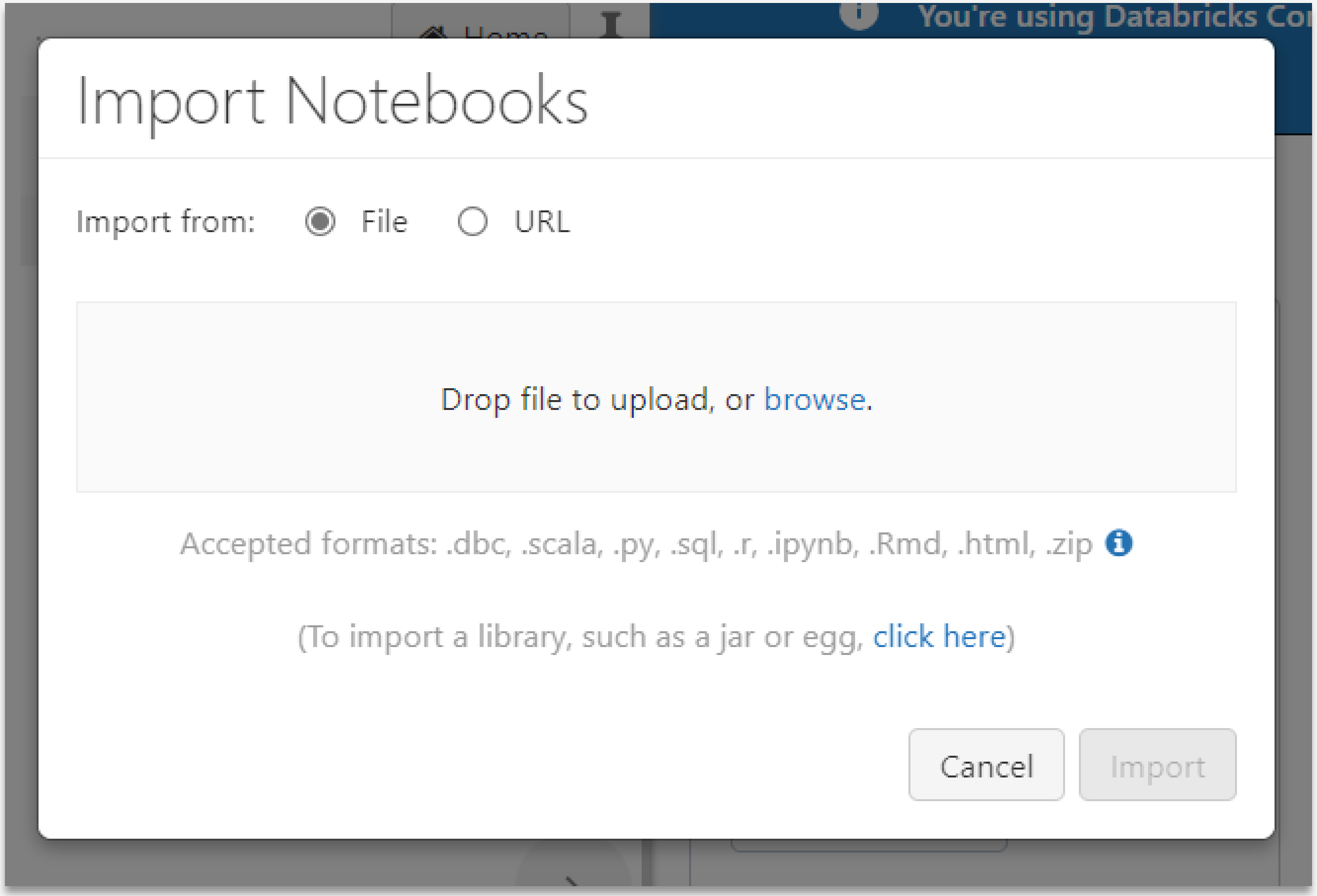


D. Zburivsky and L. Partner, *Designing Cloud Data Platforms*. Shelter Island, NY: Manning Publications Co., 2021.

ADASTRA

Databricks Lab


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
Home
Users
sirakorn.lamyai@adastragrp.com
sirakorn.lamyai@adastragrp.com
ATH Workshop 2024
Create
Import
Permissions



Import Notebooks

Import from: ● File ○ URL

Drop file to upload, or browse.

Accepted formats: .dbc, .scala, .py, .sql, .r, .ipynb, .Rmd, .html, .zip

(To import a library, such as a jar or egg, click here)

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