Data Governance

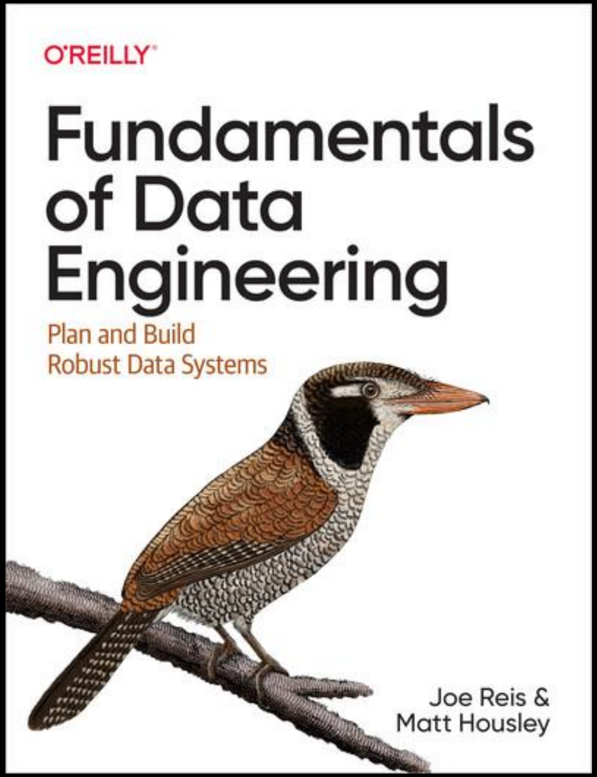
Kristo Raun

Data Engineering 2024 Fall

UNIVERSITY OF TARTU

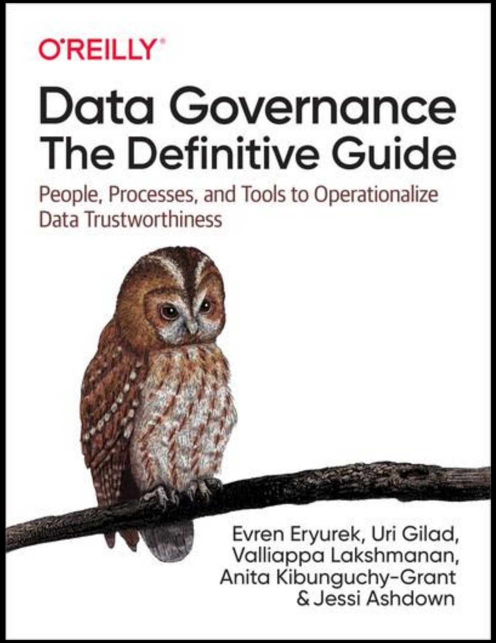
Reading

- Chapter II
  - Data Governance


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

Data Governance

"Data governance is, first and foremost, a data management function to ensure the **quality**, **integrity**, **security**, and **usability** of the data collected by an organization."

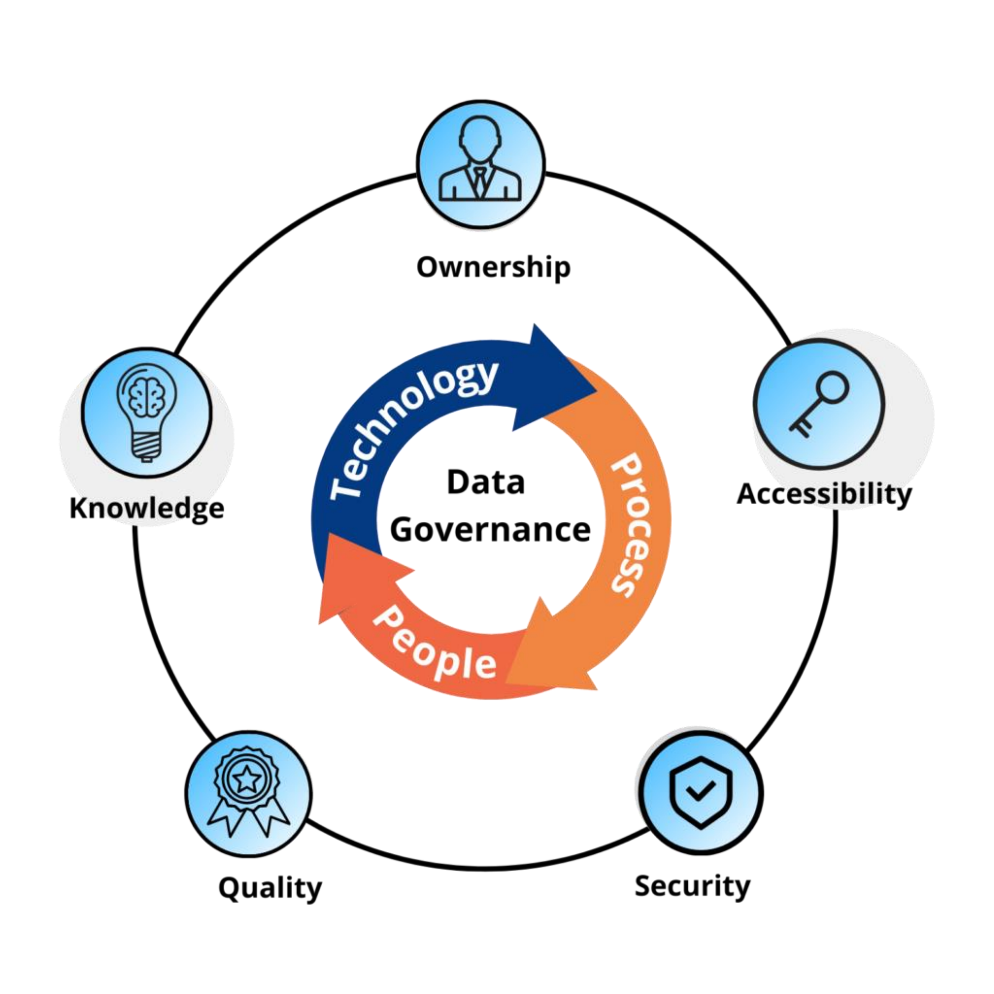


O'REILLY

Data Governance
The Definitive Guide

People, Processes, and Tools to Operationalize Data Trustworthiness

Evren Eryurek, Uri Gilad,
Valliappa Lakshmanan,
Anita Kibunguchy-Grant
& Jessi Ashdown



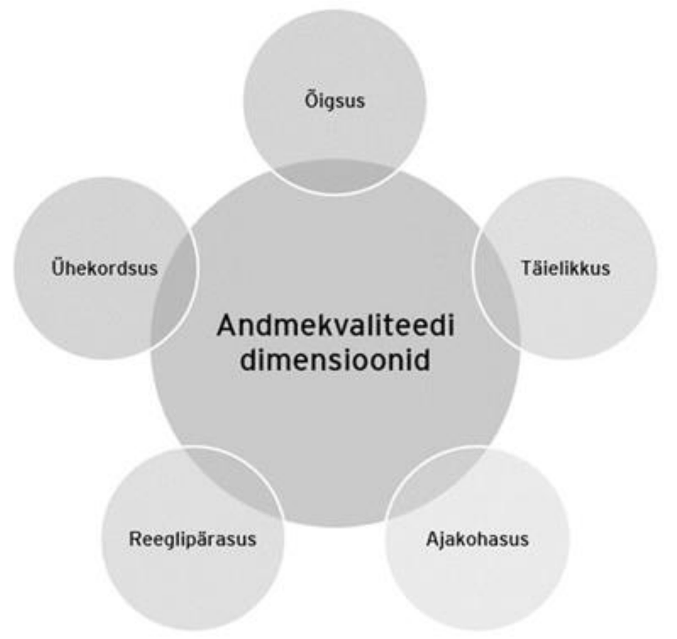
https://saxon.ai/services/data-governance/

Agenda

- Data Quality
- Data Usability (Data Discoverability)
- Quiz session

Data Quality

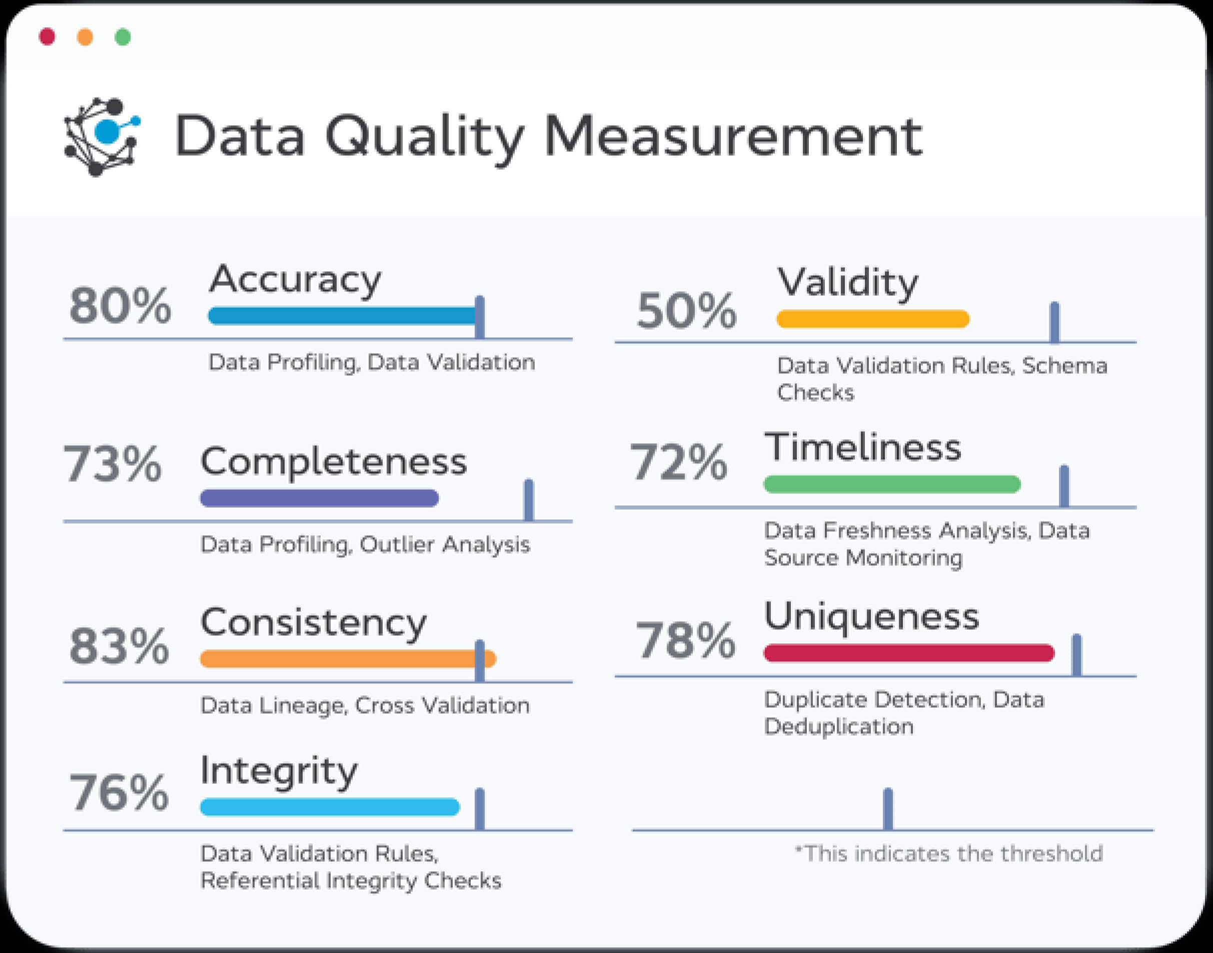
1. Accuracy (Õigsus), e.g. Does the data entry match real-world values?
   - Correct: Name: "Tõnu," Gender: "M" (matches actual values).
   - Incorrect: Name: "T6nu," Gender: "N" (typo, incorrect gender).
2. Completeness (Täielikkus), e.g. Are all required fields filled?
   - Correct: Customer record with all fields (Name, Email, Phone) populated.
   - Incorrect: Customer record missing "Email" field.
3. Timeliness (Ajakohasus), e.g. Is the data available when needed?
   - Correct: Delivery status updated in real time for tracking.
   - Incorrect: Delivery status updated only weekly, causing delays in tracking.
4. Orderliness (Reeglipärasus), e.g. Does the data follow the agreed format?
   - Correct: Date of birth recorded as "2024-11-09" (ISO format).
   - Incorrect: Date of birth recorded as "11/9/24" (ambiguous, non-standard).
5. Uniqueness (Ühekordsus), e.g. Are there duplicate entries in the dataset?
   - Correct: Only one record exists per customer (e.g., John Doe has one entry).
   - Incorrect: Same customer has multiple records (e.g., "John Doe" and "J. Doe" for the same individual).



Andmekvaliteedi juhis, mai 2023
https://www.kratid.ee/en/juhised

Data Quality Metrics

- Perfect Data Quality is impossible
  - How do we evaluate data quality?
  - What level is acceptable?



Data Quality Measurement

80% Accuracy
Data Profiling, Data Validation

50% Validity
Data Validation Rules, Schema Checks

73% Completeness
Data Profiling, Outlier Analysis

72% Timeliness
Data Freshness Analysis, Data Source Monitoring

83% Consistency
Data Lineage, Cross Validation

78% Uniqueness
Duplicate Detection, Data Deduplication

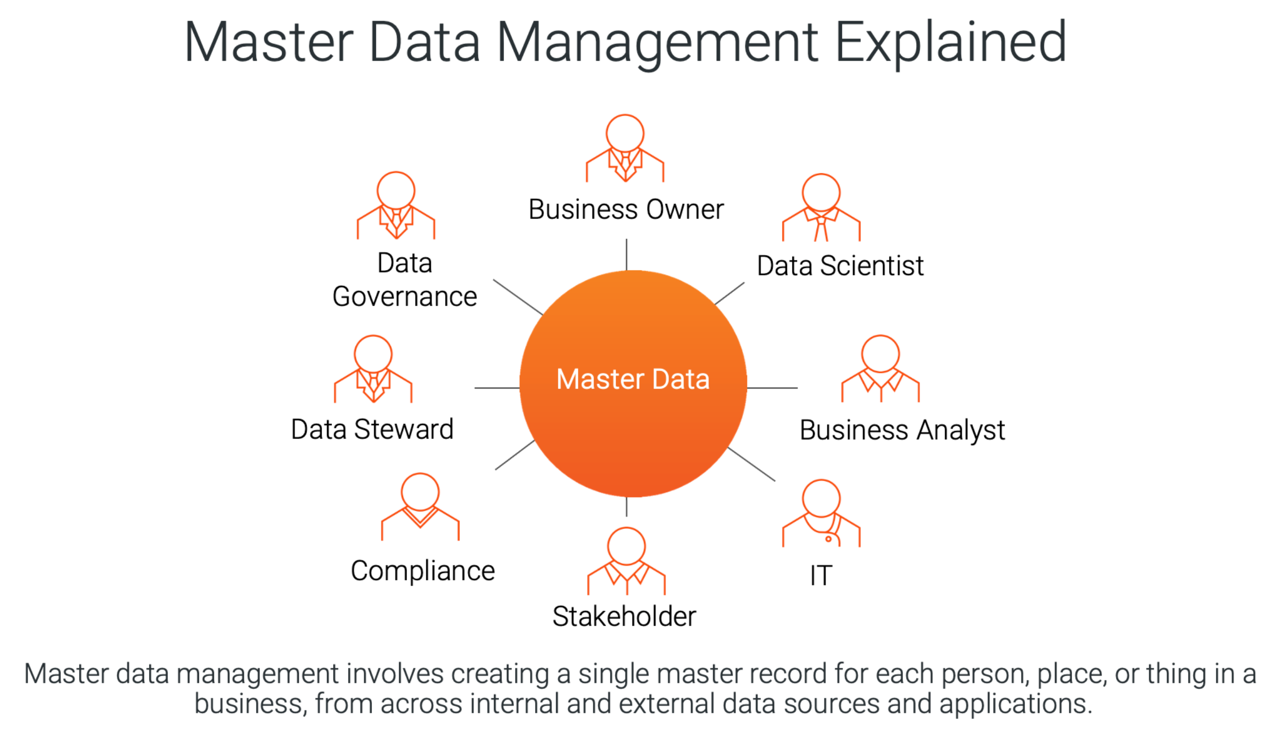
76% Integrity
Data Validation Rules, Referential Integrity Checks

*This indicates the threshold

https://www.ovaledge.com/blog/data-quality-metrics

Data Quality: Master Data Management

- Consistent entity definitions
  - Standard formats
  - Single source of truth (for key entities)
- E.g.:
  - Customer
  - Product
  - Location



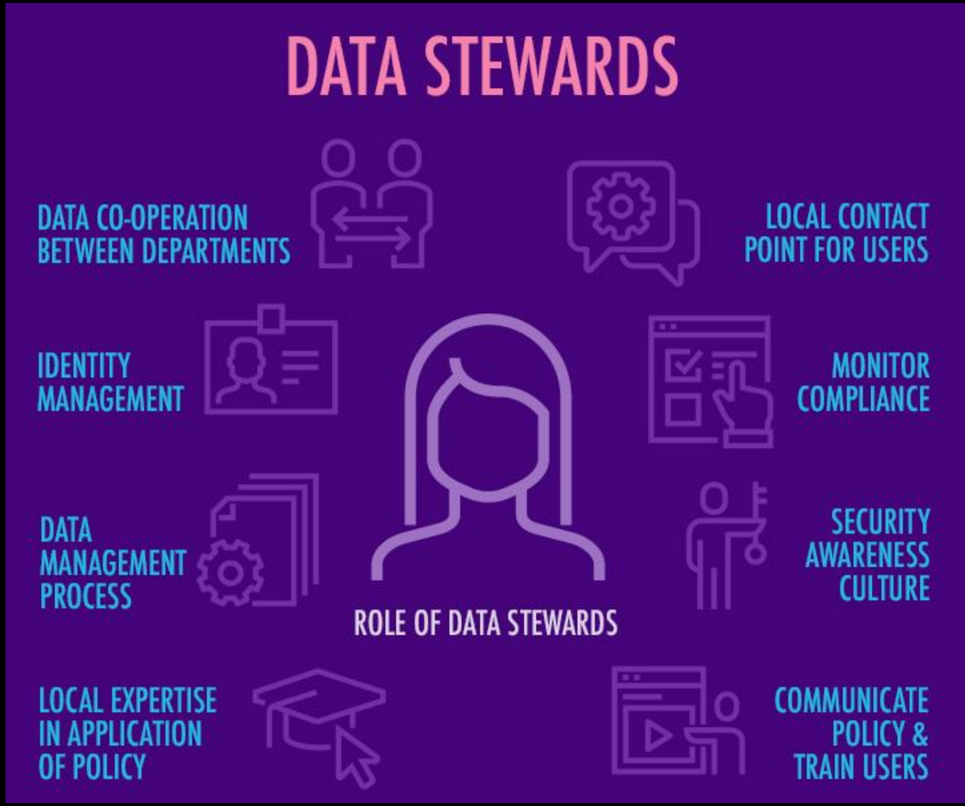
Master Data Management Explained

Master data management involves creating a single master record for each person, place, or thing in a business, from across internal and external data sources and applications.

https://www.informatica.com/resources/articles/what-is-master-data-management

Data Quality: Data Stewardship

- Accountability: who is “in charge” of the data?
- Data Steward: Managing, overseeing, and enforcing data governance policies within an organization.
  - Data Quality Monitoring
  - Data Compliance
  - Helping data users


DATA STEWARDS
DATA CO-OPERATION BETWEEN DEPARTMENTS
LOCAL CONTACT POINT FOR USERS
IDENTITY MANAGEMENT
MONITOR COMPLIANCE
DATA MANAGEMENT PROCESS
SECURITY AWARENESS CULTURE
ROLE OF DATA STEWARDS
LOCAL EXPERTISE IN APPLICATION OF POLICY
COMMUNICATE POLICY & TRAIN USERS

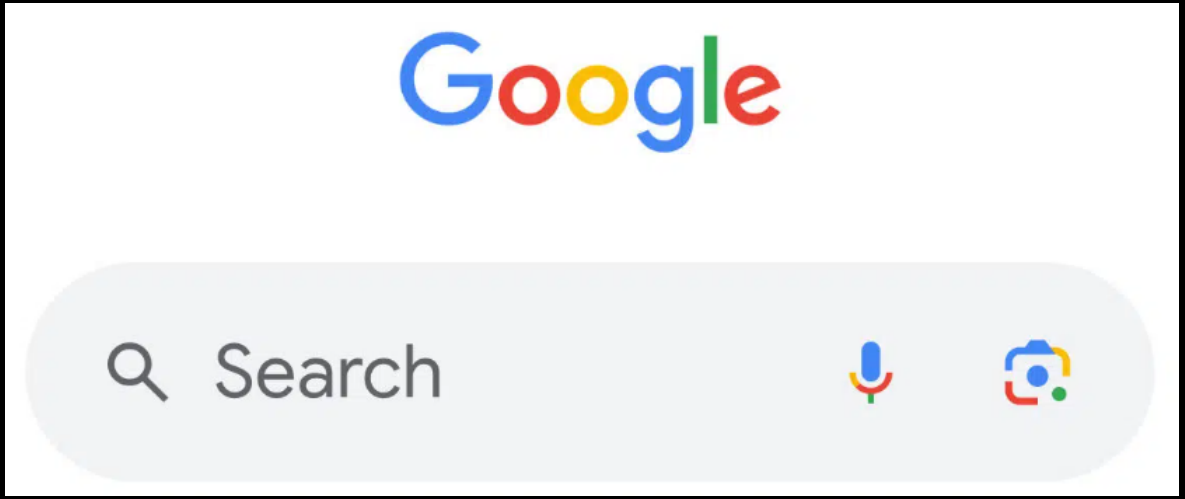
https://data-vault.com/data-stewards-gdpr-ambassadors-infographic/

Agenda

- ~~Data Quality~~
- Data Usability (Data Discoverability)
- Quiz session

Data Usability & Discoverability

- Data must be available (usable) and discoverable.
  - Data serves a purpose: “End users should have quick and reliable access to the data they need to do their jobs.”
  - “They should know where the data comes from, how it relates to other data, and what the data means.”



Metadata

- What is metadata?
  - Data about data
  - Why do we need metadata?

organized

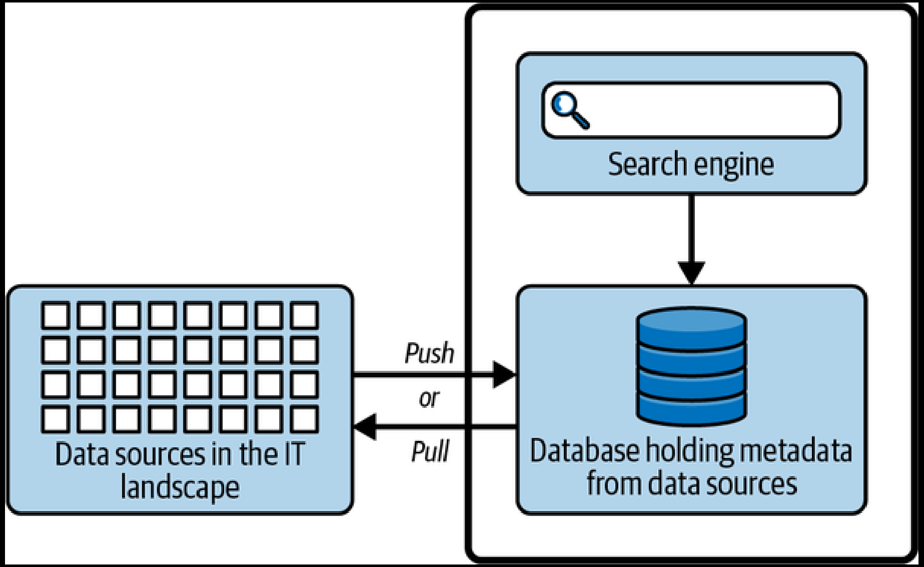
unorganized

Metadata categories

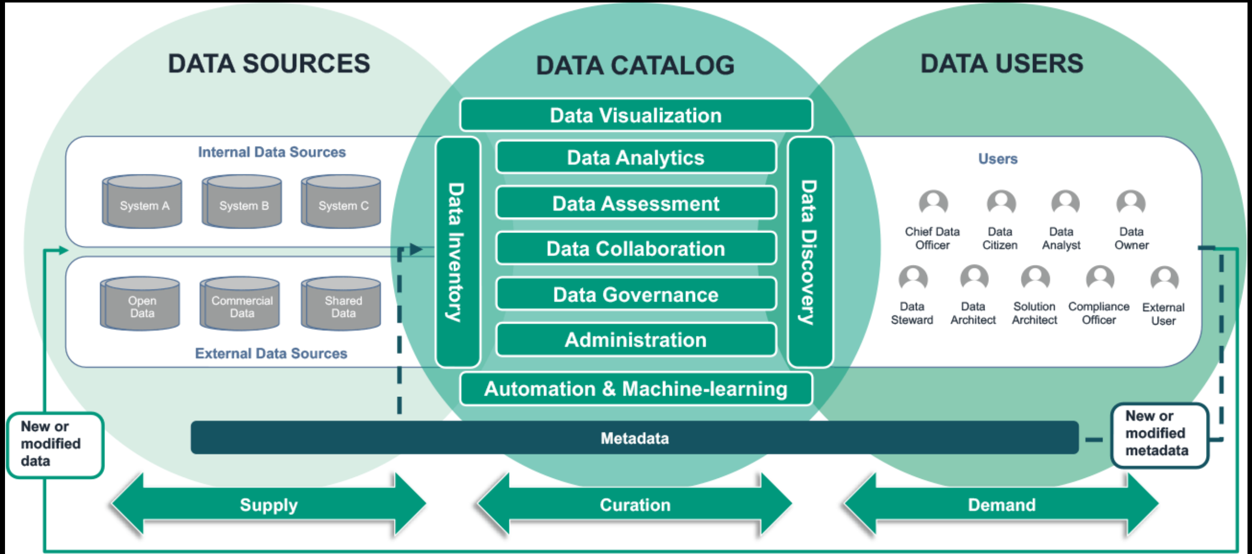
Category	Description	Examples
Business metadata	Relates to how data is used within the business context, including definitions, rules, and ownership.	• Data Definitions (who is a “customer”?) • Data Rules (phone number must be 8 digits) • Data Ownership (registered by, owned by, …)
Technical metadata	Describes the technical aspects of data, including structure, lineage, and system interactions.	• Schema Metadata (Database schemas, table structures) • Data Lineage (Tracking data origin and transformations) • Pipeline Metadata (Workflow schedules, system dependencies)
Operational metadata	Captures information about the operations and processes that handle the data, including performance and error logs.	• Process Statistics (Job IDs, execution times) • Runtime Logs (Application logs, error messages) • Process Status (Success, fail)
Reference metadata	Used to classify and standardize other data, often serving as lookup data for consistency.	• Geographic Codes (EE, LV, LT) • Units of Measurement (g, kg, tk) • Internal Codes (ATI)

Data Catalog

- An organized inventory of the data in your company.


Search engine
Data sources in the IT landscape
Push
or
Pull
Database holding metadata from data sources

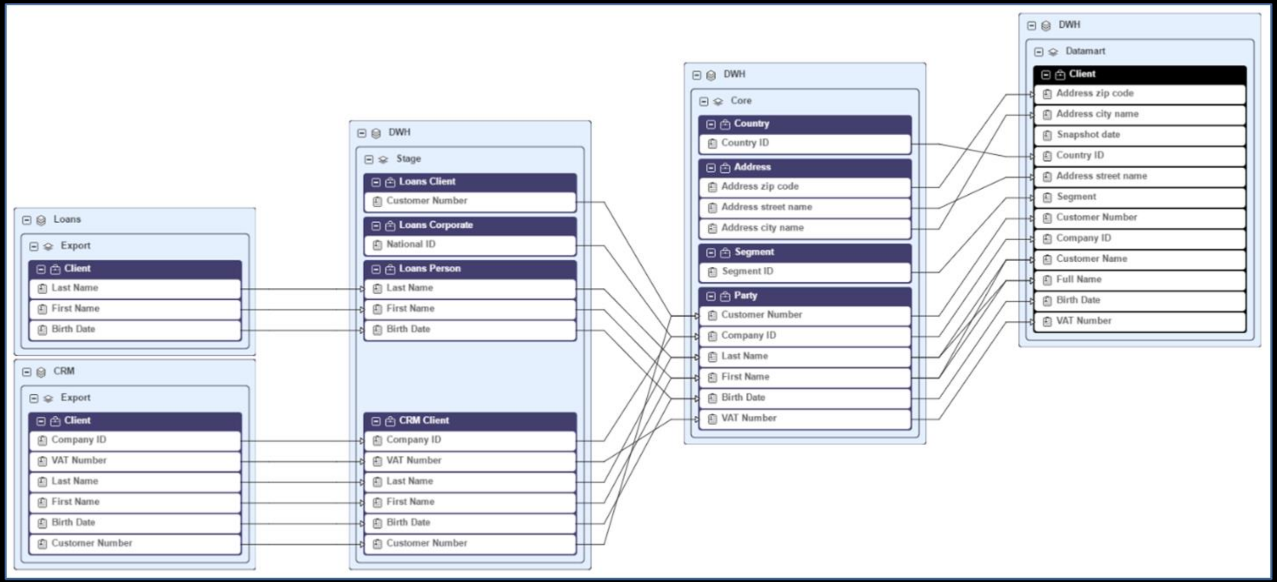
The Enterprise Data Catalog, Ole Olesen-Bagneux



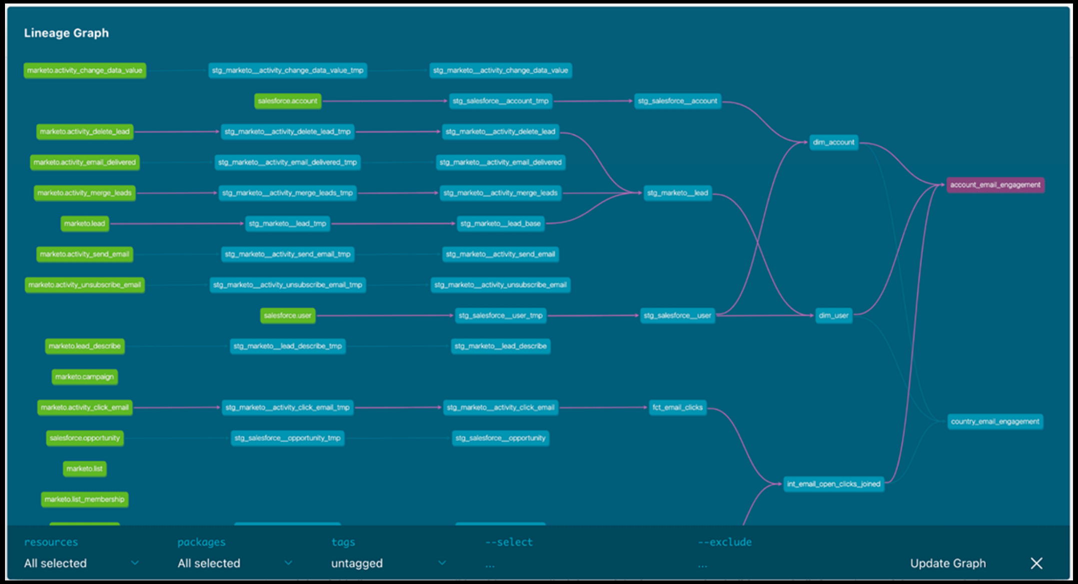
hevodata.com

Data Lineage

- Audit trail of data through its lifecycle
- Tracking across systems
  - Origins
  - Transformations
  - Consumers of data



datamanagement.wiki


Lineage Graph
marketo.activity_change_data_value
stg_marketo__activity_change_data_value_tmp
stg_marketo__activity_change_data_value
salesforce.account
stg_salesforce__account_tmp
stg_salesforce__account
marketo.activity_delete_lead
stg_marketo__activity_delete_lead_tmp
stg_marketo__activity_delete_lead
dim_account
marketo.activity_email_delivered
stg_marketo__activity_email_delivered_tmp
stg_marketo__activity_email_delivered
account_email_engagement
marketo.activity_merge_leads
stg_marketo__activity_merge_leads_tmp
stg_marketo__activity_merge_leads
stg_marketo__lead
marketo.lead
stg_marketo__lead_tmp
stg_marketo__lead_base
marketo.activity_send_email
stg_marketo__activity_send_email_tmp
stg_marketo__activity_send_email
marketo.activity_unsubscribe_email
stg_marketo__activity_unsubscribe_email_tmp
stg_marketo__activity_unsubscribe_email
salesforce.user
stg_salesforce__user_tmp
stg_salesforce__user
dim_user
marketo.lead_describe
stg_marketo__lead_describe_tmp
stg_marketo__lead_describe
marketo.campaign
marketo.activity_click_email
stg_marketo__activity_click_email_tmp
stg_marketo__activity_click_email
fct_email_clicks
salesforce.opportunity
stg_salesforce__opportunity_tmp
stg_salesforce__opportunity
country_email_engagement
marketo.list
marketo.list_membership
int_email_open_clicks_joined
Update Graph
✕

resources: All selected
packages: All selected
tags: untagged
--select: ...
--exclude: ...

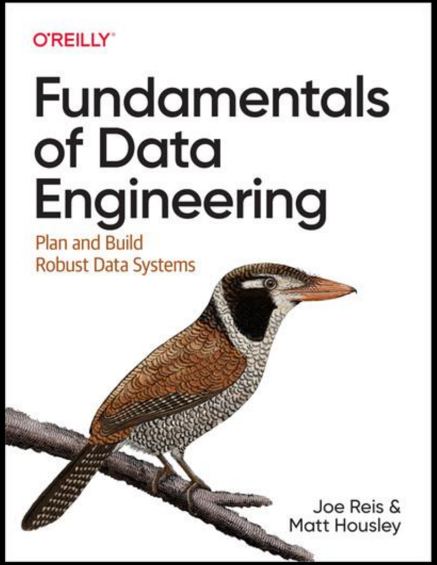
dbt

Agenda

- ~~Data Quality~~
- ~~Data Usability (Data Discoverability)~~
- Quiz session

Further reading

- Chapter II
  - Data Governance
- Book
  - The Enterprise Data Catalog
- Article
  - https://www.datacamp.com/cheat-sheet/data-governance-fundamentals-cheatsheet

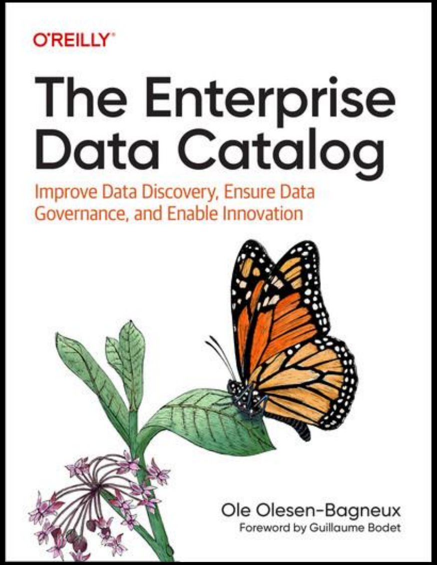


O'REILLY

Fundamentals of Data Engineering

Plan and Build Robust Data Systems

Joe Reis & Matt Housley



O'REILLY

The Enterprise Data Catalog

Improve Data Discovery, Ensure Data Governance, and Enable Innovation

Ole Olesen-Bagneux

Foreword by Guillaume Bodet