# Data Warehousing

Jens Teubner, TU Dortmund
jens.teubner@cs.tu-dortmund.de

Summer 2018

# Part II

# Overview

# Data Warehouse

So what **is** a data warehouse?

- A data warehouse is a **database**
  - → Typically a rather **large** one
  - → Think of multiple terabytes ∼ a few petabytes
- Data warehouses are **tuned for analytics**

✎ **“Tune”?**

# OLTP ↔ OLAP

**OLTP** (**O**nline **T**ransaction **P**rocessing):

- Day-to-day business operations
  - → Mix of insert, update, delete, and read operations
  - → *e.g.*, enter orders, maintain customer data, etc.
- System sometimes called **operational data store (ODS)**
  - → Up-to-date state of the data
- From a database perspective:
  - → **Short-running** operations
  - → Most queries known in advance
  - → Often **point access**, usually through indexes
  - → write access ⇝  ACID principles

# OLTP ↔ OLAP

**OLAP** (**O**nline **A**nalytical **P**rocessing):

- Provide data for **reporting** and **decision making**
  - → Mostly **read-only** access
  - → *e.g.*, resource planning, marketing initiatives
- Need **archive data**; (slightly) outdated information might be okay
  - → Report **changes over time**
  - → Can use separate data store (non-ODS)
- From a database perspective:
  - → **Long-running** operations, mostly **read-only**
  - → Queries not known in advance, often complex (⇝ indexing?)
  - → Might need to **scan** through large amounts of data
  - → Data is (almost) **append-only**.

# Transactional vs. Analytical Workloads

| | OLTP | ODS | OLAP | DM / DW |
|---|---|---|---|---|
| Business Focus | Operational | Oper./Tact. | Tactical | Tact./Strat. |
| DB Technology | Relational | Relational | Cubic | Relational |
| Transaction Count | Large | Medium | Small | Small |
| Transaction Size | Small | Medium | Medium | Large |
| Transaction Time | Short | Medium | Medium | Long |
| DB Size in GB | 10–400 | 100–800 | 100–800 | 800–80,000 |
| Data Modeling | Trad. ERD | Trad. ERD | N/A | Dimensional |
| Normalization | 3–5 NF | 3 NF | N/A | 0 NF |

source: Bert Scalzo. *Oracle DBA Guide to Data Warehousing and Star Schemas*.

# Data Warehouse

**Typically:**

- **One** large repository for **entire company**
- Dedicated **hard- and software**
  - Enterprise-grade DBMS
  - Often: **database appliances** (*e.g.*, Teradata, Oracle Exadata, IBM Netezza, ...)

**Goal:**

- Single source of truth for analysis and reporting

Requires **data cleansing** and **conflict resolution**

# Example: Oracle Exadata X4-8

- 240 CPU cores, up to 12 TB RAM **per rack**
- 44.8 TB “Exadata Smart Flash Cache”
- up to 672 TB per rack raw disk capacity (300 TB usable)
- InfiniBand 40 Gb/s interconnect
- data load rate: 20 TB/hour

# Full Data Warehouse Architecture

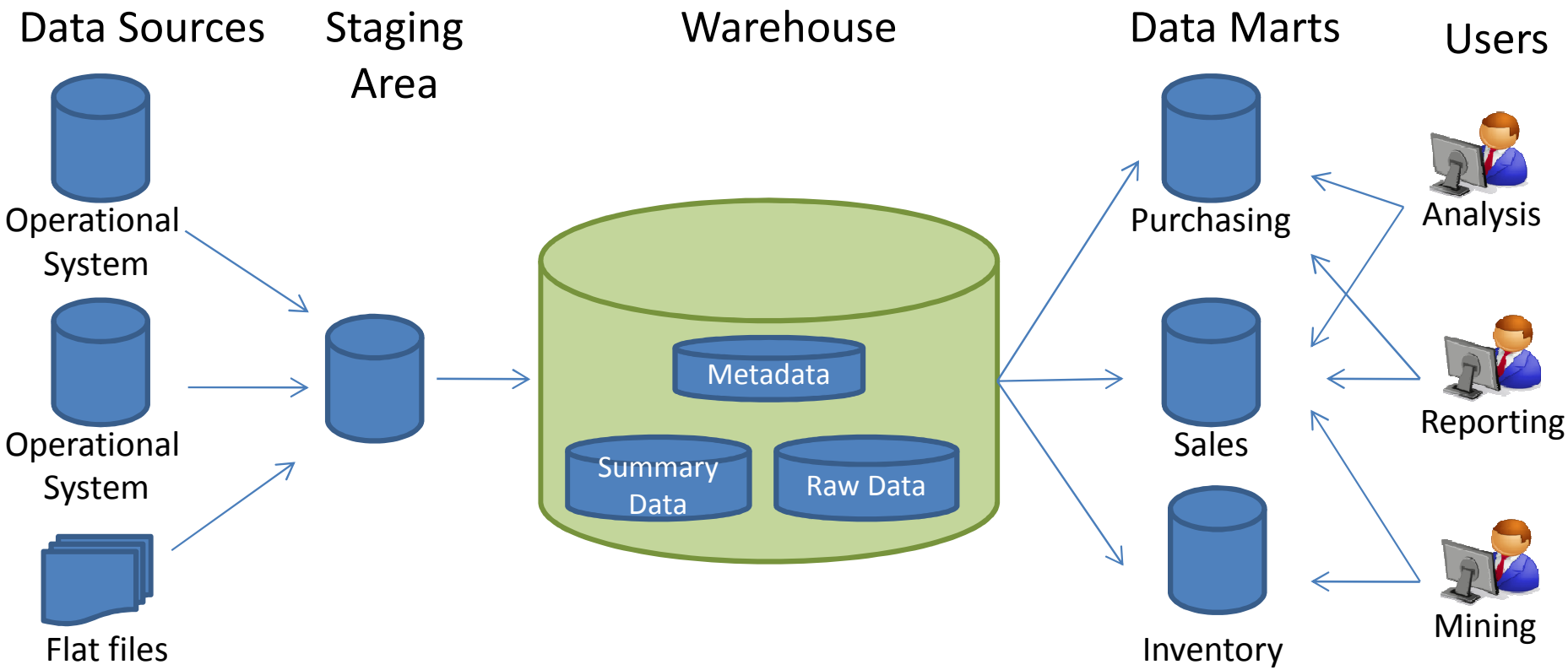



image source: Wolf-Tilo Balke. *Data Warehousing & Mining Techniques*.

# Architecture Alternatives

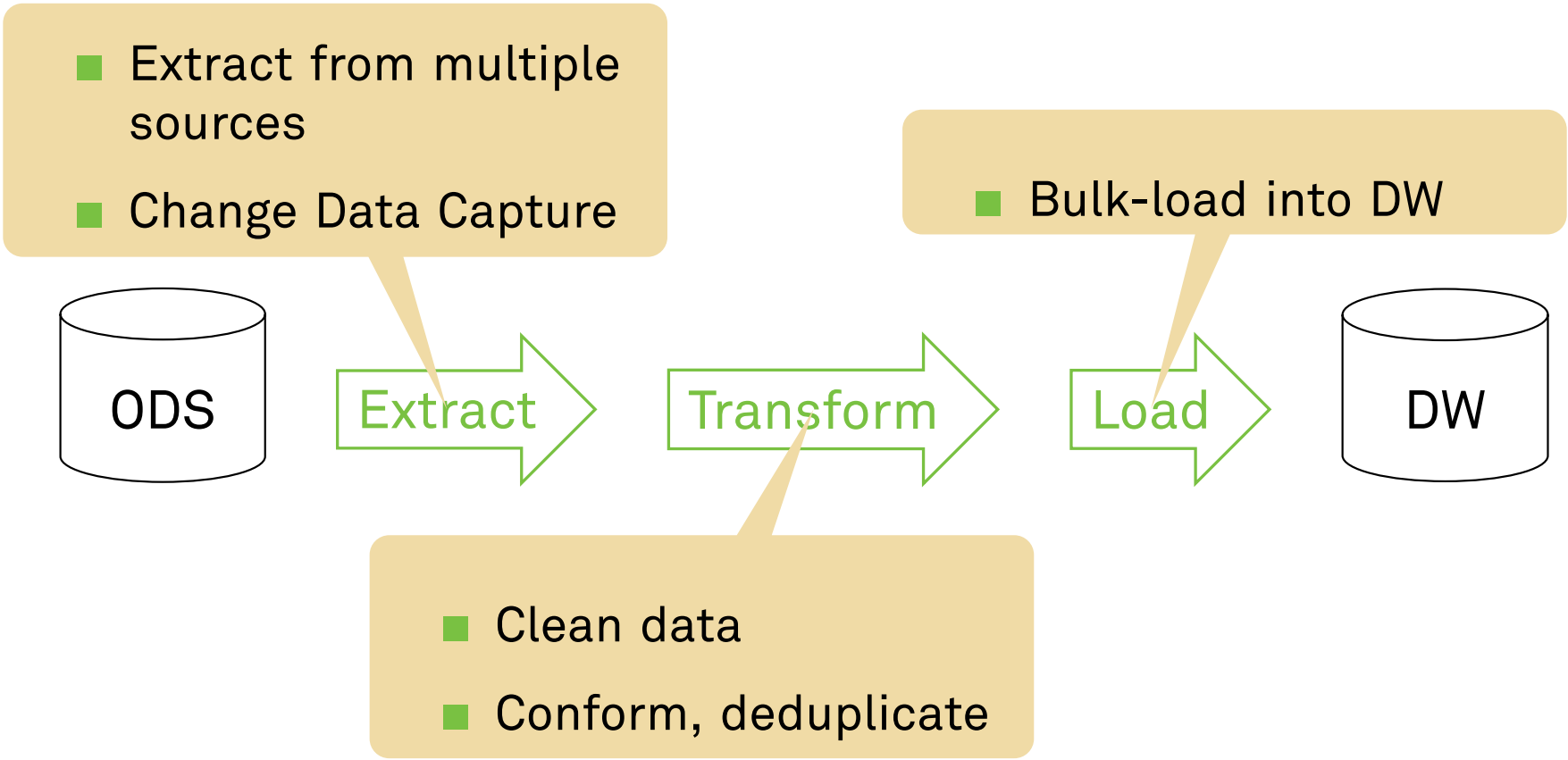
Variants of the full data warehouse architecture:

1. **Independent data marts** (no central warehouse)
   - Populate data marts directly from sources
   - Like several “mini warehouses”
   - Redundancy, no “single source of truth”
2. **Logical data marts** (no explicit, physical data marts)
   - Data mart just a logical view on full warehouse
   - Easier to provide integrated, consistent view across the enterprise

→ Data marts (and warehouse) might also reside at **different geographic locations.**

# Data Warehouse Architecture

Data is periodically brought from the ODS to the data warehouse.




→ This is also referred to as **ETL Process.**

# Data Warehouse Users

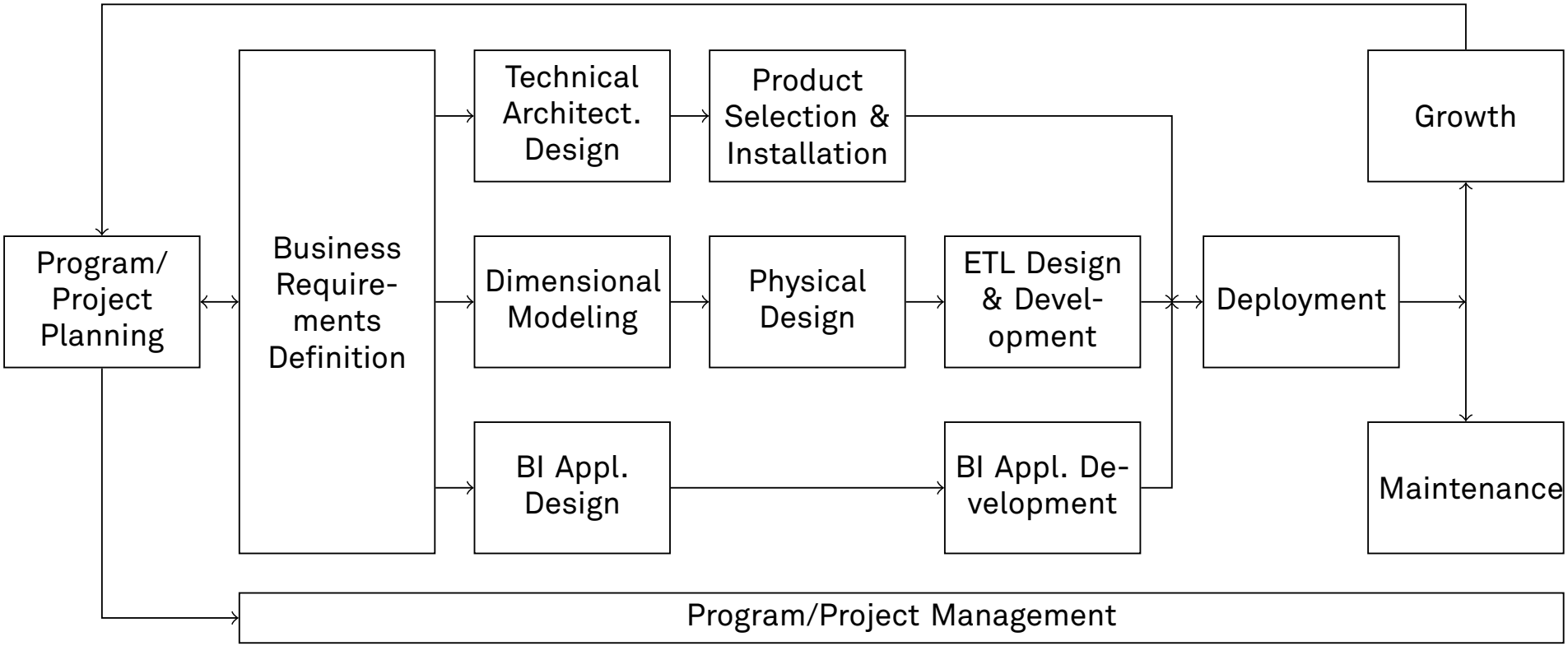
**Business analysts:**

- **Explore** data to **discover** information
- Use for **decision making**
  - → **“Decision Support System (DSS)”**

**Consequences:**

- Workloads and access patterns **not known in advance**
- For exploration, data representation must be **easy to understand** (even by business analysts)
- Design and usage driven by **data**, not applications

# Data Warehouse Lifecycle




↗ Kimball *et al.* The Data Warehouse Lifecycle Toolkit.