

# Introduction to Data Engineering

Kristo Raun

Data Engineering 2024 Fall



UNIVERSITY OF TARTU

# Story I

- Making pizza
- 10 000 pizza outlets worldwide
- Order pizzas on:
  - Website
  - Food apps
  - Social Media (Facebook, Twitter)
  - Smart watches
  - Smart TV
  - In-car entertainment
- Who is the customer?
- What is the customer behavior?
- Dominos pizza
- 85 000 data sources into data warehouse

## Dominos: Data-driven decision making at the world's largest pizza delivery chain

23 July 2021

With over 10,000 outlets serving up millions of pizzas each year, Dominos is the largest pizza delivery chain in the world.

### How Dominos uses Big Data in practice

The company has consistently pushed its brand onto new and developing tech, and it's now possible to order pizzas on Twitter, smart watches and TVs, in-car entertainment systems such as Ford's Synch, and social media platforms like Facebook. This drive to keep a Dominos order button at customers' fingertips at all times is referred to as Dominos AnyWare.



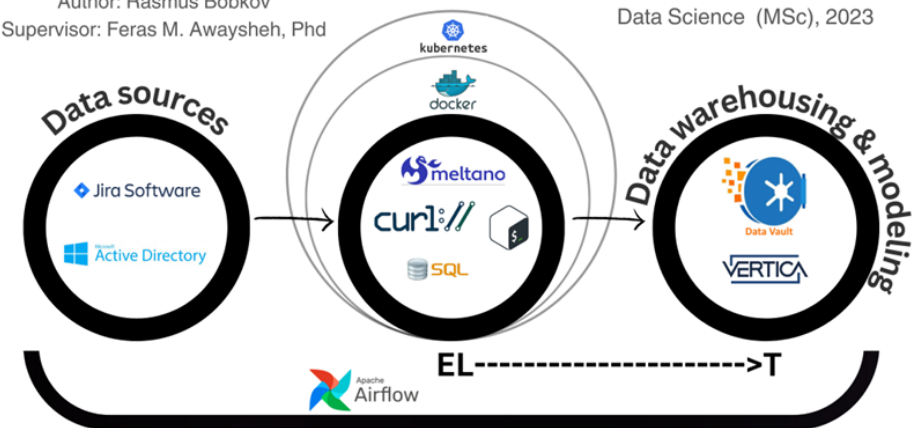
# Story II

- Management of health and welfare IT systems
- 300 000+ pensions, 200 000+ family benefits, 1 million page visits to personal health data
- Jira ticketing (incidents and requests)
- Jira: not meant for (custom) analytics
- Incorrect history
- TEHIK  
Health and Welfare Information Systems Centre

## Design and Implementation of an Incremental ELT Pipeline for a Jira Data Warehouse using Data Vault 2.0 Methodology and HP Vertica

Author: Rasmus Bobkov  
Supervisor: Feras M. Awaysheh, PhD

Data Science (MSc), 2023




# Story III

- Railway company
  - Relay-based railway crossings sometimes malfunction
  - Log data – hard to analyze, no insights
- 
- Eesti Raudtee (EVR)


**A Functional Prototype and General Architecture of Analytic Data Management for a Railway Company**

**Mait Metelitsa**  
Data Science (MSc) 2024  
Institute of Computer Science  
University of Tartu  
Supervisors: Kristo Raun, MSc;  
Prof. Ahmed Awad, PhD


#UniTartuCS




1. Describe AS-IS state of analytic data management in a railway sector company




2. Perform gap-analysis




3. Describe TO-BE architecture of analytic data management in a railway sector company



4. TO-BE based use case - purchase e-invoices row-level data analysis



5. TO-BE based use case - level crossing log data analysis



# Agenda

- ~~Background info~~
- Course intro
- Data engineering intro
- Docker lab
- Quiz session

# Course

## Instructors

Kristo Raun



PhD student  
Data Engineering (~9y)

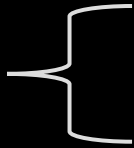
Riccardo Tommasini



Visiting professor  
Stream processing  
Knowledge graphs  
Query Languages

# Course

Group of 3  
Project proposal



Week	Lecture (16:15-18:00)		Practice (10:15-12:00)	
1	2024-09-02	No class	2024-09-05	No class
2	2024-09-09	Intro	2024-09-12	Docker+Postgres
3	2024-09-16	Data processing and orchestration	2024-09-19	Airflow
4	2024-09-23	Data modelling	2024-09-26	Kimball
5	2024-09-30	Data transformation	2024-10-03	dbt
6	2024-10-07	Data storage	2024-10-10	DuckDB
7	2024-10-14	NoSQL	2024-10-17	MongoDB
8	2024-10-21	Data Lakes	2024-10-24	Delta, Iceberg
9	2024-10-28	Graph Databases	2024-10-31	Neo4j
10	2024-11-04	Security and privacy	2024-11-07	Security and privacy
11	2024-11-11	Key-Value stores	2024-11-14	Redis
12	2024-11-18	Data governance	2024-11-21	Open Metadata
13	2024-11-25	Data visualization	2024-11-28	Streamlit
14	2024-12-02	Exam	2024-12-05	Working in class (tutoring for project)
15	2024-12-09	Working in class (tutoring for project)	2024-12-12	Working in class (tutoring for project)
16	2024-12-16	Project presentation (poster session)	2024-12-19	Redo exam

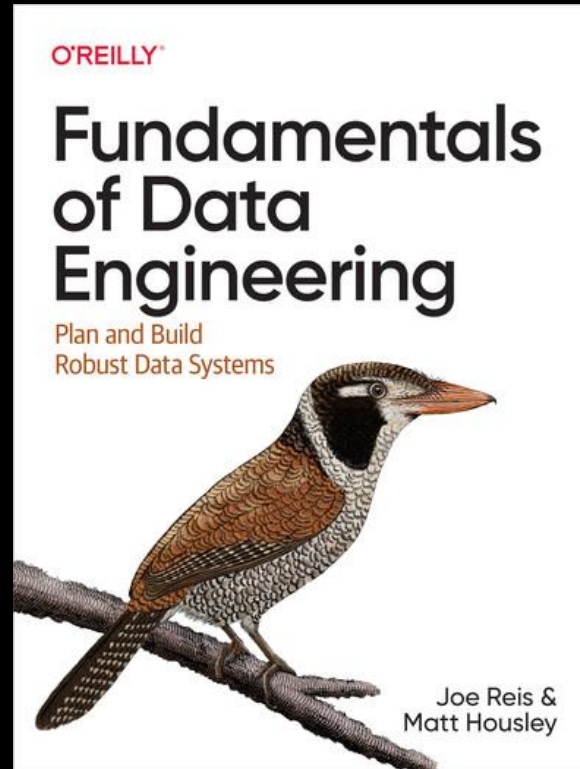
# Course

- Grade
  - Exam – 30%
    - TBD
      - ~ 10 questions
      - ~ 1.5h
  - Project – 70%
    - TBD
      - Group of 3
      - Project proposal (~ 2024-10-24)
      - Min passing grade
        - Join and clean 2 datasets
        - Use orchestration
        - Dimensional modeling
      - Project submission deadline 2024-12-12
      - Poster session 2024-12-16



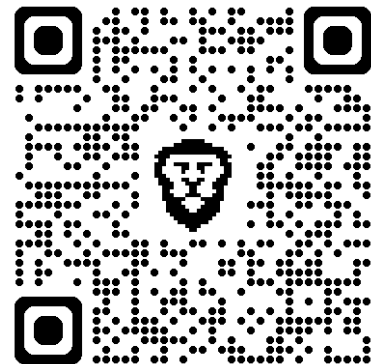
# Course

## Reading



<https://learning.oreilly.com/library/view/fundamentals-of-data/9781098108298/>

<https://utlib.ut.ee/en/oreilly>

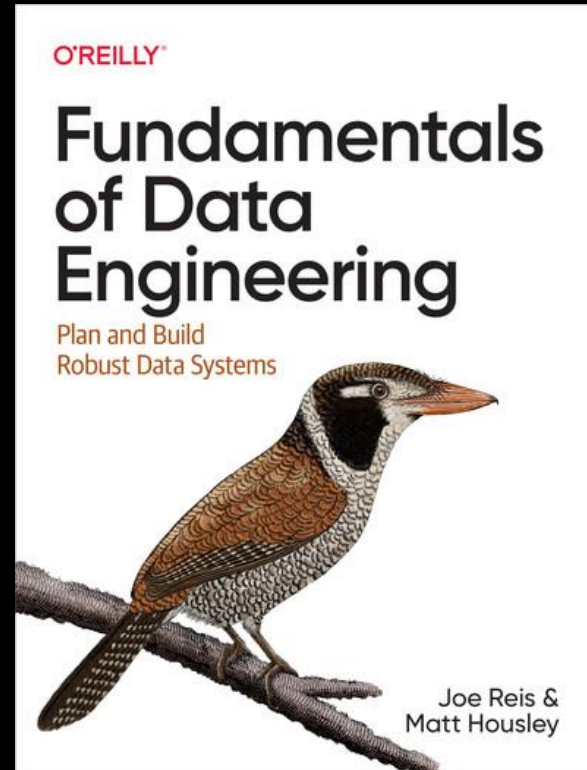


# Agenda

- ~~• Background info~~
- ~~• Course intro~~
- Data engineering intro
- Docker lab
- Quiz session

# Reading

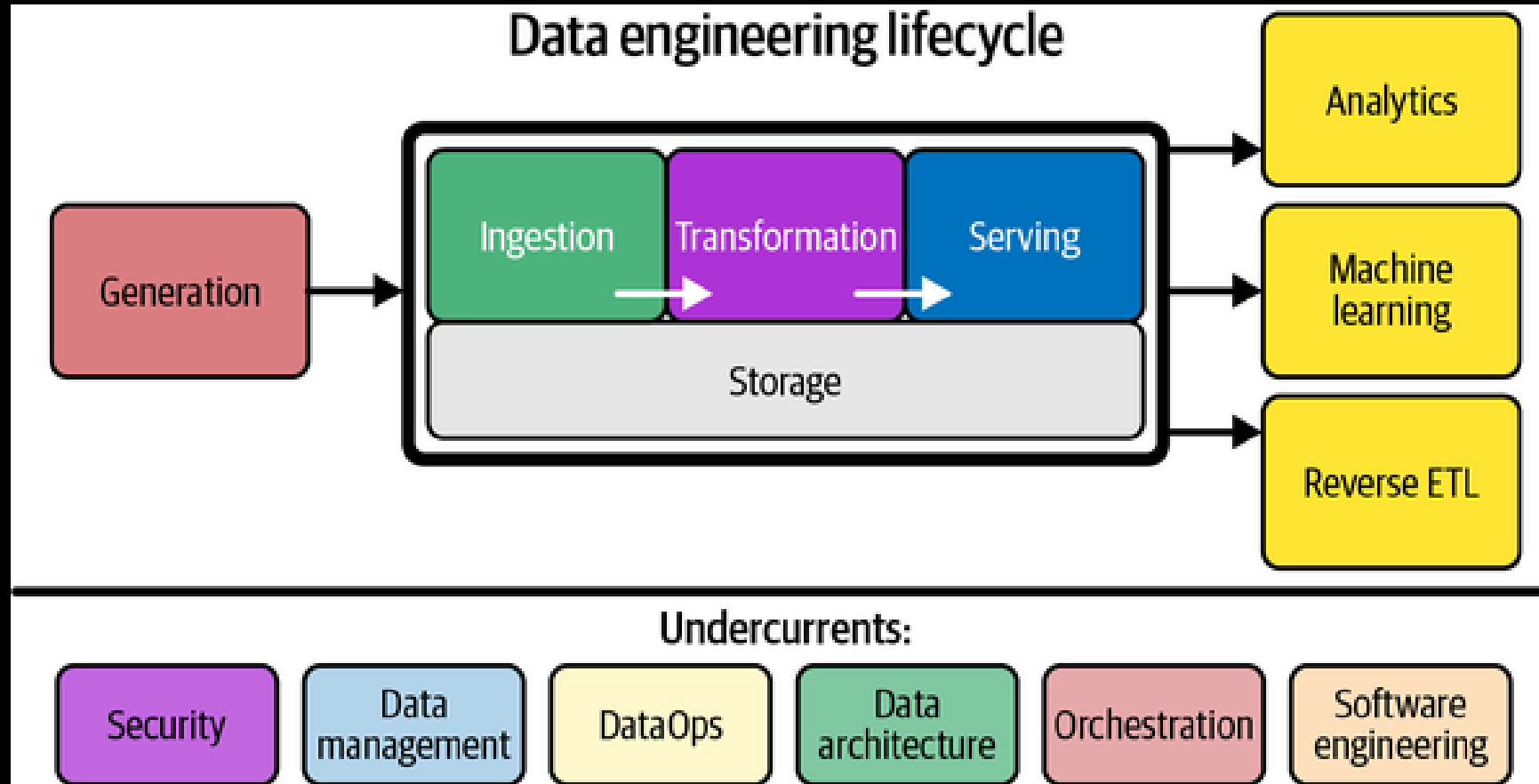
- Chapters I and II



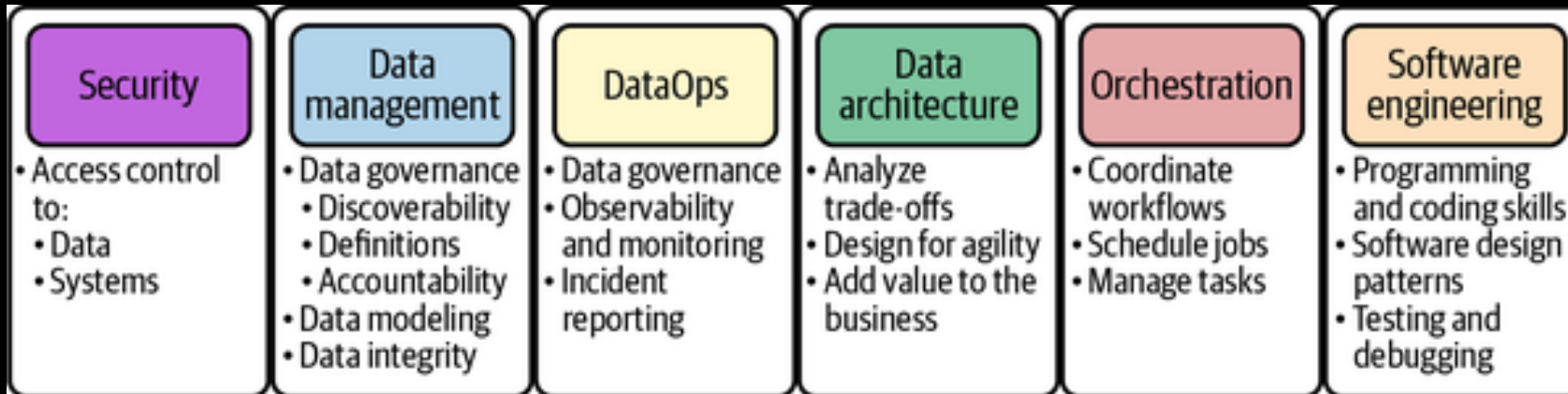
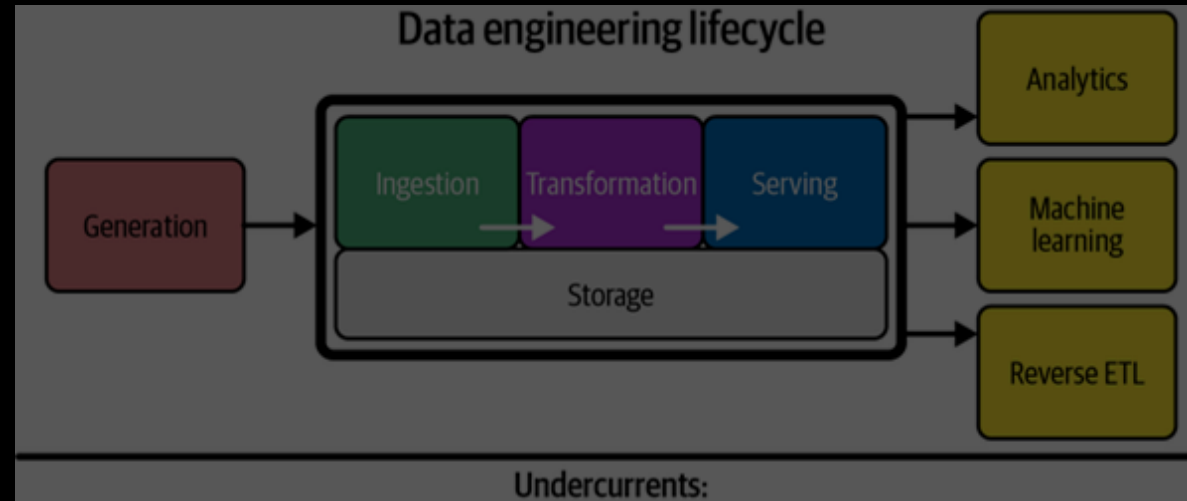
# What is Data Engineering?

Data engineering is the development, implementation, and maintenance of systems and processes that take in raw data and produce high-quality, consistent information that supports downstream use cases, such as analysis and machine learning. Data engineering is the intersection of security, data management, DataOps, data architecture, orchestration, and software engineering. A data engineer manages the data engineering lifecycle, beginning with getting data from source systems and ending with serving data for use cases, such as analysis or machine learning.

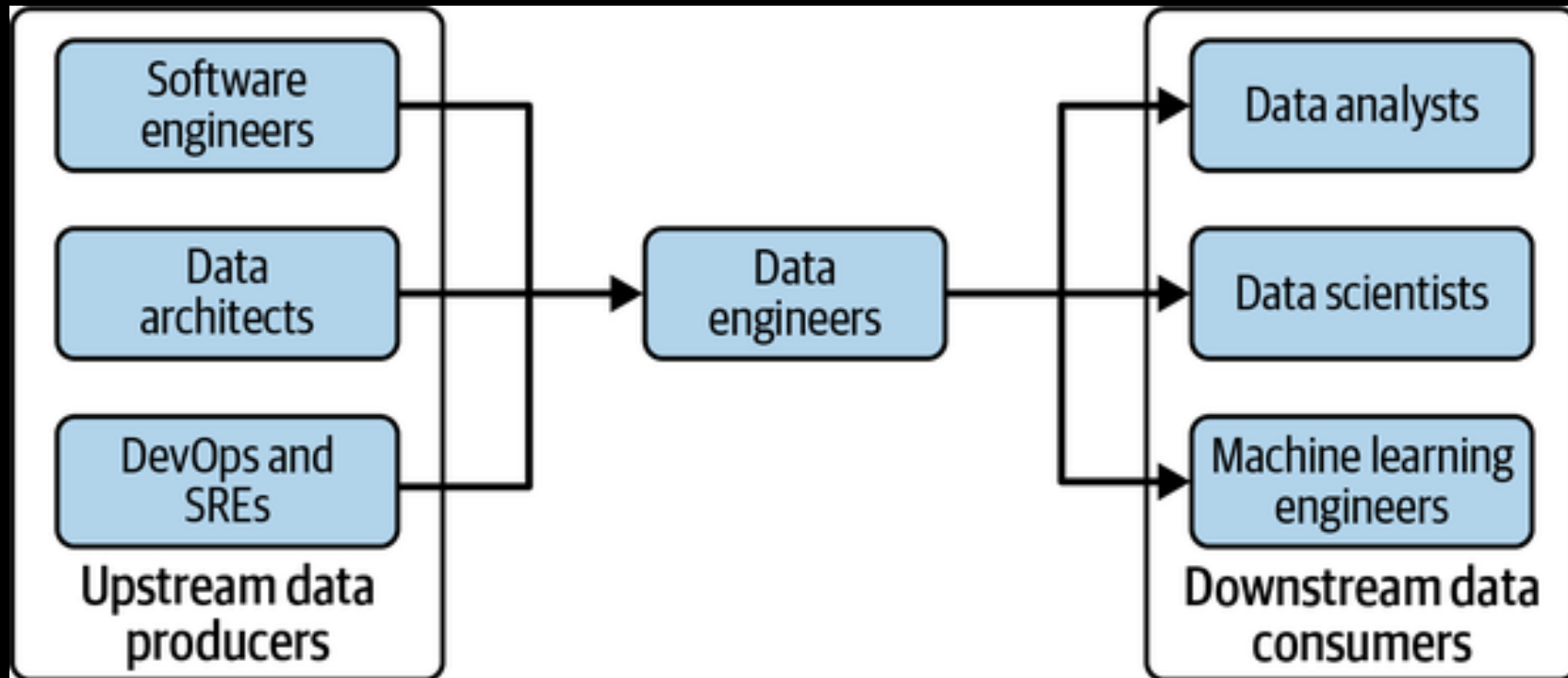
# Data Engineering Lifecycle



# Data Engineering Lifecycle



# Data Engineers in an organization



# Read more:

- Chapter I
  - DE history
  - DE and DS
  - DE skills (business, technical)
  - Whom data engineers work with
- Chapter II
  - Data Engineering Lifecycle in detail

