Data-Intensive Distributed Computing
CS 431 (Fall 2026)



UNIVERSITY OF WATERLOO

Rubber, Meet Road
(v1.00)

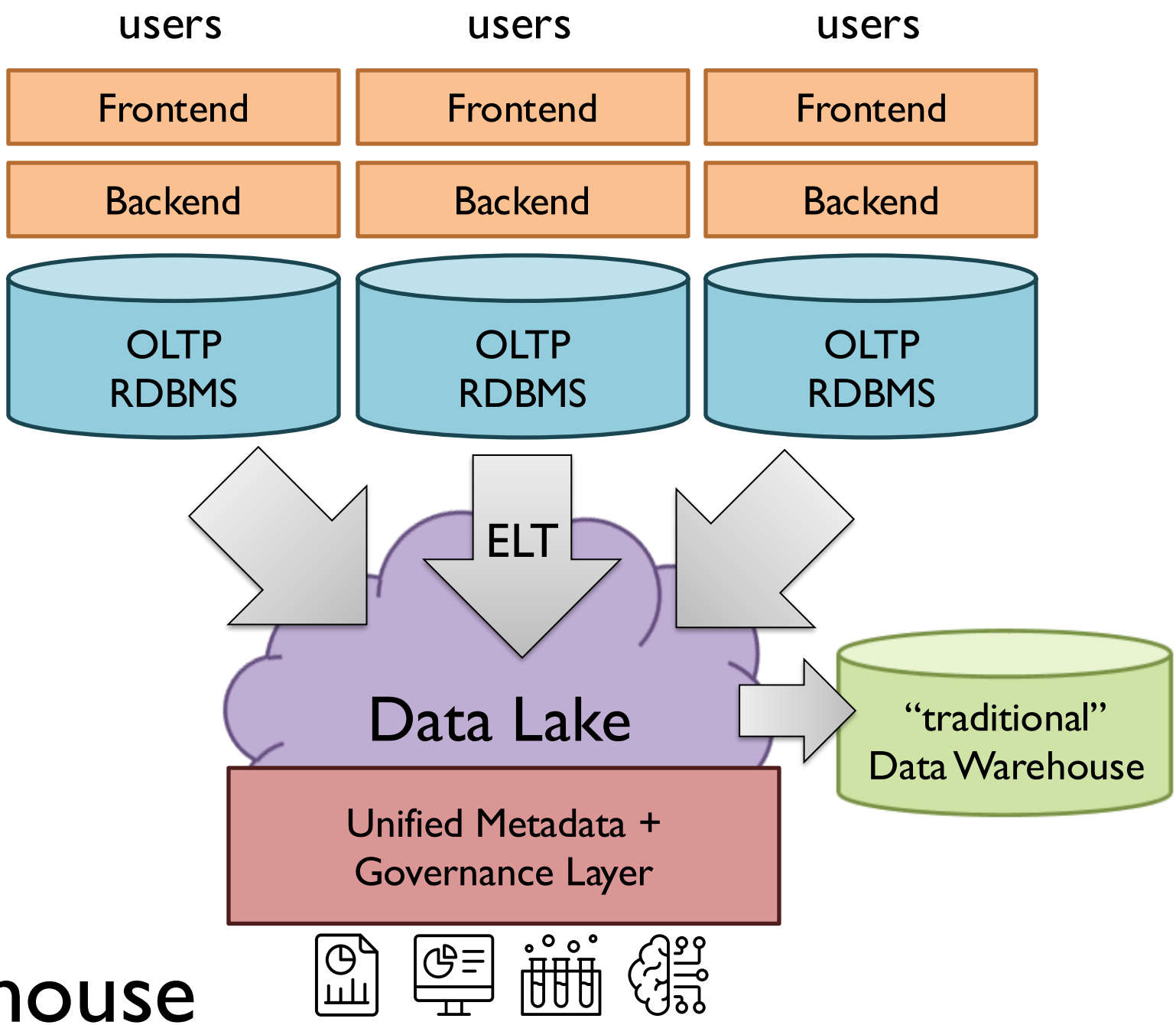
Week 7: October 20, 2026

Jimmy Lin
David R. Cheriton School of Computer Science
University of Waterloo

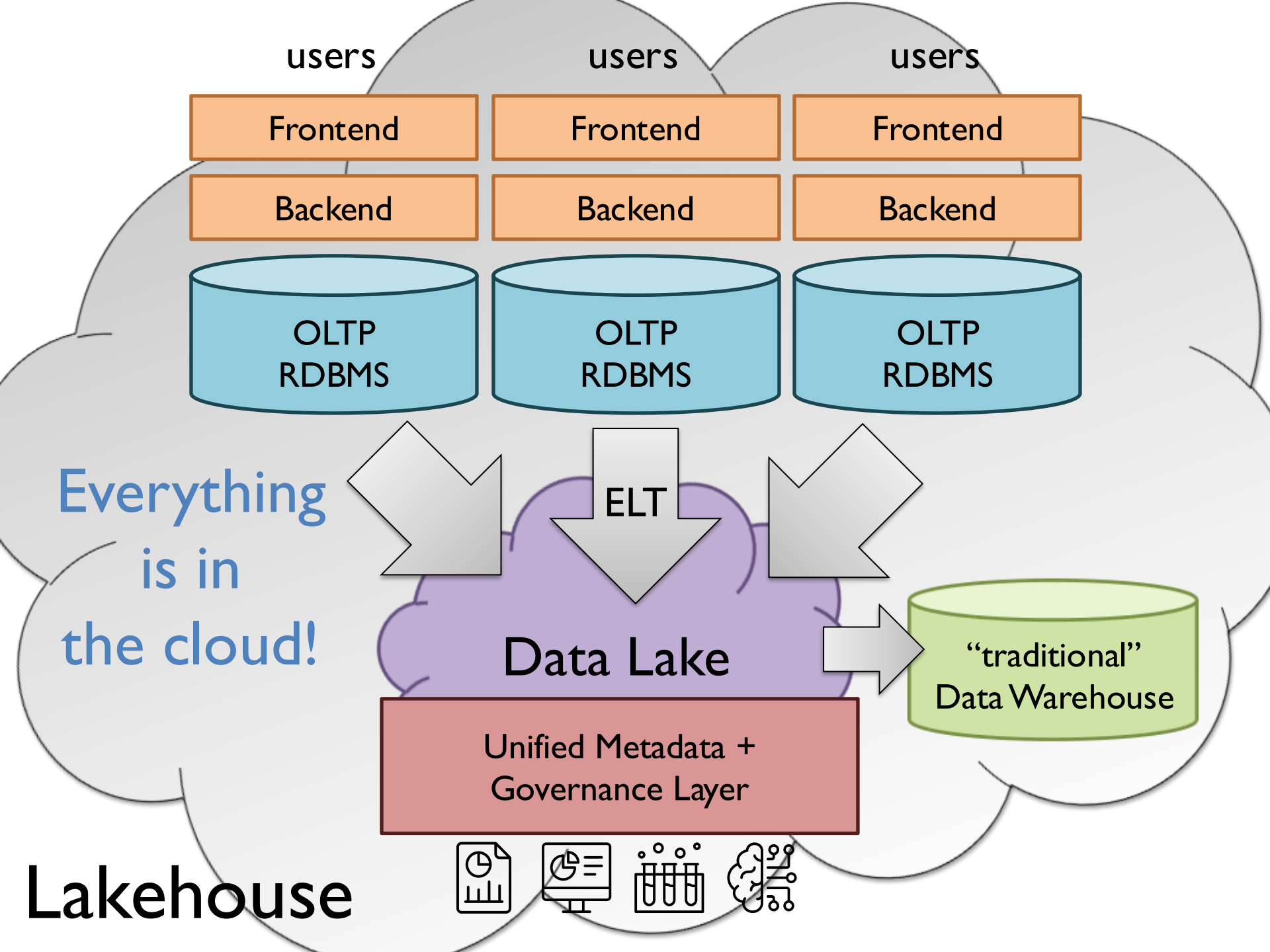
These slides are available at `https://lintool.github.io/cs431-2026f/`



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International
See https://creativecommons.org/licenses/by-nc-sa/4.0/ for details

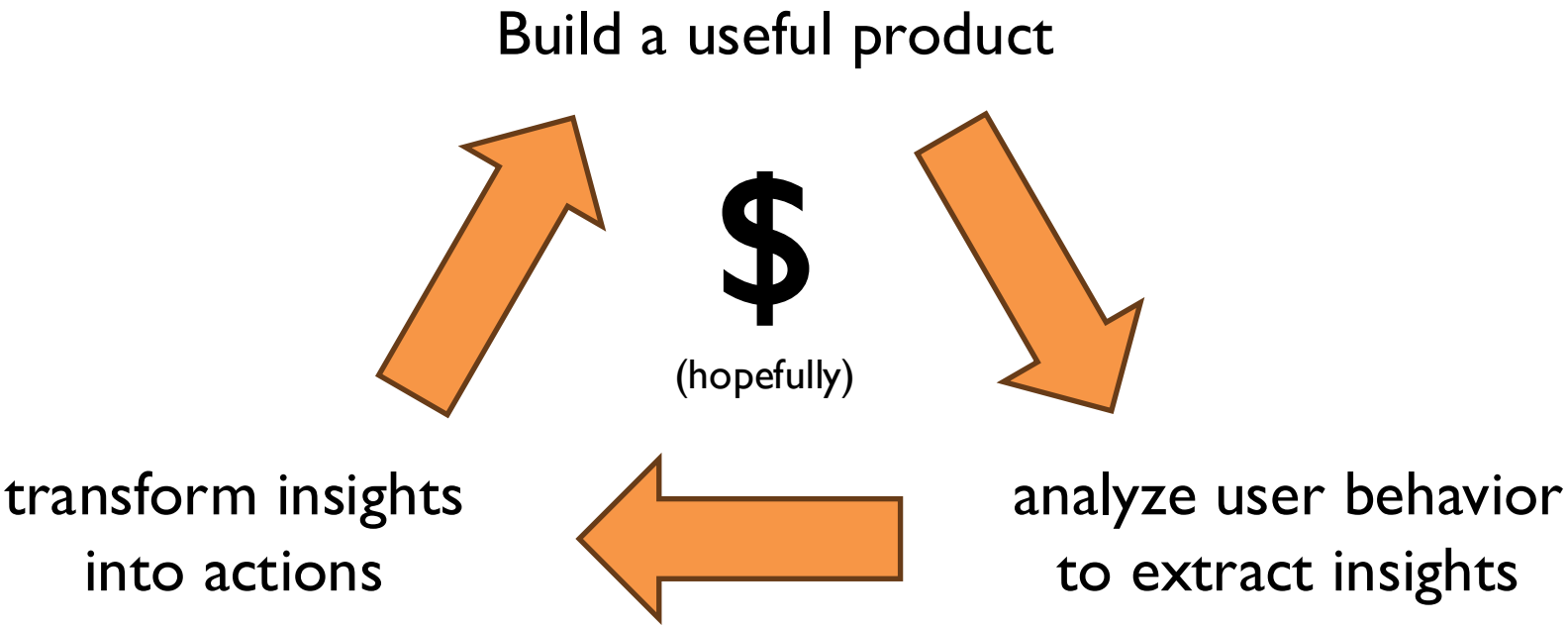

users
users
users
Frontend
Frontend
Frontend
Backend
Backend
Backend
OLTP RDBMS
OLTP RDBMS
OLTP RDBMS
ELT
Data Lake
"traditional" Data Warehouse
Unified Metadata + Governance Layer

Lakehouse

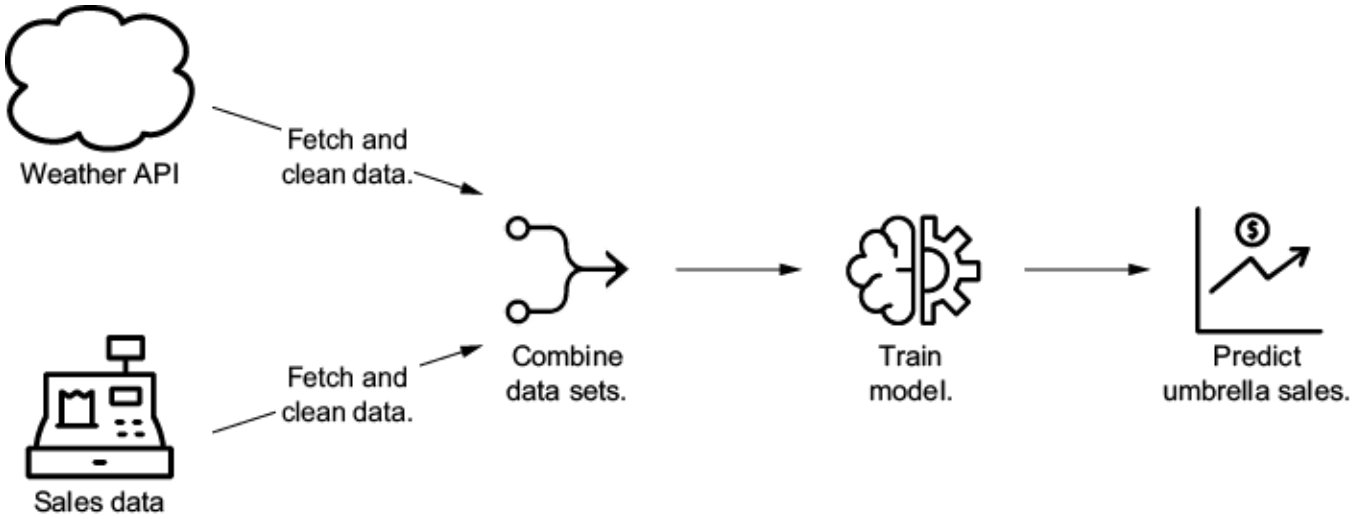

users
users
users
Frontend
Frontend
Frontend
Backend
Backend
Backend
OLTP RDBMS
OLTP RDBMS
OLTP RDBMS
ELT
Everything is in the cloud!
Data Lake
"traditional" Data Warehouse
Unified Metadata + Governance Layer
Lakehouse

The Data Flywheel

(a virtuous cycle)



Google. Facebook. Twitter. Amazon. Uber.

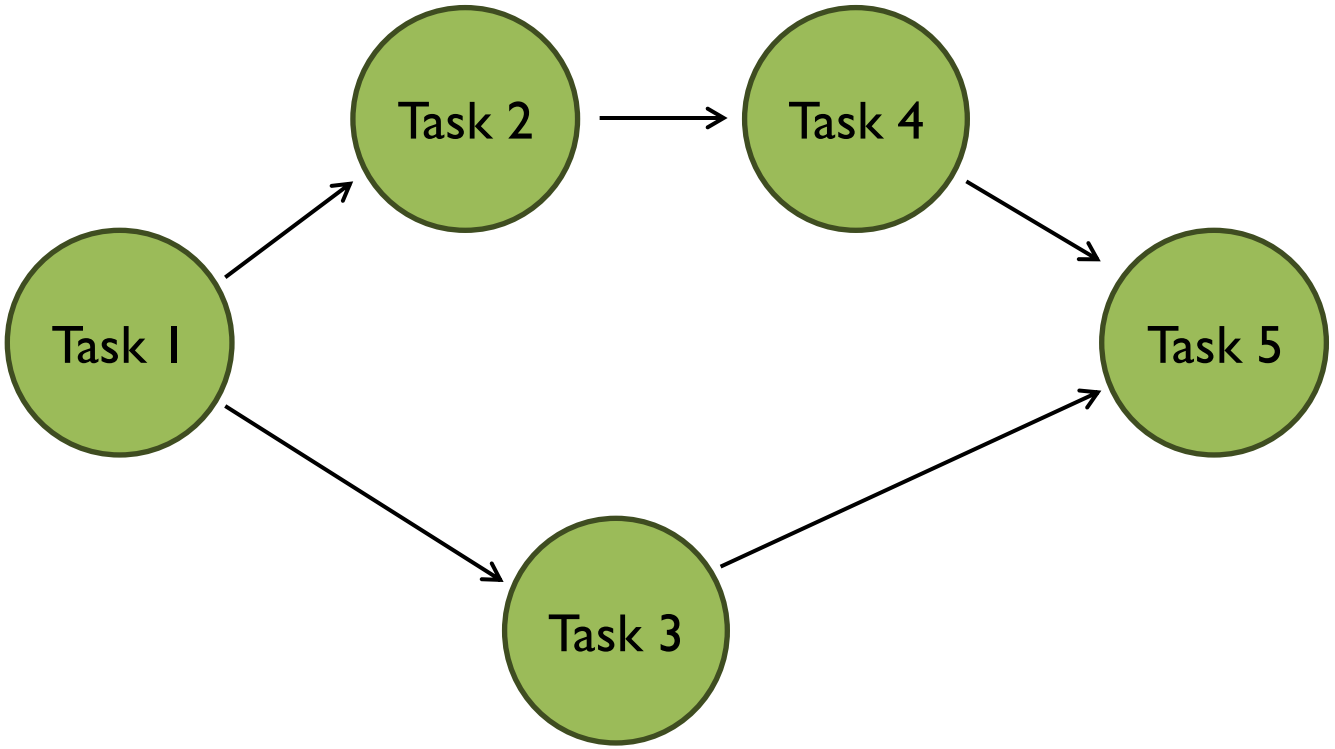


From Cabral et al. Data Pipelines with Apache Airflow (2026)



From Cabral et al. Data Pipelines with Apache Airflow (2026)

It's just a DAG



Heterogeneous Tasks

Input	**Processing**	**Output**
CSV	Spark	CSV
JSON(L)	Flink	JSON(L)
Parquet	DuckDB	Parquet
JDBC row	custom binaries	JDBC row
tar.gz	shell scripts	tar.gz
html	…	html
docx		docx
pptx		pptx
png		png
…		…

At Scale



(circa 2022) Since its inception, Pinterest's philosophy has always been centered around data. As a data driven company, that means all data ingested is stored for further use. This looks like 600 terabytes of new data every day, encompassing over 500 petabytes of total data. At this scale, big data tooling plays a critical role in enabling our company to gather meaningful insights. This is where the workflow team comes in. We help facilitate over 4000 workflows, which produce 10,000 daily flow executions and 38,000 daily job executions on average.

https://medium.com/pinterest-engineering/spinner-pinterests-workflow-platform-c5bbe190ba5

Solution?

You need a workflow orchestrator like Apache Airflow

A platform created by the community to
programmatically author, schedule and monitor workflows.

At every point in time: What should we run next?
Taking into account:

What's the goal?
What resources are available?
What tasks have dependencies satisfied?
What tasks can run in parallel?
What tasks have failed?
What tasks can be retried?

Requirements

State management
Dependency management
Resource management
Scheduling
Failure handling
Observability

Just three additional nuances…

Temporal granularity
Backfills
Late arrivals


EXPRESSWAY
TOYOTA
COROLLA LE
ONTARIO
HADOOP