# Data Transformation

Kristo Raun

Data Engineering 2024 Fall

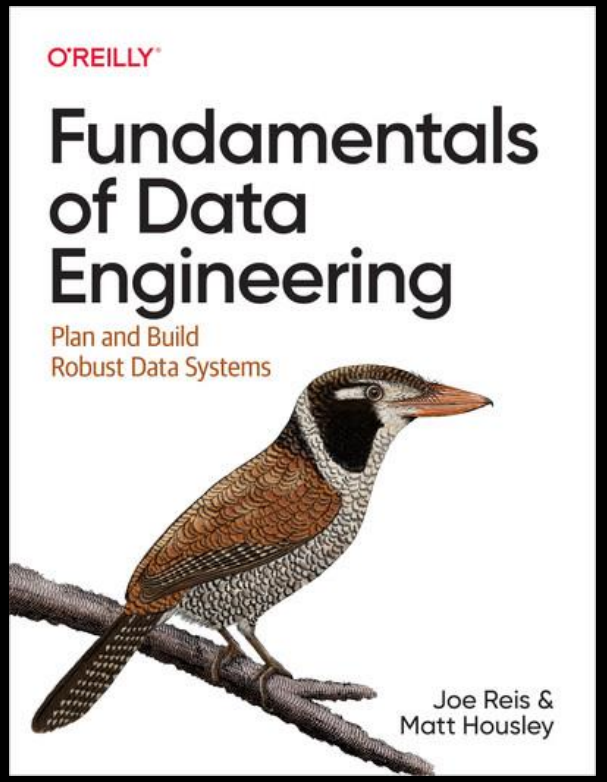

UNIVERSITY OF TARTU

# Agenda

- Data querying
- Data transformation
- Data wrangling
- Quiz session

# Reading

- Chapter VIII



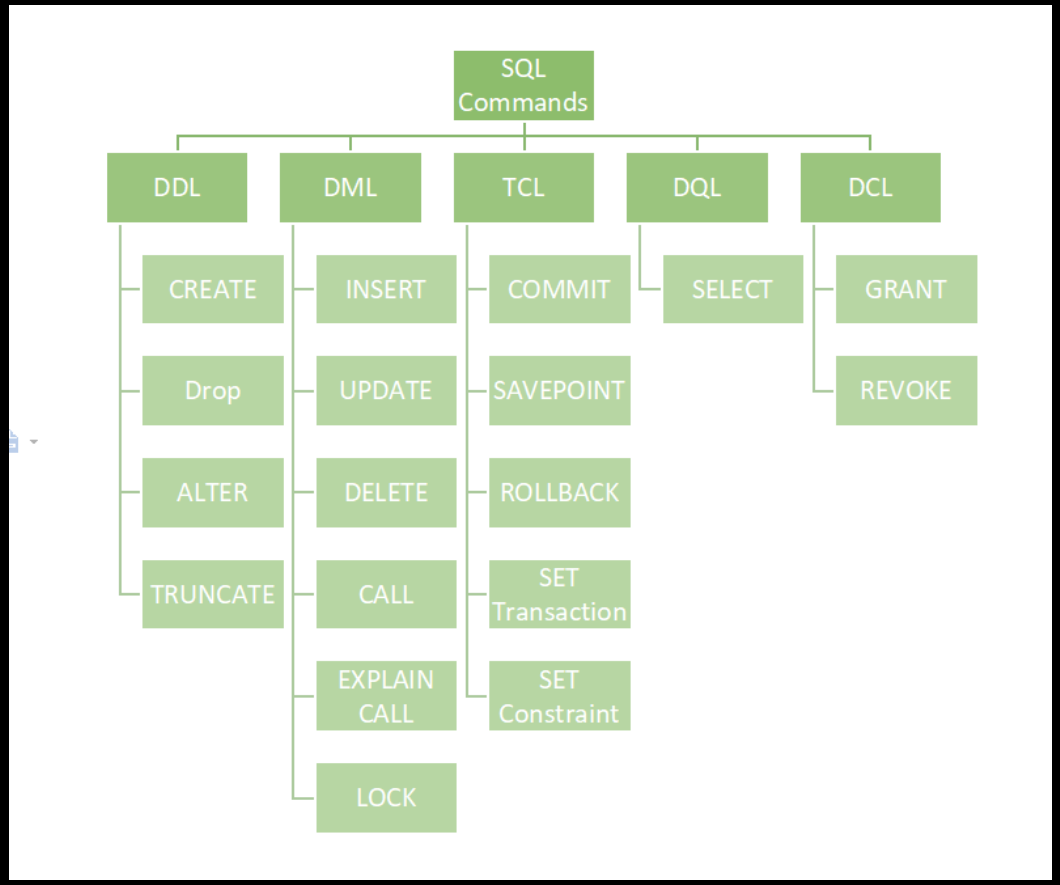
# Data querying

- What is a query?
- Query → Retrieve and act on data
- CRUD
  - Create-Read-Update-Delete
- Most basic query in SQL (Read)
  - “SELECT [columns] FROM [table];”

# Data querying

- Categorization
  - DDL – Data Definition Language (create/modify/delete db objects)
  - DML – Data Manipulation Language (CRUD for actual data)
  - DCL – Data Control Language (access rights)
  - TCL – Transaction Control Language (commits, rollbacks)
  - *DQL – Data Query Language ("SELECT", debated)*

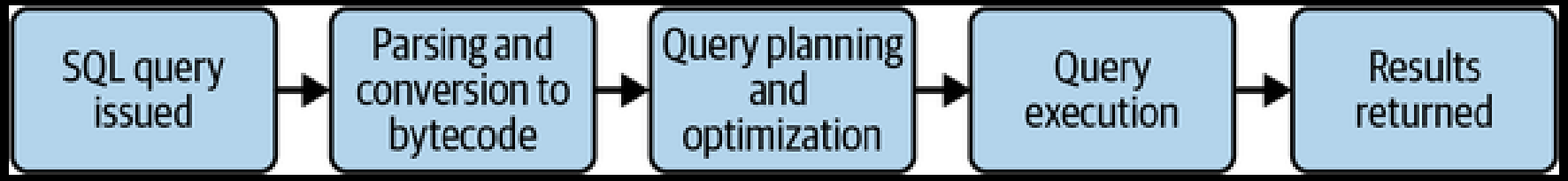
# Data querying




https://www.geeksforgeeks.org/sql-ddl-dql-dml-dcl-tcl-commands/

# Data querying

Life of a query



# Agenda

- ~~Data querying~~
- Data transformation
- Data wrangling
- Quiz session

# Data transformation

- Query vs transformation
  - “SQL query”
  - Query → read (SELECT)
  - SELECT customerId, SUM(salesInEur) FROM … …… ?
- Generally speaking, transformation:
  - Manipulates, enhances, and *saves* data
  - For downstream use
  - Increases the value of data

# Data transformation

- Transformation (vs query) persists the results
  - Ephemerally
  - Permanently
- Transformation (vs query) – generally more complex
  - Query – single CTEs, scripts, etc
  - Transformation – “bundle” of queries

# Data transformation

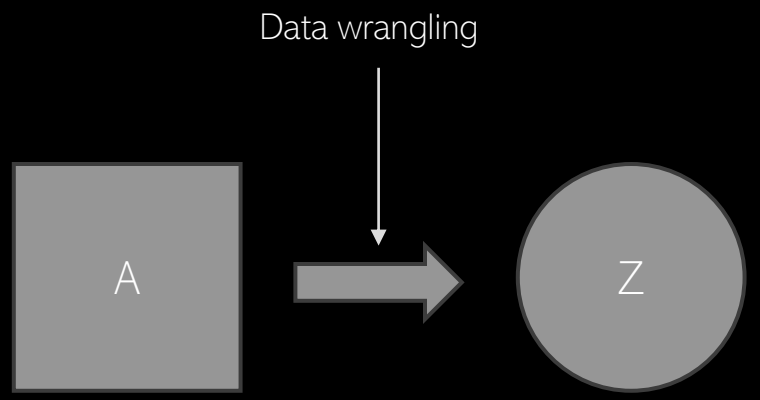
- Persistence types
  - Table
  - View
  - Materialized view
  - Federated query (Data virtualization)

# Agenda

- ~~Data querying~~
- ~~Data transformation~~
- Data wrangling
- Quiz session

# What is data wrangling?

- Data cleaning
- Data cleansing
- Data munging
- Data preprocessing
- Data preparation
- Data mapping
- Data transformation
- . . .



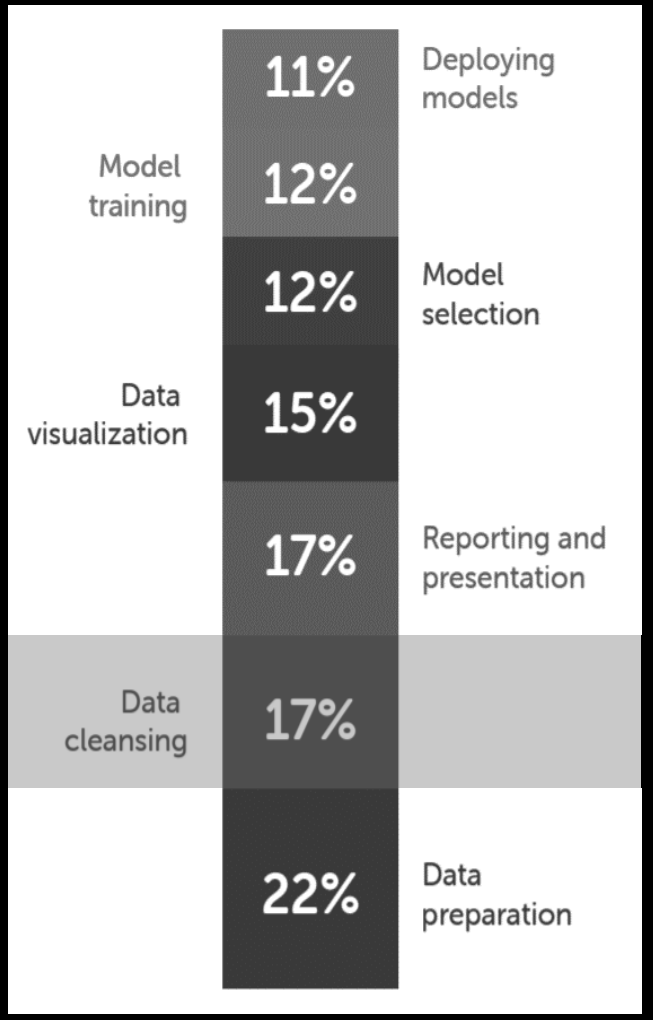

We need data in format Z
We have data in format A

# What is data wrangling?

- We want to change
  - Data format
  - Data type
- We want to fix
  - Missing values
  - Duplicates
- We want to
  - Augment
  - Group
  - Aggregate
  - Filter

# Why is data wrangling necessary?

- The ML view

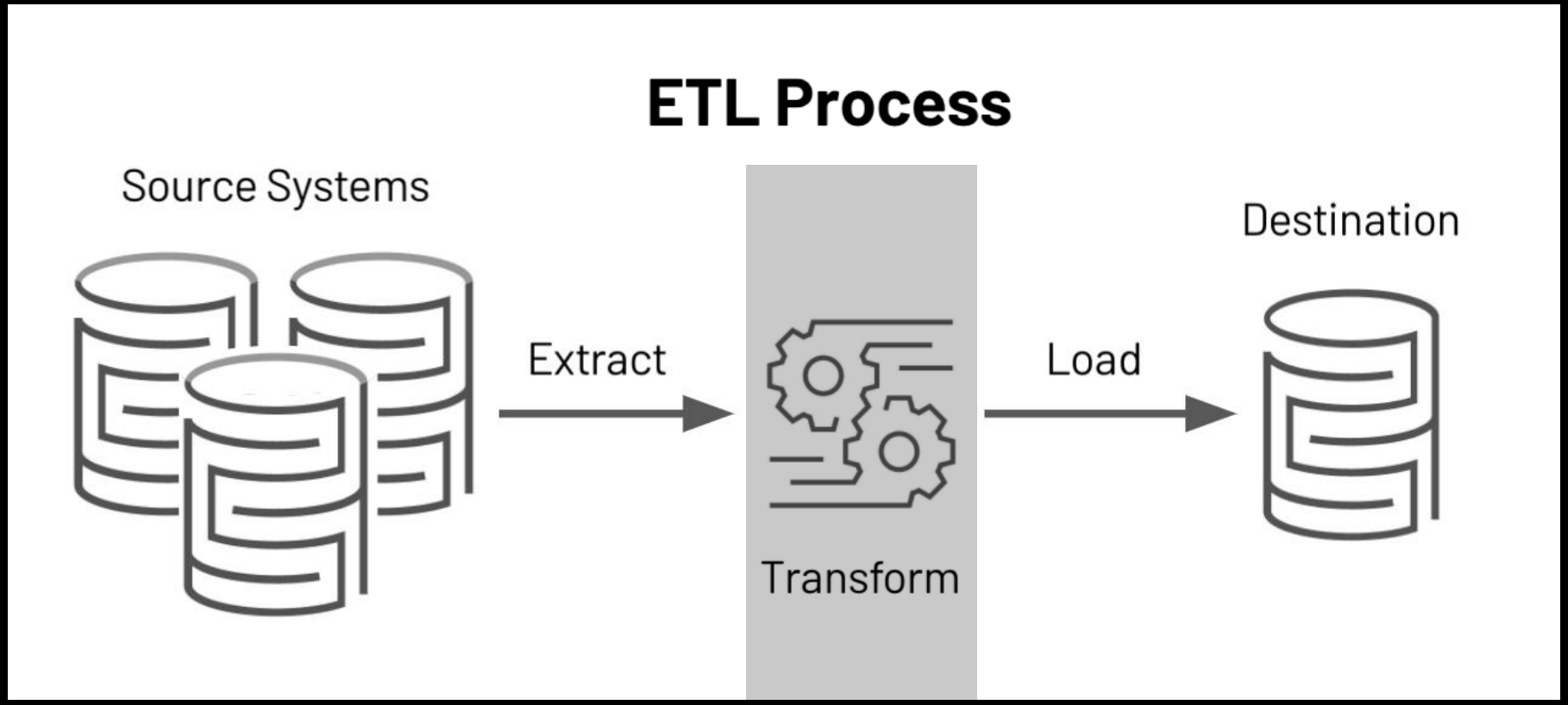



How do data scientists spend their time?

Source: 2021 State of Data Science Report
Anaconda
N = 2 030

# Why is data wrangling necessary?

- The BI view



# Types of data wrangling

- Validity of values
  - Phone numbers
    - “+372 51234567”
    - “00372 51234567”
    - “512 345 67”
    - “51234567”
    - 51234567
  - Checksum
    - SSN
    - Invoice reference number
    - Credit card

# Types of data wrangling

- Consistent values
  - Does zip code + city make sense?
  - Can there be a sales order worth more than 1 million euros?
  - If customer has conflicting emails in different systems, which system is correct?

# Types of data wrangling

- Duplicates
  - Do we accept duplicates?
  - Is it possible to set validity of data (updated timestamp)?
  - Keep only one row (try to make the choice idempotent)
  - Can we trust the source system to have unique values?

# Types of data wrangling

- Business rules
  - Does a premium customer have the associated premium services?
  - Is this product allowed to have this discount?

# Types of data wrangling

- Conforming values
  - If customer has conflicting emails in different systems, which system is correct?
  - Is there a unique code across systems for defining a … (product, customer, location, …)
    - Eg product name in ERP vs sales system vs website?

# Types of data wrangling

- Missing data
  - What is NULL?
  - Is NULL acceptable?
    - Aggregation over NULL is (usually) correct
    - -1 (or similar) to use for referential integrity
  - In data science/ML:
    - Delete data
    - Impute data
      - Easiest/fastest: median

# Types of data wrangling

- Wrong data type
  - Schema definition
    - Load in csv without schema - everything is a string
    - Inferring schema (eg Spark) can end up with wrong type
  - Unit
    - String vs integer vs decimal type
      - A hundred pieces
      - 100 pieces
      - 100.55 pieces
  - Timestamps
    - UTC vs local
    - UNIX timestamp
      - Avro files – date/timestamp is integer. Eg how many days since 1 January 1970 (ISO calendar)

# Types of data wrangling

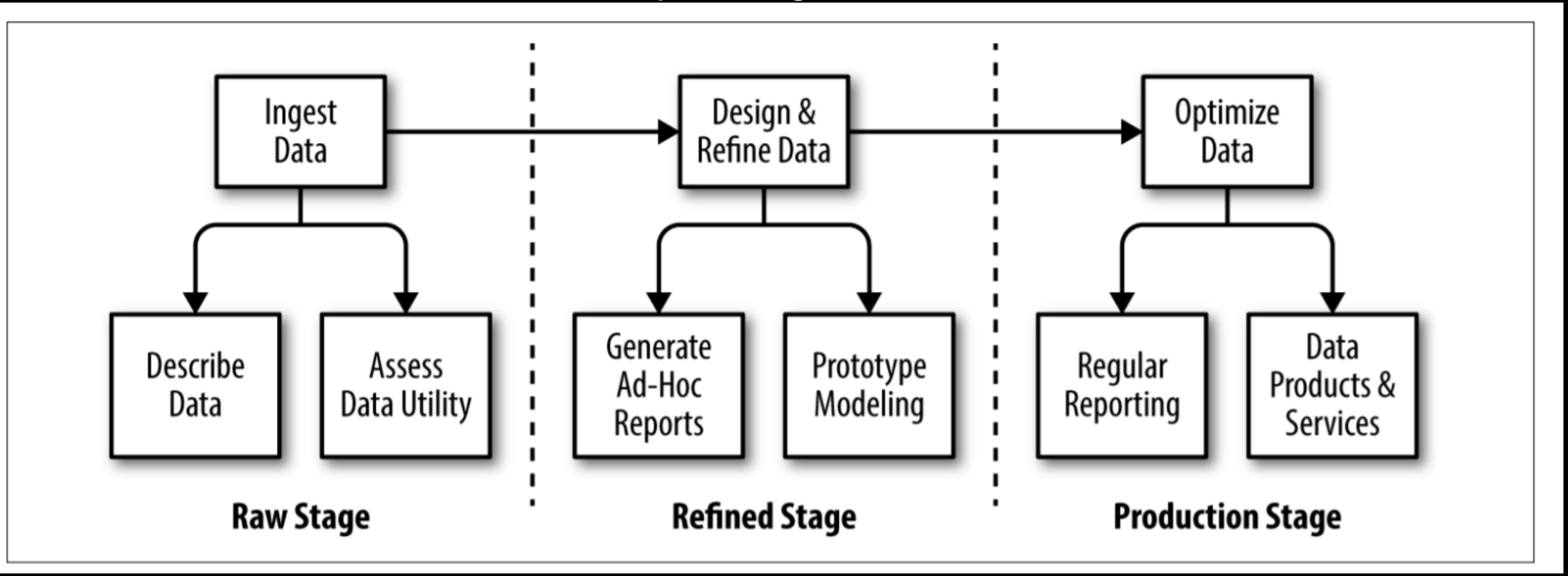
- Wrong data structure
  - Structured data
    - CSV
    - Excel
    - Conventional database
  - Nested data (semistructured data)
    - JSON
    - Parquet
    - Struct, array (in Spark, BigQuery, etc)
  - Unstructured data
    - Text
    - Images
    - Audio/Video

# Types of data wrangling

- Aggregations
  - Grouping
    - Correct grouping columns (level of slice/dice)
  - How will the data be used (visualization, reporting, ML)

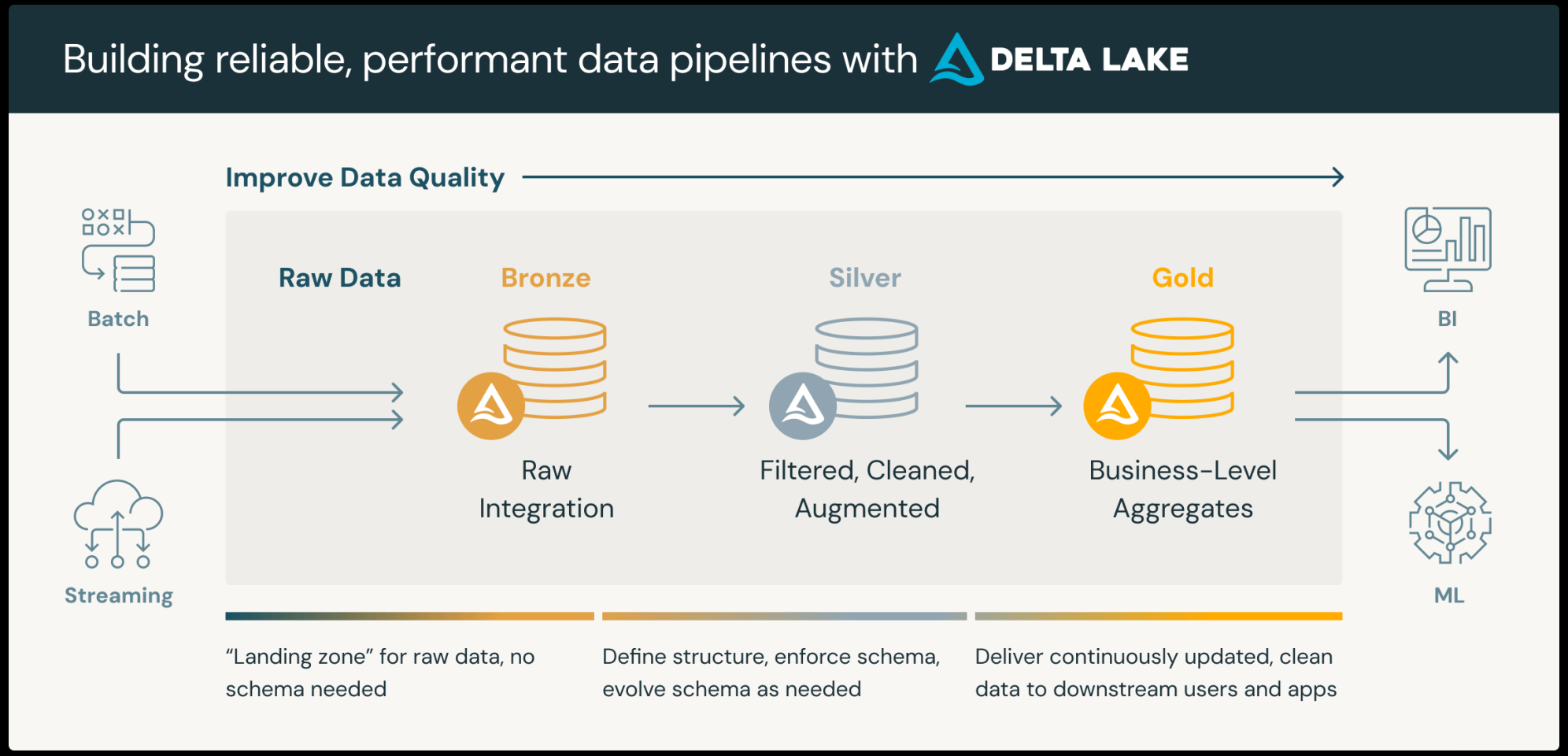
# Levels of data wrangling

- Various definitions. Commonly 3 stages are defined




Source: Trifacta

# Levels of data wrangling




Source: Databricks

# Agenda

- ~~Data querying~~
- ~~Data transformation~~
- ~~Data wrangling~~
- Quiz session

# Further reading

- Chapter VIII
  - Joins in distributed systems
  - Update patterns
- https://www.databricks.com/glossary/what-is-data-transformation
- https://www.getdbt.com/blog/what-exactly-is-dbt