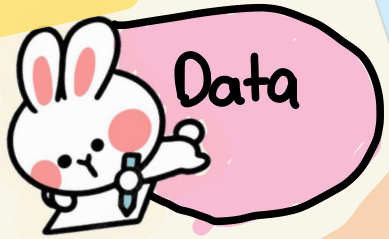
Data

Types

- Structured
  - predefined model
  - fields / tables / columns / spreadsheet
  - Numerical [Interval] [Ratio]
  - Categorical [nominal] [Ordinal]
- Unstructured
  - no field / attribute
  - binary files – audio / video / images
  - [Textual]
  - [Multimedia]
  - [XML / JSON]
- Semi-structured
  - have markers / tags to seperate elements
  - xml / HTML / email / webpage

Challenges

- Silos
- Availability
- Technology
- Quality
- Chaos



Database

- Impact
  - Business
  - Service industry
  - Education
  - Personalized application
- Background
  - entity
  - attribute
  - relationship
  - key (primary, secondary, foreign)
  - chen / crow's foot notation
- Concept
  - model
  - ER diag.
  - SQL
  - DDL
  - DML

Pipeline

Data source → Collection → Processing → Storage → Consumption → Governance ↕ monitoring

ETL

Extracting

1. extract from sources
2. Reading & comprehension
3. Copying & transfer

Issues
- source ident.
- method extrac.
- extraction freq.
- time window
- job sequence.
- exception handling

Transformation

1. Selection
2. Splitting / joining
3. Conversion
4. Summarization
5. Enrichment

Loading

* Initial load
* Incremental load
* Full refresh

- Suitable for monolithic, legacy data sources

Data warehouse → enterprise DW
→ data mart → dependent
→ independent
↓ metadata repo

Component
- OLAP
- Source system
- End user → data-mining tools, reporting tools, statistical tools
- Data staging → extraction, transform [clean, integ., aggreg.], Loading

Data cube → 2D, 3D

OLAP → Operation
- drill through
- drill across
- roll up (drill-up)
- Drill down (roll down)
- slice dice
- pivot (rotate)

Schema
- Star Schema
- Fact constellation schema
- Snowflake Schema