Data Processing and Orchestration

Kristo Raun

Data Engineering 2024 Fall

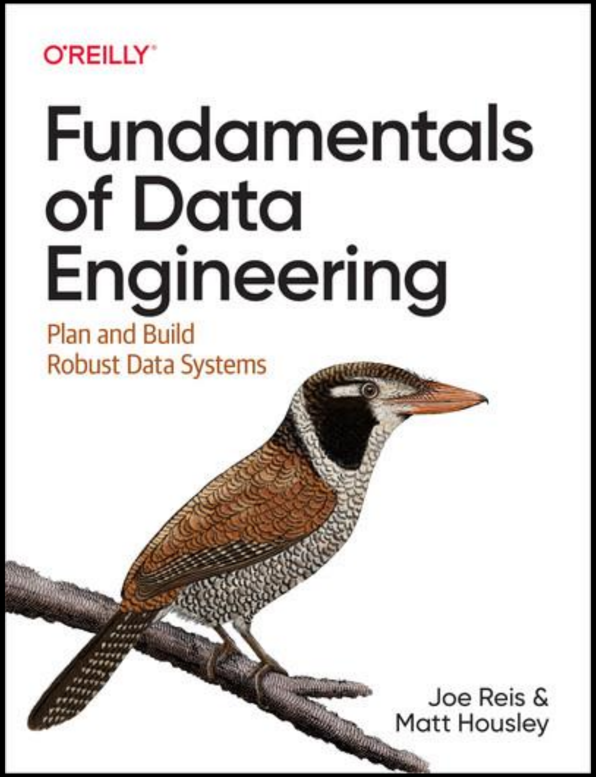
UNIVERSITY OF TARTU
1632

Agenda

- Data processing: ETL, ELT, CDC
- Data orchestration
- Airflow setup
- Quiz session

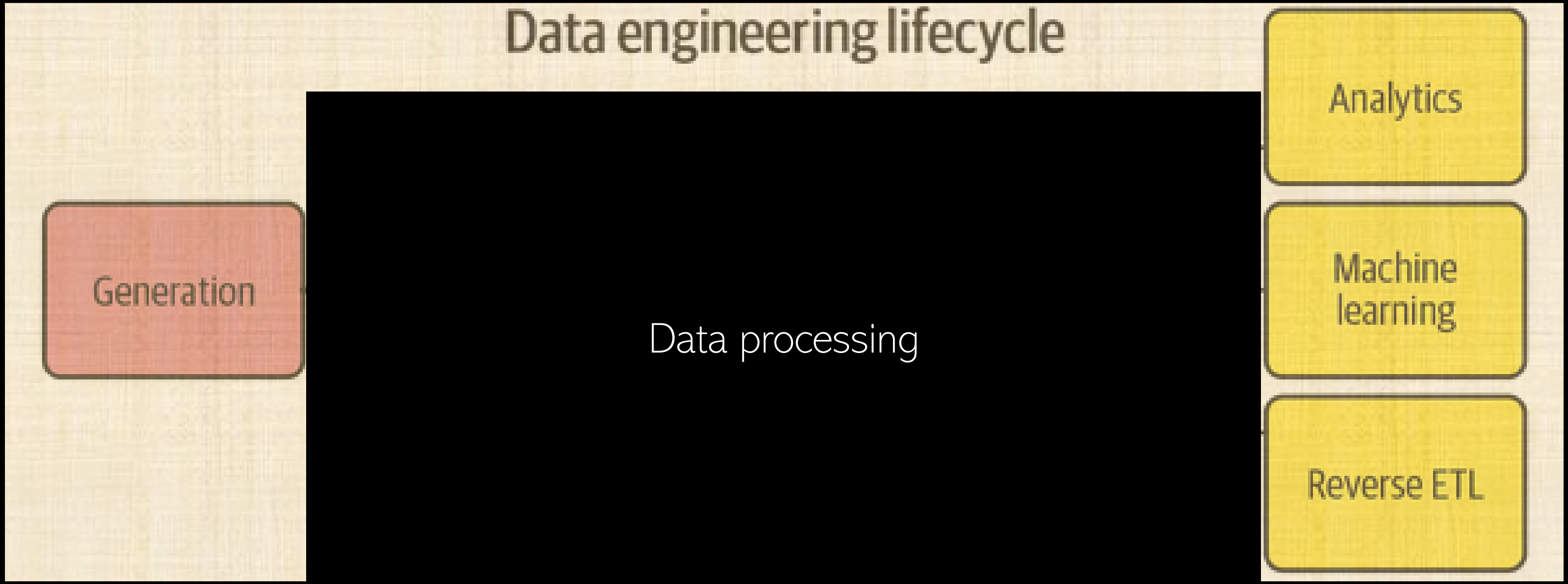
Reading

- Chapters V and VII

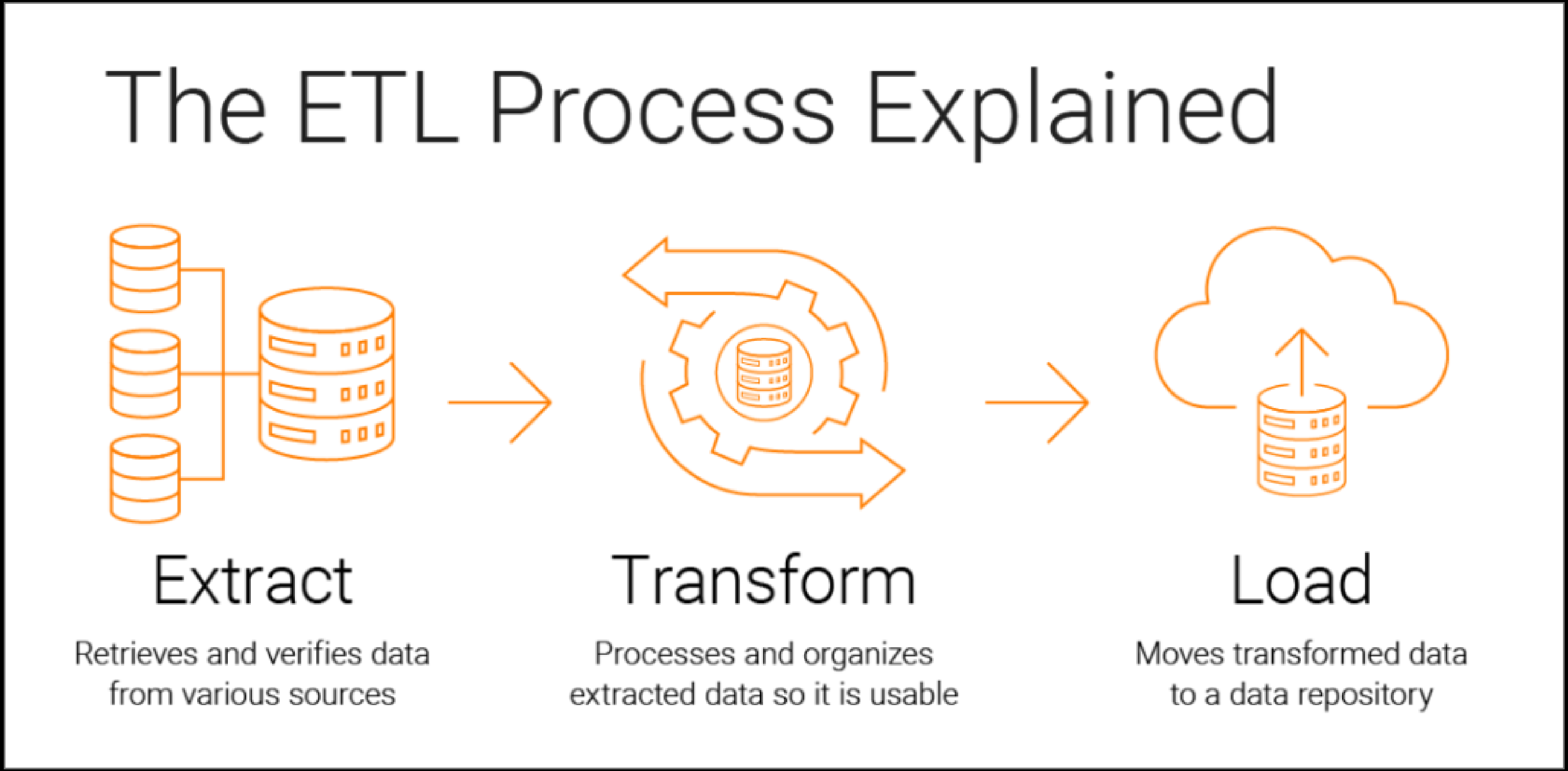

O'REILLY
Fundamentals of Data Engineering
Plan and Build Robust Data Systems
Joe Reis & Matt Housley

Data processing

- From raw data
- To usable information



Data processing: ETL



The ETL Process Explained

Extract

Retrieves and verifies data from various sources

Transform

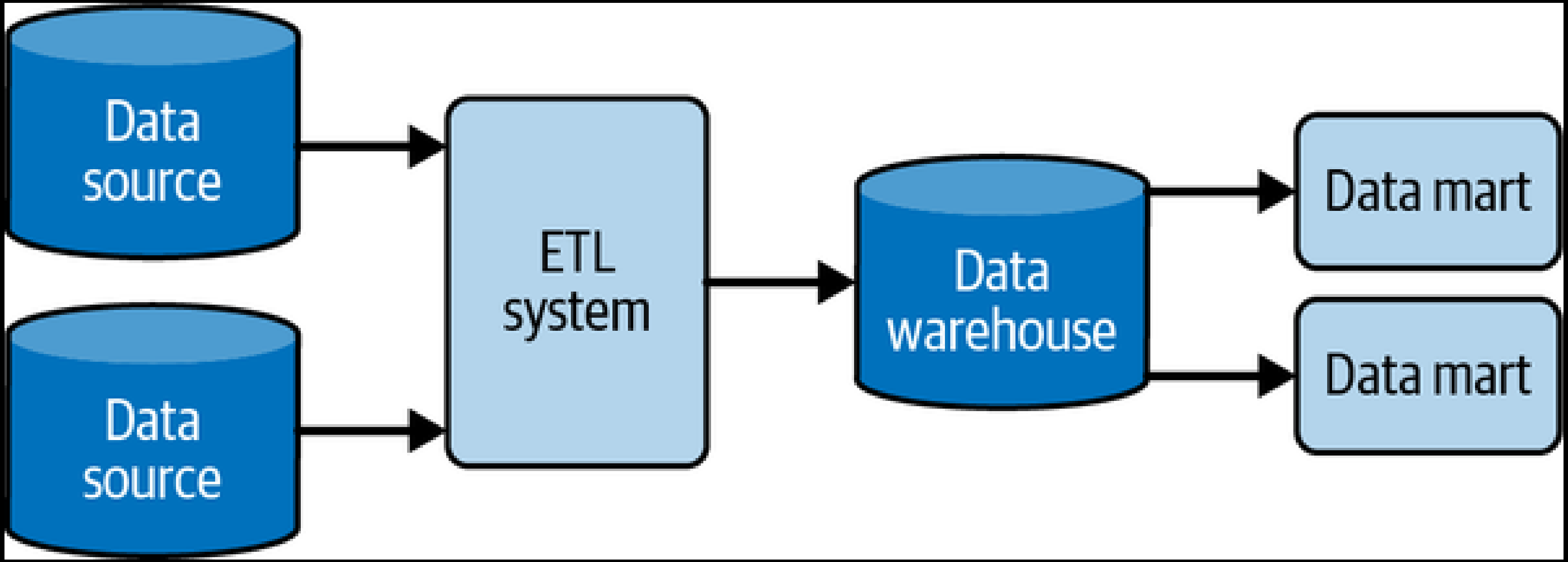
Processes and organizes extracted data so it is usable

Load

Moves transformed data to a data repository

https://www.thedataops.org/list-of-etl-extract-transform-load-tools/

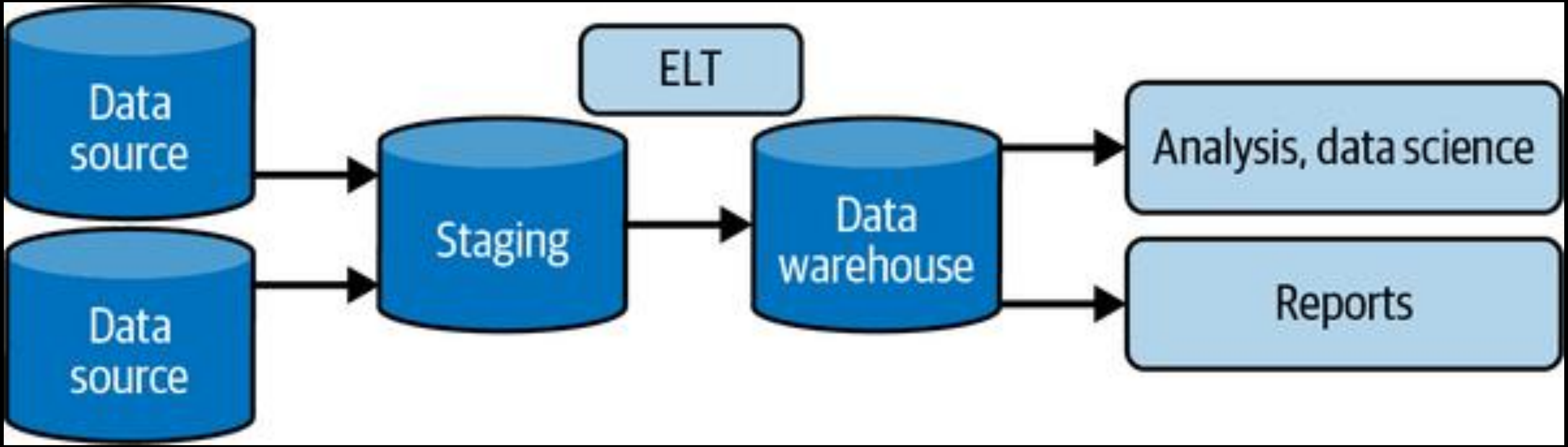
Data processing: ETL



Fundamentals of Data Engineering, ch III

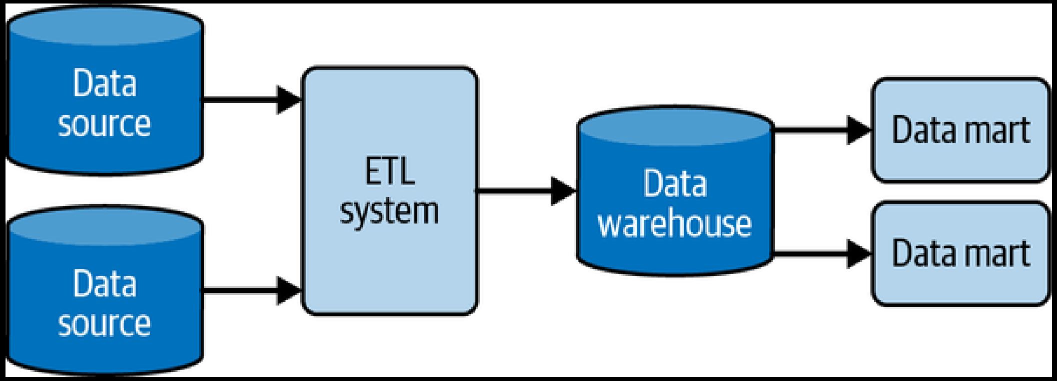
Data processing: ELT

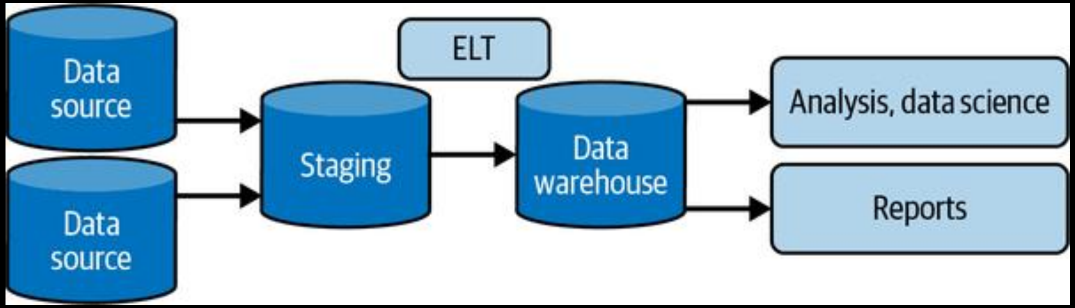
- Cloud data warehouses
  - Cheap and elastic storage
  - Increased processing power



Fundamentals of Data Engineering, ch III

Data processing: ETL vs ELT





Fundamentals of Data Engineering, ch III

Data processing: CDC

- Change Data Capture
- Extract each change in the source system
- Used for near real-time processing

Data processing: comparison

	ETL	ELT	CDC
Flow	Data cleaned between source and DWH/lake	Data loaded to DWH/lake, then cleaned	Incremental changes
Type	Batch	Batch	Streaming
Use case	Legacy, or specific privacy/business req.	State-of-the-art cloud warehousing	Near real-time updates
Scalability	Low scalability (high requirements on transformation)	High scalability	Depends on source system

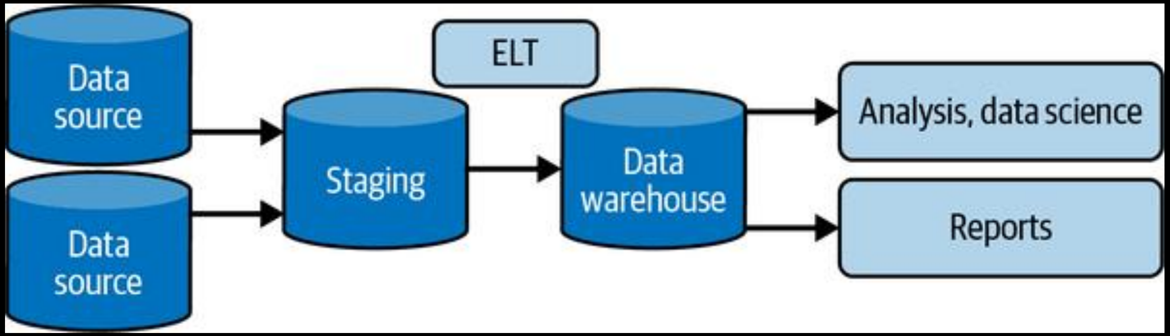
Agenda

- ~~Data processing: ETL, ELT, CDC~~
- Data orchestration
- Airflow setup
- Quiz session

Data orchestration

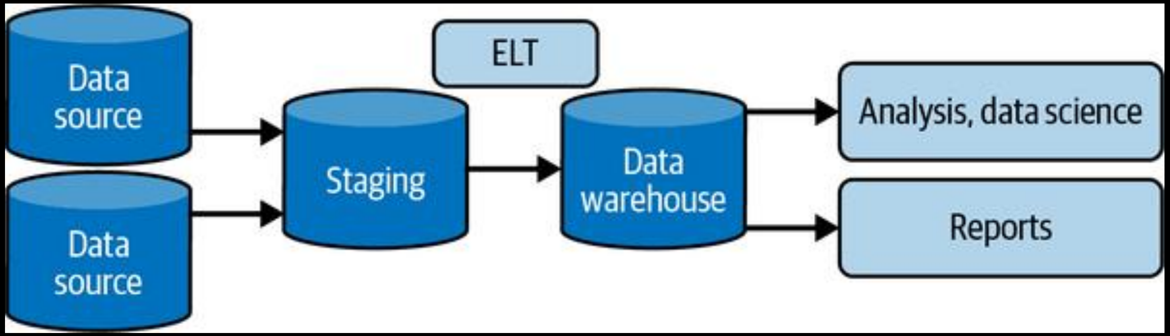
- Coordinating many jobs
- DAG
  - Directed Acyclic Graph
- Batch-oriented

Data orchestration – use case 1



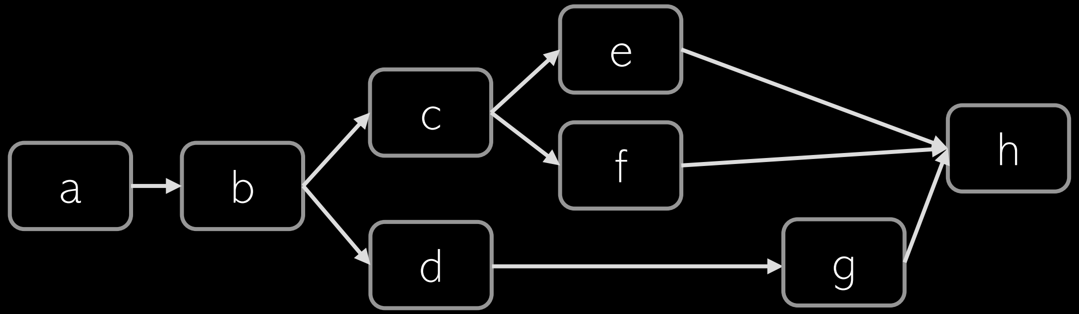
- Sales report
  - Sources:
    - CRM
    - Sales system
  - Both need to be loaded to DWH before transformations can be applied
  - Error handling

Data orchestration – use case 2



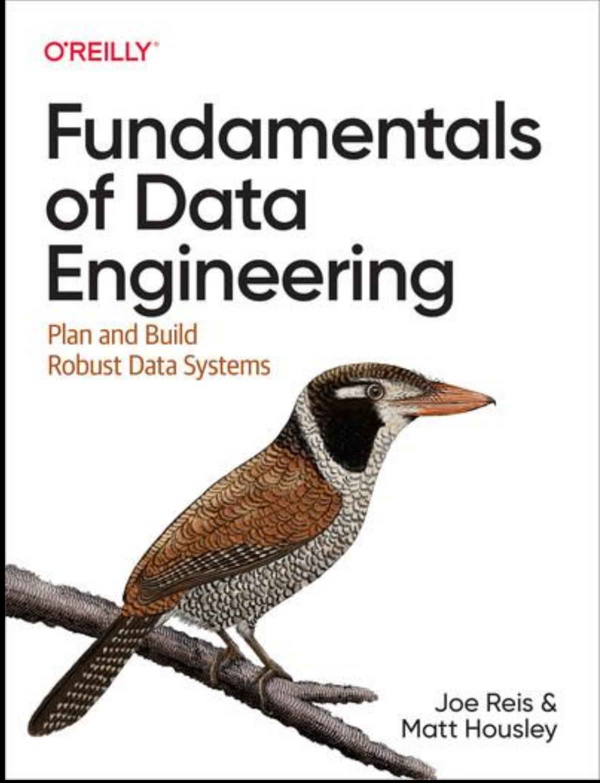
- External report upload
  - Source:
    - .csv uploaded to object storage daily
    - The file upload is controlled by the external vendor
  - The rest of the workflow should only start when file has been uploaded
  - Error handling

Data orchestration – DAG



- DAG
  - Directed
    - Determines task orders and dependencies
  - Acyclic
    - You can't loop back to an already completed task (avoids paradoxes, infinite loops)
- Control flow:
  - Sequential
  - Parallel
  - Conditional (branching)
  - *Various subtypes depending on the tool*

Read more:


O'REILLY
Fundamentals of Data Engineering
Plan and Build Robust Data Systems
Joe Reis & Matt Housley

- Chapter V: Data generation in source systems
  - How is data created?
  - Types of data in source systems
- Chapter VII: Ingestion
  - Batch ingestion considerations
  - Ways to ingest data
  - "At times, the minutiae of ingestion may feel tedious, but the exciting data applications (e.g., analytics and ML) cannot happen without it."
- Why Data Engineers LOVE/HATE Airflow (by Seattle Data Guy) https://www.youtube.com/watch?v=h5X3124R61U