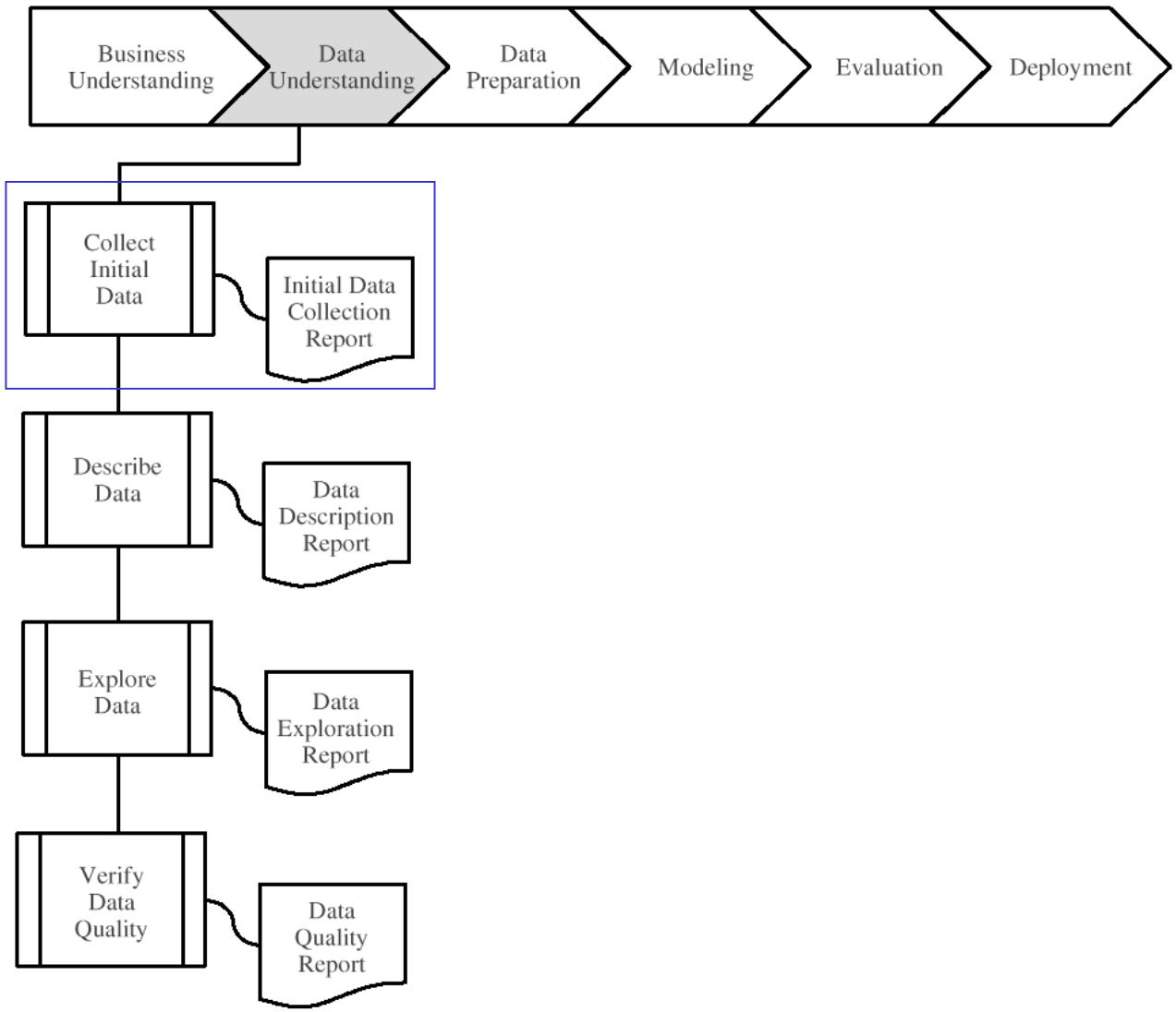
Kapitel II

Datenbereitstellung

II. Datenbereitstellung



II. Datenbereitstellung

- **Collect Initial Data**
  - identify relevant attributes
  - identify inconsistencies between sources
- **Describe Data**
  - characterize attributes (relevance, statistical characteristics, ...)
- **Explore Data**
  - querying, visualization
- **Verify Data Quality**
  - identify errors in data
  - number of missing values
    - identify false encodings of missing values (e.g. 1.11.[19]11 as birthday)

II.1 Grundlagen

- Anspruch an **Datenbanken** in Unternehmen ist vielschichtig.
- Man kann sie - je nach Einsatzzweck - in **operative** und **informationelle Systeme** einteilen.

    - **Operative Systeme**
        - eingesetzt von Sachbearbeitern, am Bankschalter, etc.
        - dienen der täglichen Arbeit

    - **Informationelle Systeme**
        - helfen dem Management, (strategische) Entscheidungen zu finden.
          (durch DSS = Decision Support Systems)
          (Systeme für Business Intelligence Anwendungen)
        - bieten Grundlage für weitere Analysen mit OLAP / Data Mining

- **Informationelle Systeme**

  - zugeschnitten auf **Gegenstandsbereiche** (sog. ***Subjects***), z.B.
    - Kunde,
    - Produkt,
    - Vertriebsregion

  - unterstützen **Informations- und Analyseaufgaben**,
    d.h. das Management in der Entscheidungsfindung

  - wenige Zugriffe, aber mit relativ **hohem Datenvolumen**

  - Datenbankeinträge werden nicht geändert (**keine Updates**)

  - Antwortzeitverhalten spielt untergeordnete Rolle

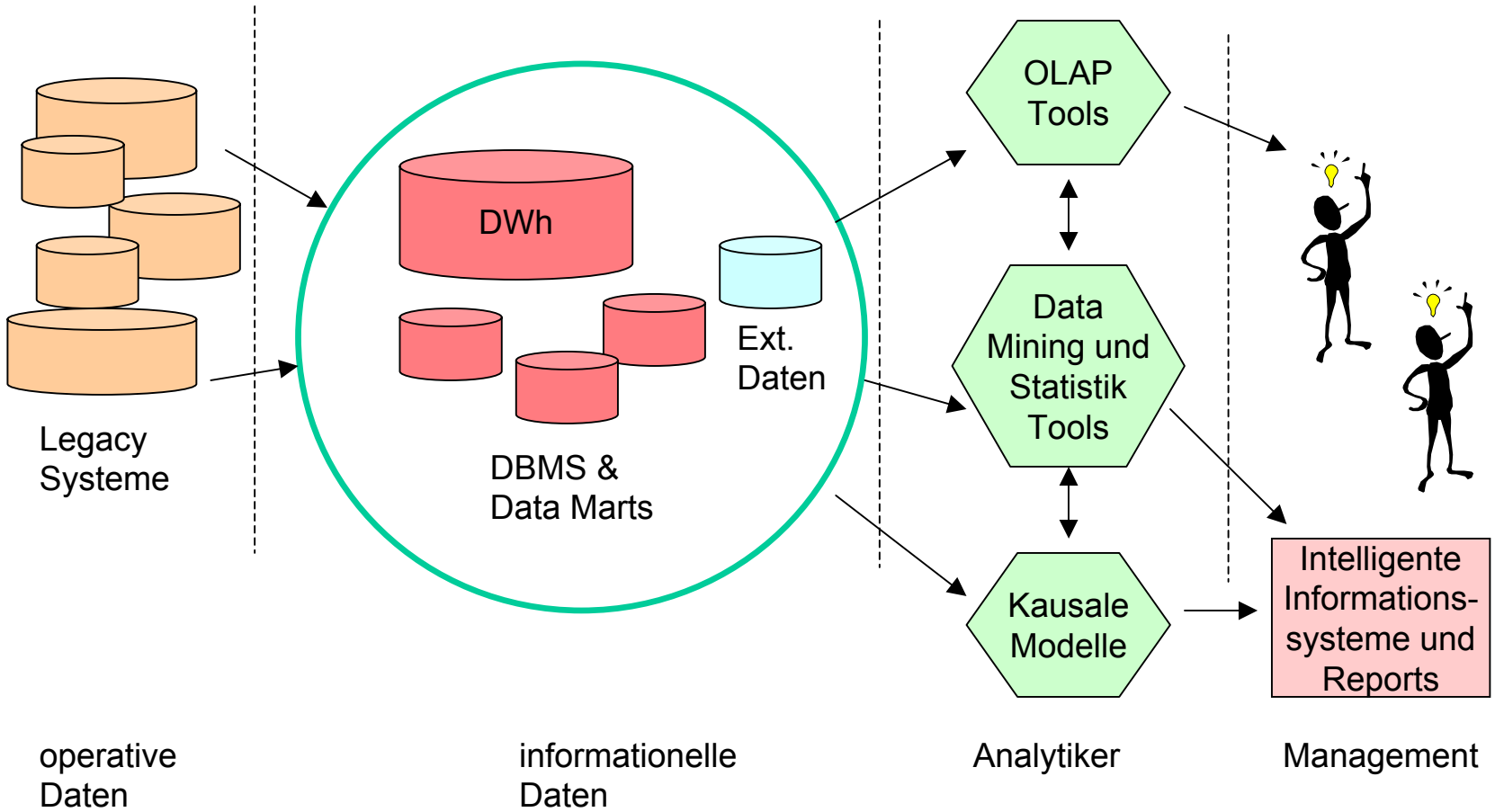
- **Informationelle Systeme** (forts.)
  - enthalten sehr große Datenmengen
  - enthalten zum großen Teil **historische, zusammengefasste Daten**
    - Historie aus Daten der operativen Systeme ist nachvollziehbar
  - relativ hohe **Redundanz**
  - Überblick über alle relevanten Unternehmensdaten
  - **komplexe,** oft heuristische **Ad-hoc-Anfragen**
    - z.B. auf der Basis von OLAP-Funktionalitäten
  - Daten sind **wohl strukturiert, integriert** und **konsolidiert**
  - Anzahl Benutzer ist eher klein („Power-User“)

Aus diesen Charakteristiken ergeben sich **fundamentale Gegensätze:**

Operative Systeme	Informationelle Systeme
• Schnelle Antwortzeit • Anwendungsorientiert • Aktuelle Daten • Detaillierte, primäre Daten • Häufige Änderungen • Dient täglicher Arbeit	• Hohe Speicherkapazität • Gegenstandsorientiert • Historische Daten • Auch zusammengefasste, abgeleitete Daten • Keine Updates • Dienst als Datenspeicher für Analyse und Entscheidungsfindung

⇒ Man muss beide Systeme trennen.

Data Warehouse für den informationellen Systemteil


DWh
Ext.
Daten
Legacy
Systeme
DBMS &
Data Marts
OLAP
Tools
Data
Mining und
Statistik
Tools
Kausale
Modelle
Intelligente
Informations-
systeme und
Reports
operative
Daten
informationelle
Daten
Analytiker
Management

II.2 Was ist ein Data Warehouse?

- „Mit dem Begriff Data Warehouse wird eine von den operationalen DV-Systemen isolierte Datenbank umschrieben, die als unternehmensweite Datenbasis für Management-Unterstützungssysteme dient.“ [Muksch et al. 1996]

- „A Data Warehouse is a
  - subject-oriented,
  - integrated,
  - time-variant,
  - nonvolatile

  collection of data in support of management‘s decision-making process.“

[Inmon, Hackathron 1994]

A Data Warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management‘s decision-making process.

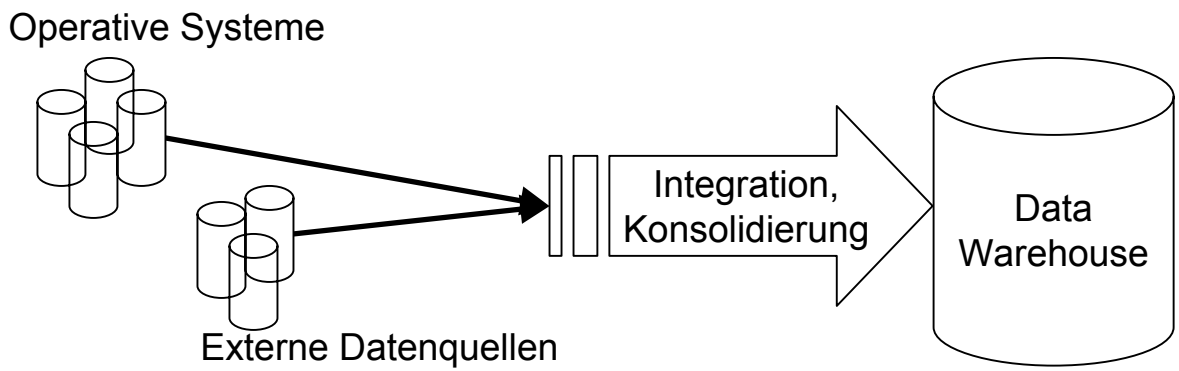
[Inmon, Hackathron 1994]

II.2.1 Gegenstandsorientierung (subject-oriented)

- DWh ist an **Gegenstandsbereichen** des Unternehmens orientiert,
  - z.B. Produkten, Kunden, Lieferanten
- Gegensatz zu **Funktions-** oder **Anwendungsorientierung** bei operativen (**legacy**) Systemen:
  - Funktionen sind z.B. Einkauf, Lagerhaltung, Verkauf
- Bei der Entwicklung eines DWh stehen die **Daten im Mittelpunkt**.
  - Bei operationalen Systemen muss auch der Prozess berücksichtigt werden.
- DWh enthält nur solche Daten, die für DSS-Analysten/Manager relevant und interessant sind, werden oder sein könnten.

> A Data Warehouse is a **subject-oriented**, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process.
>
> [Inmon, Hackathron 1994]

II.2.2 Integration



- In zwei verschiedenen operativen Systemen können
  - die **gleichen Daten** unter **verschiedenen Bezeichnern** abgelegt sein
  - die **gleichen Bezeichner** für **verschiedene Zwecke** benutzt werden
  - der **gleiche Sachverhalt** auf **verschiedene Weise** kodiert sein

A Data Warehouse is a subject-oriented, **integrated**, time-variant, nonvolatile collection of data in support of management's decision-making process.

[Inmon, Hackathron 1994]

II.2.2 Integration

- Daten aus **verschiedenen Quellen** werden im DWh vereinheitlicht, u.a. durch
  - **konsistente** Vergabe und **Definition** von Bezeichnern
  - **einheitliche Kodierung**
    - z.B. wird jedes Datum in der Form <YYYY-MM-DD> gespeichert
  - **einheitliches** Festlegen der **Maßeinheiten** von Attributen
    - z.B. werden Preise in Dollar angegeben
  - Auflösung von **strukturellen** Konflikten
    - z.B. Schema-Wert-Konflikt

- Integration führt dazu, dass alle Daten im DWh in einer einzigen, **allgemein akzeptierten Art und Weise** gespeichert sind.

- Erst die Integration erlaubt die einfache und effektive Nutzung der DWh-Daten für Anwendungen z.B. im Management

- Integration ist ein **schwieriger** und **zeitaufwendiger** Prozess

Behandlung von Strukturkonflikten in relationalen Schemata: [Saltor et. al. 1993]

- **Beispiel:** Datenbank für Aktienkurse
  - Datenbank New York (ein Tupel pro Tag und Aktie)

date	stock	clsprice
991008	IBM	347
991008	HP	418
991008	GM	250
991009	IBM	350
991009	HP	420
991009	GM	215

  - Datenbank Barcelona (ein Tupel pro Tag, ein Attribut pro Aktie)

date	HP	IBM	GM
991008	418	365	250
991009	420	350	200

  - Datenbank Melbourne (eine Relation pro Aktie, ein Tupel pro Tag)

HP	date	clsprice
	991008	425
	991009	420

IBM	date	clsprice
	910408	347
	910409	350

GM	date	clsprice
	991008	385
	991009	320

II.2.3 Zeitraumbezug (time variancy)

- In **operativen Systemen** ist der **aktuelle Datenbestand** gespeichert. Er kann jederzeit geändert werden (**Update**).
- DWh enthält eine ganze **Historie von Daten**
- DWh besteht aus **Snapshots** der operativen Systeme
- DWh-Daten sind zu einem bestimmten Zeitpunkt gültig (gewesen). Der **Gültigkeitszeitraum** ist an allen Daten im DWh vermerkt (als Teil des Schlüssels)
- **Zeithorizont** des DWh: ca. 5-10 Jahre
  - operative Systeme: max. 60-90 Tage

A Data Warehouse is a subject-oriented, integrated, **time-variant**, nonvolatile collection of data in support of management's decision-making process.

[Inmon, Hackathron 1994]

II.2.4 Beständigkeit (nonvolatility)

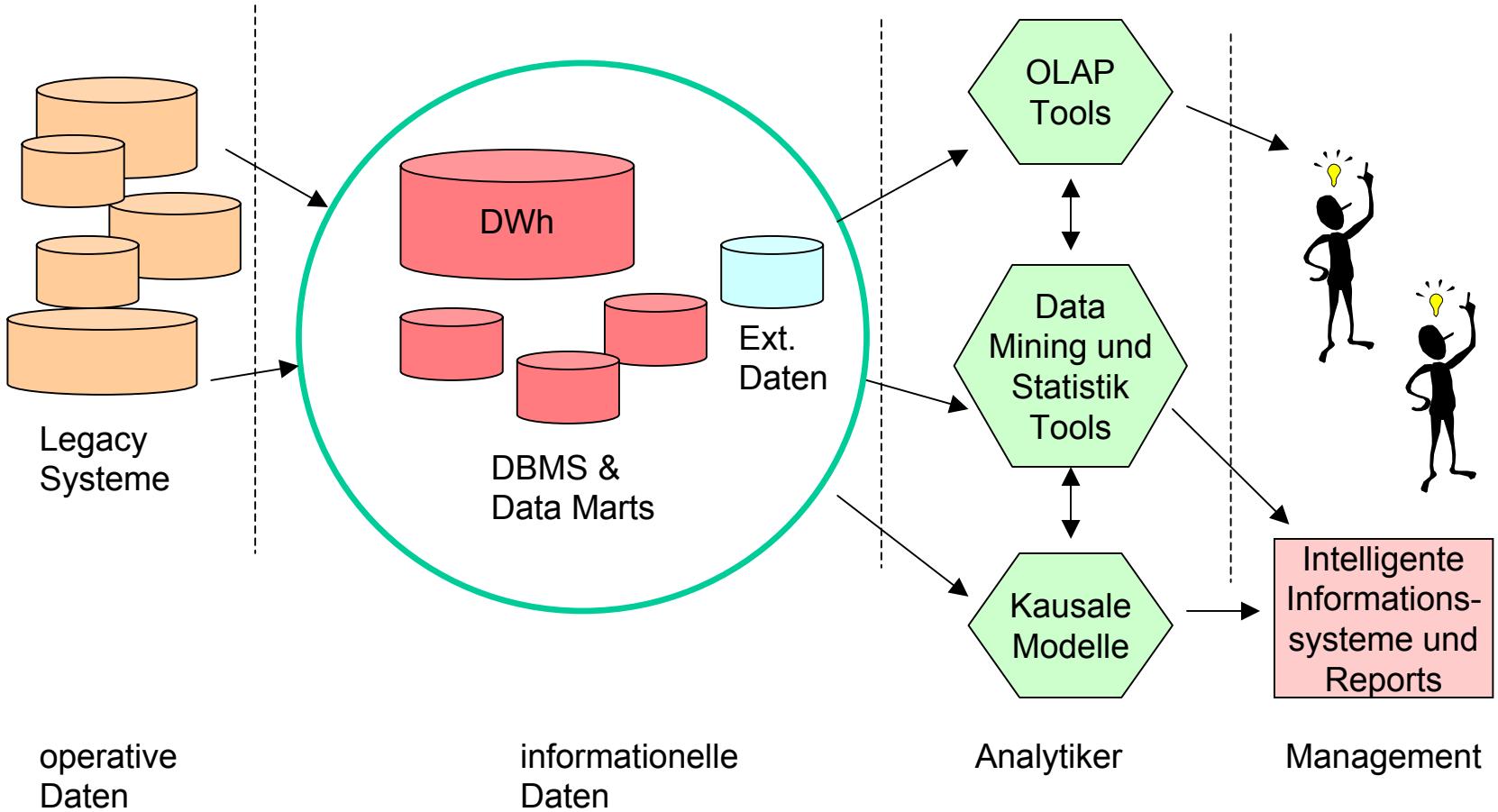
- **Operative Systeme:** Daten werden oft geändert, gelöscht, eingefügt.
  - Aufwendige Mechanismen, um Deadlocks zu vermeiden
  - Locking-Mechanismen, etc.

- **DWh**: primär nur **Leseoperationen**
  - Daten werden aus den operativen Systemen (initial) **geladen**
  - Analysesysteme greifen **lesend** auf DWh-Daten zu.
  - Es gibt **keine Updates**

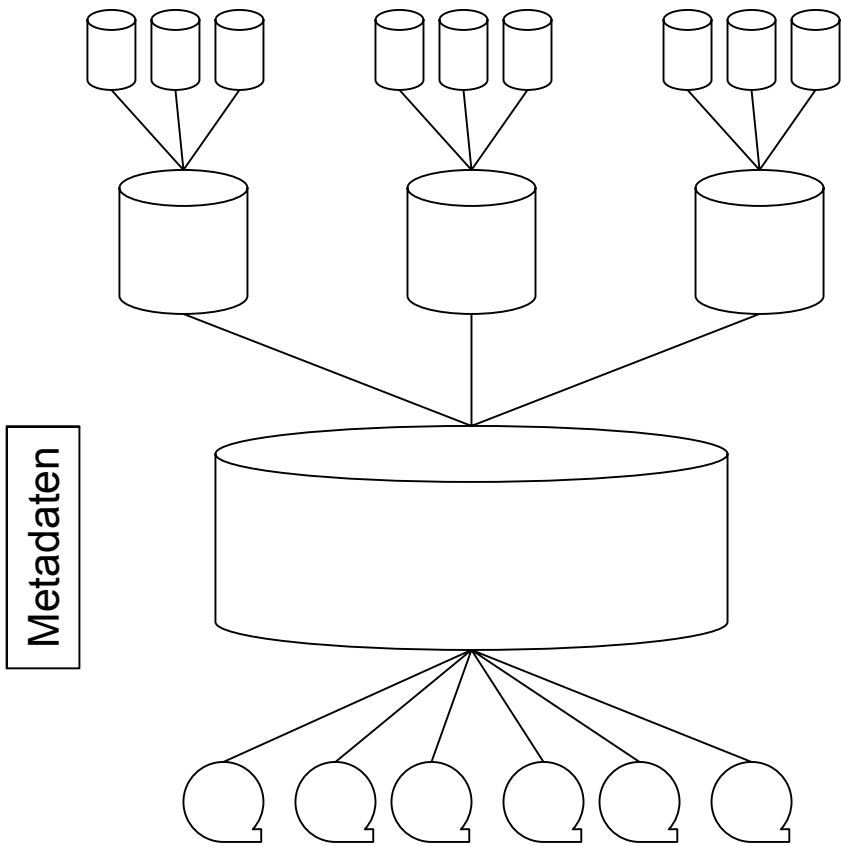
A Data Warehouse is a subject-oriented, integrated, time-variant, **nonvolatile** collection of data in support of management's decision-making process.

[Inmon, Hackathron 1994]

II.3 Architektur einer Data Warehouse Umgebung



Die innere Struktur eines Data Warehouse



Stark zusammengefasste Daten

Leicht zusammengefasste Daten

Aktuelle detaillierte Daten

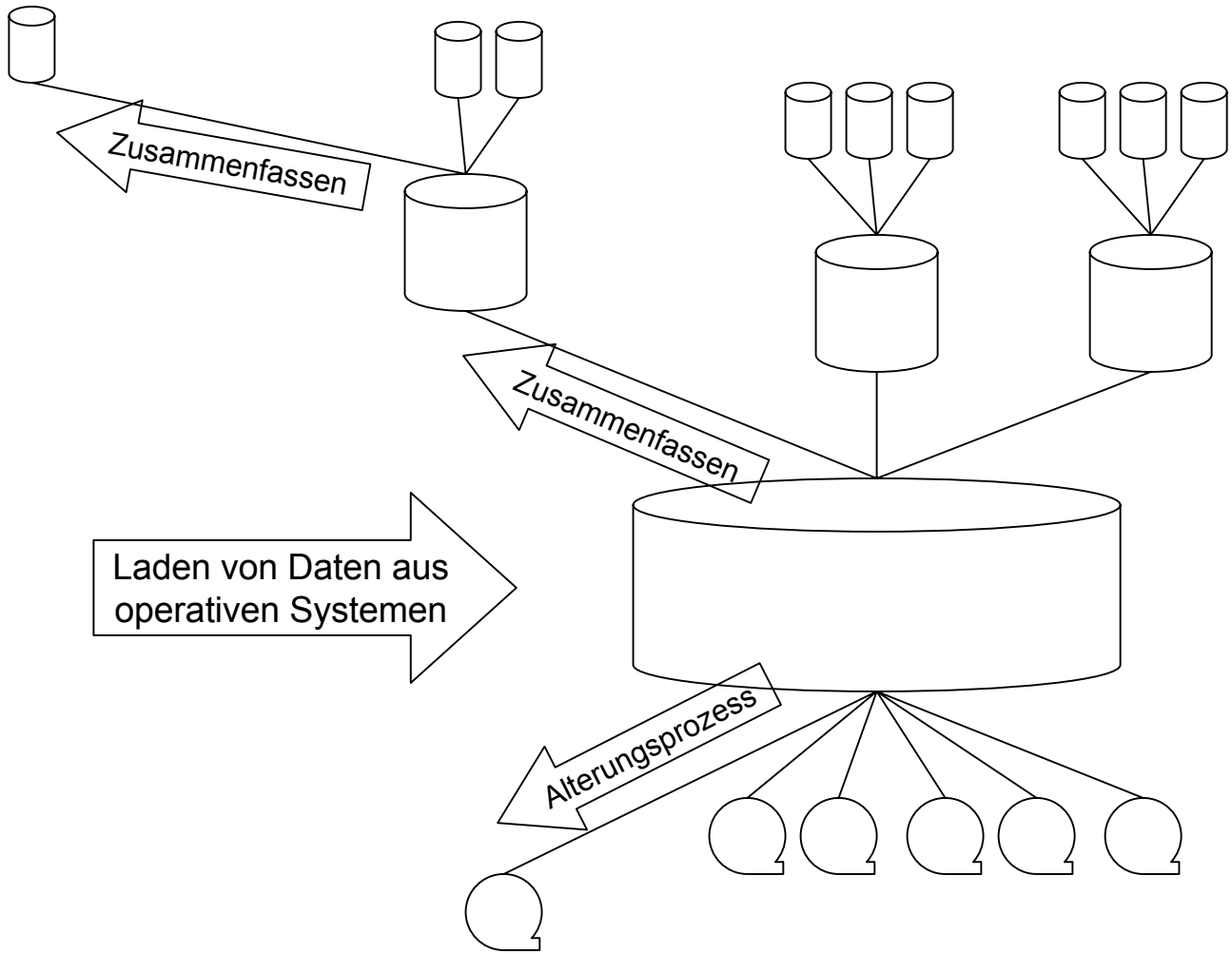
Alte (detaillierte) Daten

Beispiel: Telekommunikationsunternehmen

detailliert	leicht zusammengefasst

• Für jeden Kunden, jedes Gespräch inkl.
  • Zone
  • Teilnehmer
  • Zeitpunkt
  • Dauer
  • Gebühren
  • Art des Dienstes
∅ 45.000 Byte pro Kunde
	• Für jeden Kunden
	  • Anzahl der Gespräche insgesamt
	  • Anzahl der Ferngespräche
	  • ∅ Gesprächsdauer
	  • Umsatz je Zone
	  • Umsatz insgesamt
	
	Zusammenfassung monatlich
	ca. 200 Byte pro Kunde

Datenflüsse im Data Warehouse



Metadaten

- Metadaten sind Daten über Daten
- Metadaten lassen sich in Kategorien einteilen:

  - **semantische** Metadaten
    - Festlegung der DWh-**Terminologie**
    - **Transformations**- und **Integrationsregeln** für die Abbildung der operativen Daten in die DWh-Daten
    - **Aggregationsregeln** für das Zusammenfassen der Daten auf verschiedenen Aggregationsstufen

Metadaten (forts.):

- **verwaltungstechnische** Metadaten
  - Festlegung von **Benutzer (-gruppen)** und zugehörige **Zugriffsrechte**
  - **statische** Daten über das DWh
    - Größe von Tabellen
    - Zugriffsrechte auf Tabellen

- **schematische** Metadaten
  - **logisches** Schema des DWh
  - Abbildung zwischen logischem und physischem Schema
  - Quellen der DWh-Daten

DWh-Entwicklungszyklus unterscheidet sich vom klassischen:

- Am Anfang des Data-Warehouse-Entwicklungszyklus stehen die Daten (der Prozess ist ***datengeleitet (data-driven)***)

- Das Data Warehouse wird **schrittweise** entwickelt.

- <u>Gründe:</u>
  - genaue Ziele/ Anforderungen an das DWh sind meistens noch nicht bekannt, Größe auch schlecht abschätzbar
  - Kosten und Entwicklungszeit schlecht abschätzbar
  - benötigte Ressourcen (Mitarbeiter, Rechner, ...) sind hoch

Iterative Vorgehensweise

- **iteratives Vorgehen** und **kurze *feedback loops*** haben viele Vorteile:
  - Anwender können ihre Anforderungen erst dann detailliert artikulieren, wenn der erste DWh-Prototyp vorliegt (1. Stufe der DWh-Entwicklung)
  - Management wird erst dann größeres Projektbudget genehmigen, wenn positive Resultate sicher greifbar sind.
  - Qualität des DWh wird durch feedback loops mit Anwendern deutlich verbessert.

$\Rightarrow$ Leitmotiv: **Think big! Start small! Grow step by step!**

Monitoring der DWh-Benutzung

- Monitoring ist Voraussetzung für Anpassung des DWh an aktuelle Nutzung
  - Welche Daten des DWh werden regelmäßig genutzt?
  - In welchem Umfang wächst der Datenbestand?
  - Wer benutzt das DWh?
  - Welche Antwortzeiten treten bei welchen Anfragen auf?
  - Wie ist die Belastung des DWh?