AIG150- Week 8

Data governance over a data life cycle

Reading Text: Ch 04-06

Data Governance: The definitive guide

Agenda

- Different types of data
- Phases of data life cycle
- Data life cycle management
- Data governance framework
- Data quality
- Data transformations

Different Data Types

- Transactional Data

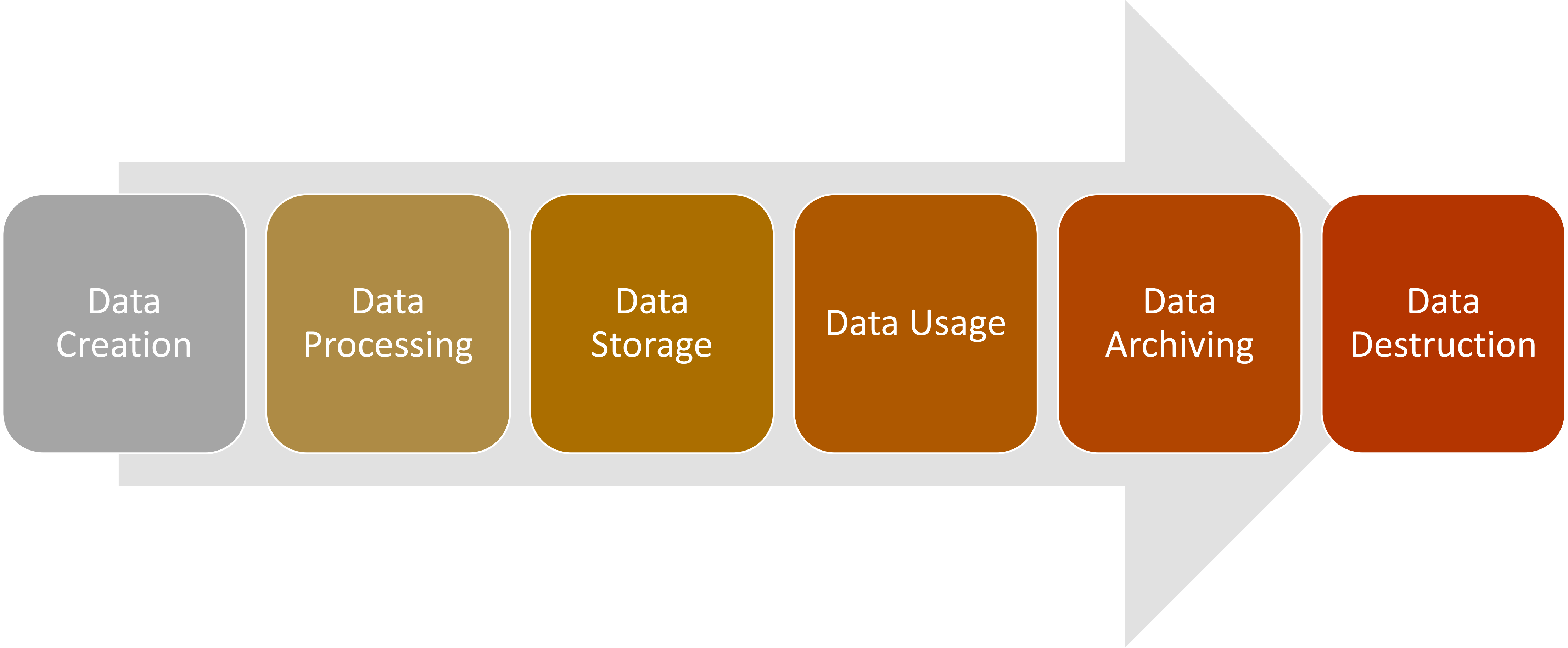
Databases that are optimized to run day-to-day transactional operations. These are fully optimized systems that allow for a high number of concurrent users and transaction types.

- Analytical Data

They are optimized to run analytical processes. These databases store historical data from various sources, including CRM, IOT sensors, logs, transactional data (sales, inventory), and many more. These systems allow data analysts, business analysts, and even executives to run queries and reports against the data stored in the analytic database.

They both have totally different life cycles.

Phases of a Data Life Cycle



Data Creation

Data Processing

Data Storage

Data Usage

Data Archiving

Data Destruction

Data Creation

- **Data is generated from multiple sources in multiple formats such as structured, semi-structured or unstructured.**

- **Data can be created by any one of the following three activities:**
  - Data acquisition
  - Data entry
  - Data capture

- **4 Vs of the data :volume, velocity, variety and veracity.**

- **Each way of generating data brings different types of data governance challenges.**

Data Processing

- Also referred to as data maintenance and involves the following activities:
  - Integration
  - Cleaning
  - Scrubbing
  - Extract-transform-load (ETL)

- Data governance implications:
  - Data lineage
  - Data quality
  - Data classification

Data Storage

- Data storage (both data and meta data) on different systems and devices with appropriate level of protection.
- Different storage mechanisms such as data mart, warehouse, data lake or data stores.
- Encryption techniques
- Backups and recovery

Data Usage

- Data is offered as a service or product by the organization.

- Proper access management and audits should be done to ensure proper and authorized access of data.

Data Archiving

- Data is removed from all active production systems and moved to warehouse for archiving.
- No maintenance or general usage occurs.
- Guidelines should be established as part of the data governance plan on the length of data retention and on the controls applied.

Data Destruction

Data Purging:

- Removal of every copy of data from the organization including the archives.

Data Life Cycle Management

- Comprehensive policy-based approach to manage the flow of data throughout its life cycle, from creation to the time when it becomes obsolete and is purged.

- When an organization can define and organize the life cycle processes and practices into repeatable steps for the company, this refers to DLM.

Data Management Plan

- A data management plan (DMP) defines how data will be managed, described, and stored.
- In addition, it defines standards you will use and how data will be handled and protected throughout its life cycle.
- Check DMPTool from MIT

Guidelines for DMP

- Identify the data to be captured or collected
- Define how the data will be organized
- Document a data storage and preservation strategy
- Define data policies
- Define roles and responsibilities

Data Governance Framework

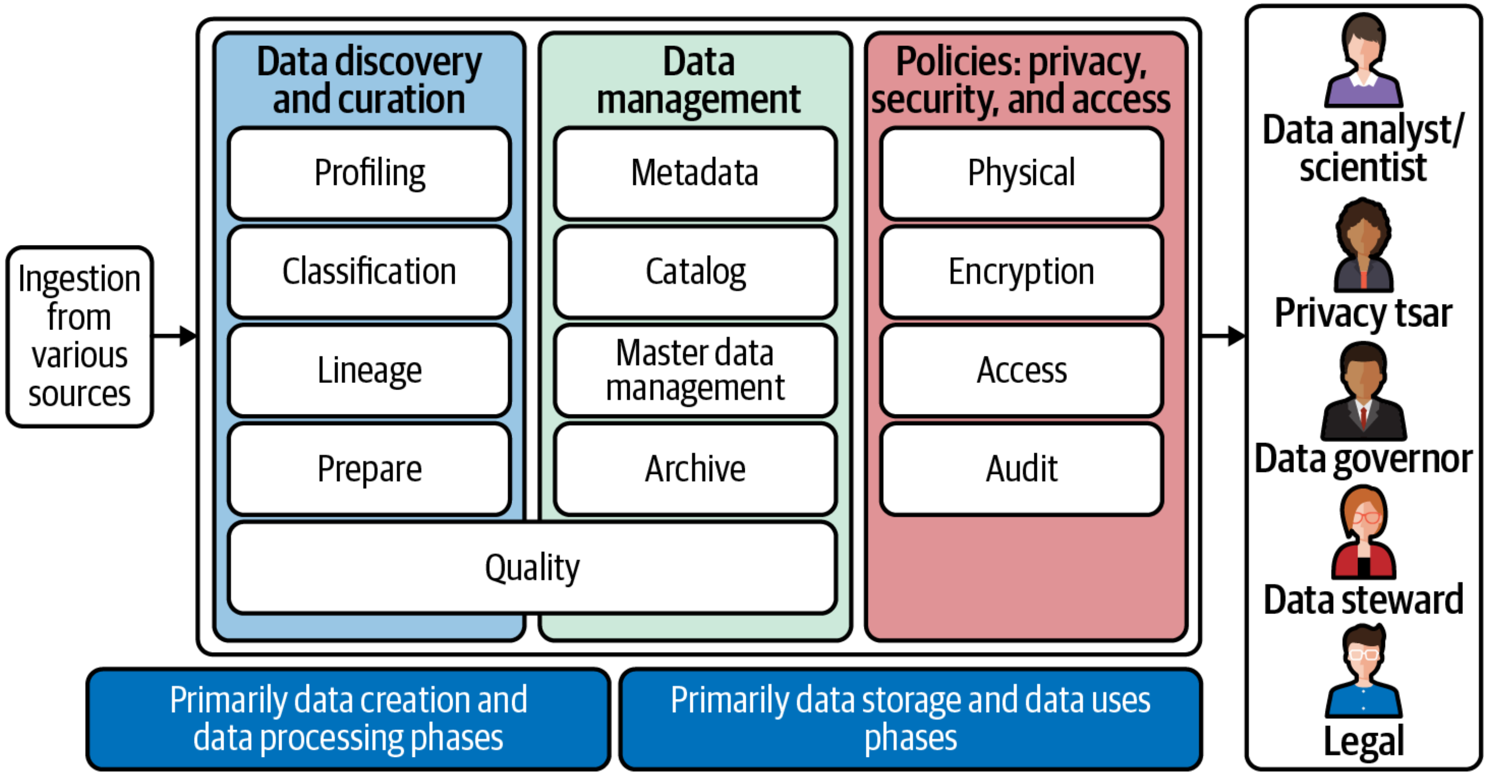
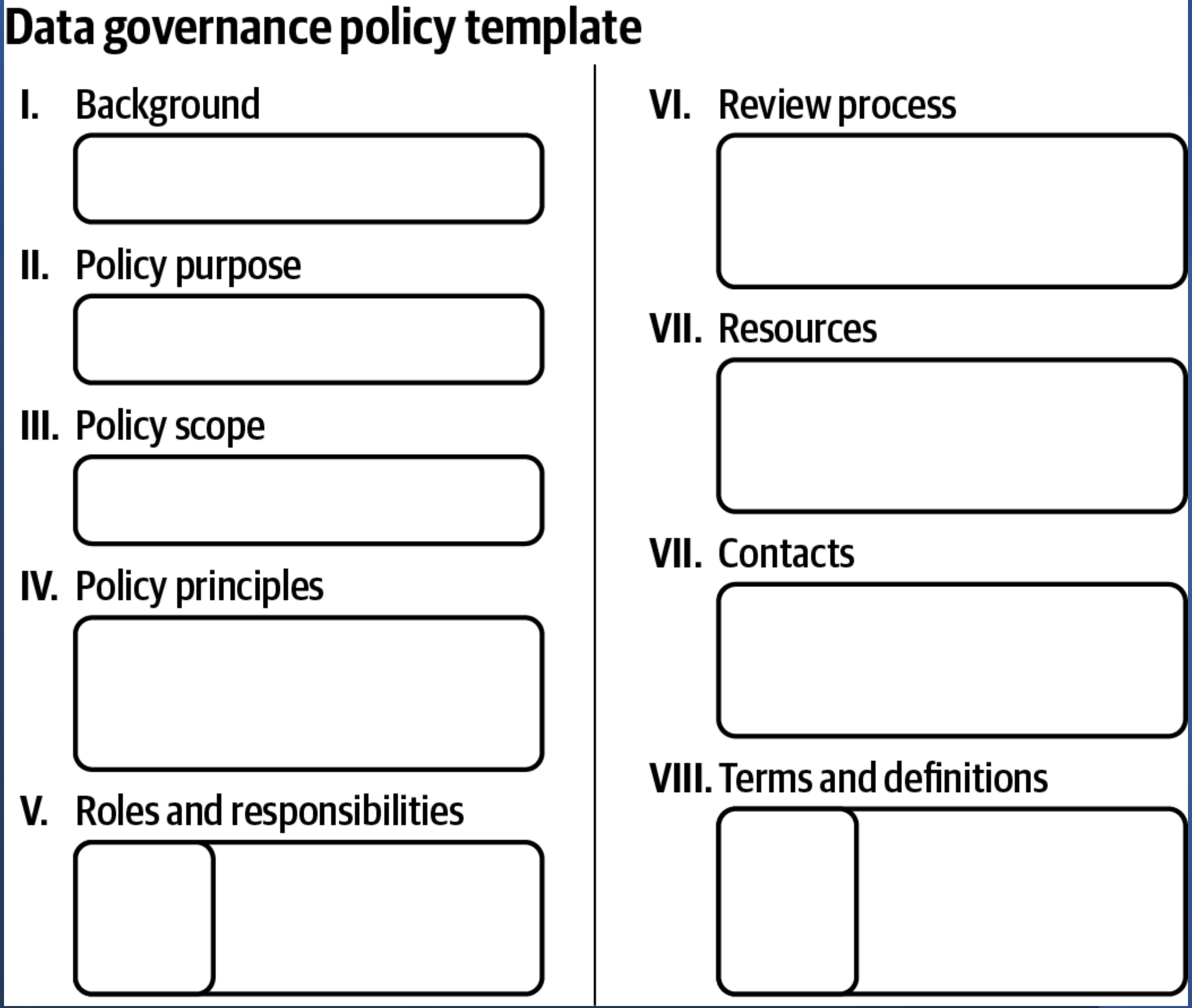


Figure 4-2. Governance over a data life cycle from the reference text

Discuss Case Studies Here

Why Data Governance Policy Is So Important ?

- Consistent, efficient, and effective management of the data assets throughout the organization and data life cycle and over time.
- The appropriate level of protection of the organization’s data assets based on their value and risk as determined by the data governance committee.
- The appropriate protection and security levels for different categories of data as established by the governance committee.



Data governance policy template

I. Background

II. Policy purpose

III. Policy scope

IV. Policy principles

V. Roles and responsibilities

VI. Review process

VII. Resources

VII. Contacts

VIII. Terms and definitions

Figure 4-5. Example data governance policy template from the reference text

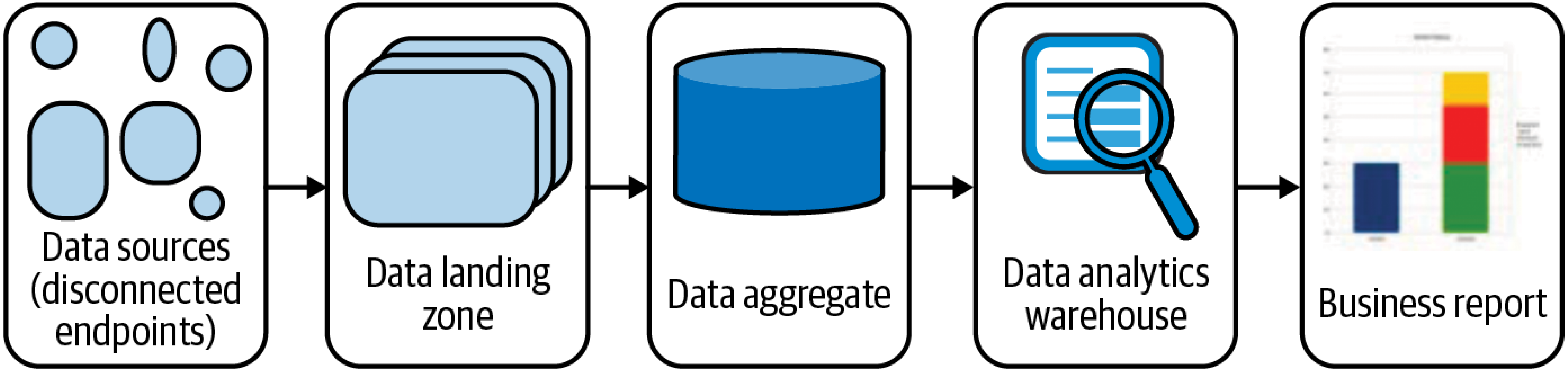
Consideration for Governance Across A Data Life Cycle

- Deployment time
- Cost and complexity
- Changing regulation environment
- Location of data
- Organizational structure

How to Define Data Quality ?

- Accuracy
- Completeness
- Timeliness
- Easily affected by outliers
- Source trustworthiness

Why Is Data Quality Important?



Data sources (disconnected endpoints)

Data landing zone

Data aggregate

Data analytics warehouse

Business report

From Reference Text
Figure 5-1. Simple data acquisition chain

Data Quality Is Different

- Big Data Analytics
    - ETL
    - Data is rarely updated but is refreshed periodically, read-only mode.
    - Support decision making processes

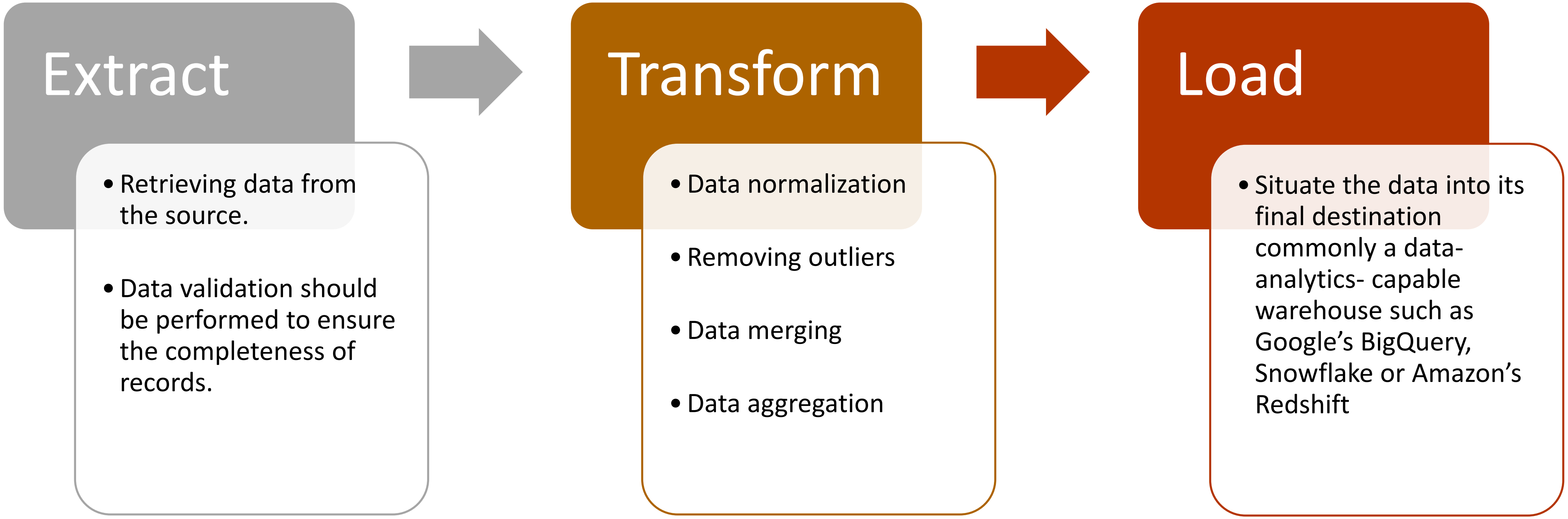
- AI\ML Models
    - Data should be error-free to help organization make right predictions
    - Data is divided into training, test and validation data sets.
    - Data should be free of biases.
    - Data should be generalized

Data Quality should be part of data governance program.

Techniques For Data Quality

- Scorecard
- Prioritization
- Annotation
- Profiling
- Deduplication
- Outliers
- Lineage tracking
- Completeness
- Merging datasets

Data Transformations



Extract

- Retrieving data from the source.
- Data validation should be performed to ensure the completeness of records.

Transform

- Data normalization
- Removing outliers
- Data merging
- Data aggregation

Load

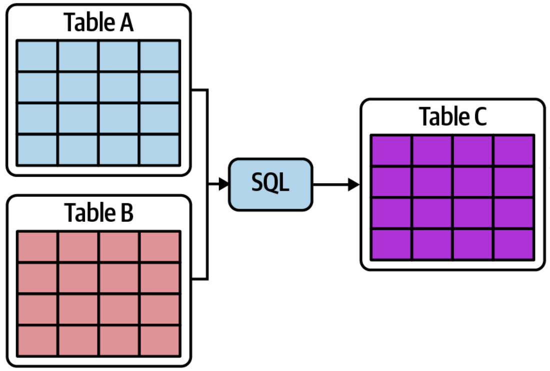
- Situate the data into its final destination commonly a data-analytics- capable warehouse such as Google’s BigQuery, Snowflake or Amazon’s Redshift

Lineage

- Lineage, or *provenance*, is the recording of the "path" that data takes as it travels through extract-transform-load, and other movement of data, as new datasets and tables are created, discarded, restored, and generally used throughout the data life cycle.

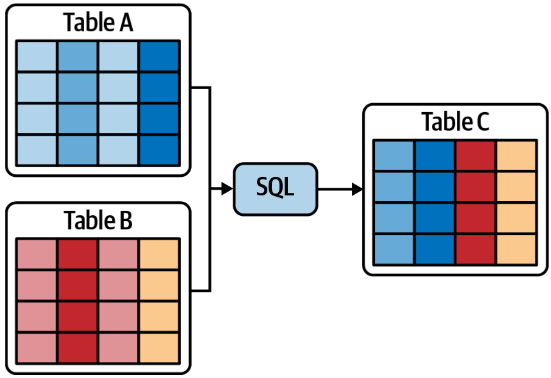
- Lineage can be a visual representation of the data origins (creation, transformation, import) and should help answer the questions "Why does this dataset exist?" and "Where did the data come from?"

Table Level Lineage



"Table C is a product of Tables A and B"

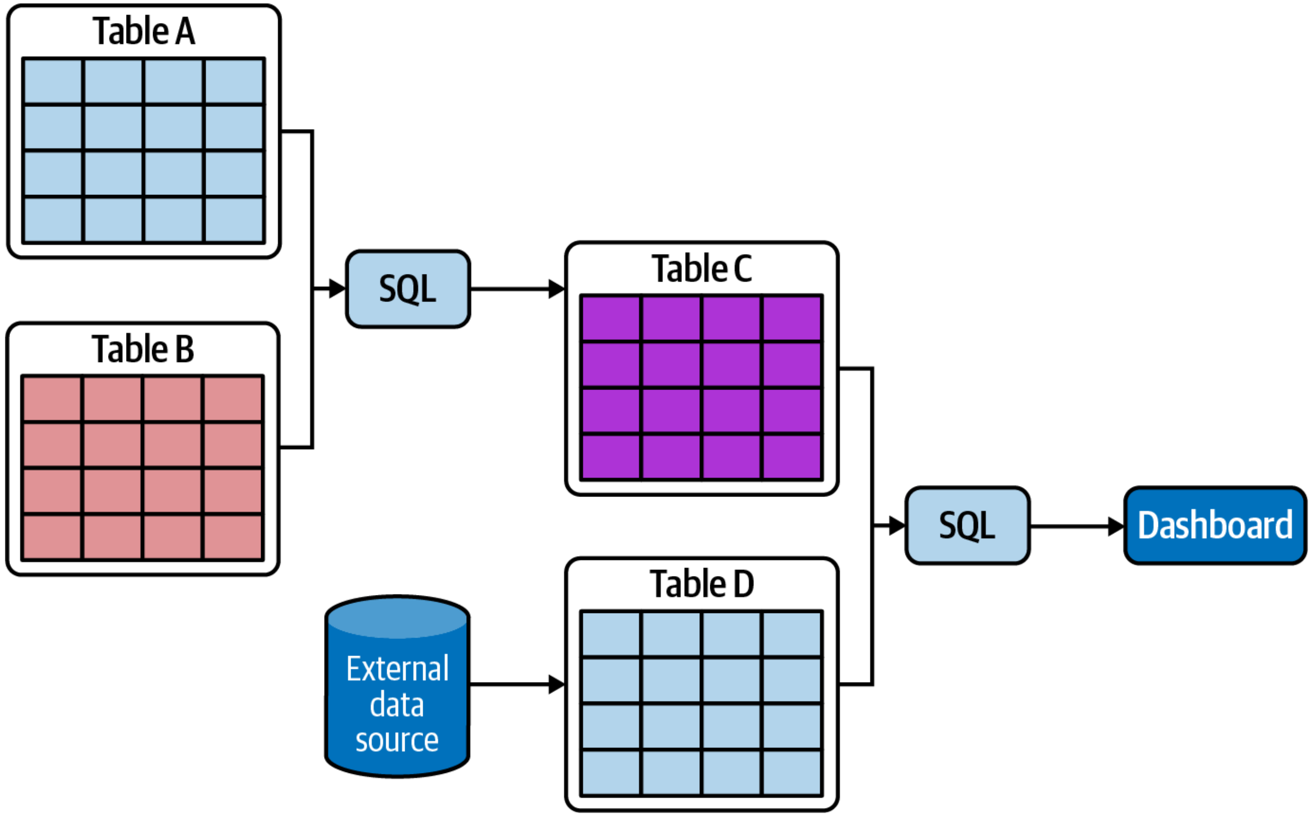
Column Level Lineage



- Columns 1 and 2 in Table C come from Table A (source columns 2, 4)
- Columns 3 and 4 in Table C come from Table B (source columns 2, 4)

Images taken from reference text

How to Govern Data ?



Reference Text
Figure 6-4. Lineage workflow—if Table B contains sensitive data, that data can potentially be found in the dashboard as well

Lineage Is Useful …

- Policy management
- Simulation
- Monitoring
- Change Management
- Audit and compliance