

## TDWI Europe 2010 Munich

### Architektur und Modellierung für das Enterprise Data Warehouse

W2A

*Mittwoch, 16. Juni 2010*

*10:00 bis 11:30 & 12:00 bis 13:15*

*Dr. Michael Hahne*

**hahne consulting gmbh**

IT solutions for your business

## Agenda

- Motivation
- Konzeptionelle Modellierung des Reporting Layers
- Zeitabhängigkeit im Reporting Layer am Beispiel Star Schema
- Staging und Core Data Warehouse Modellierung
- Architektur für Enterprise Data Warehousing

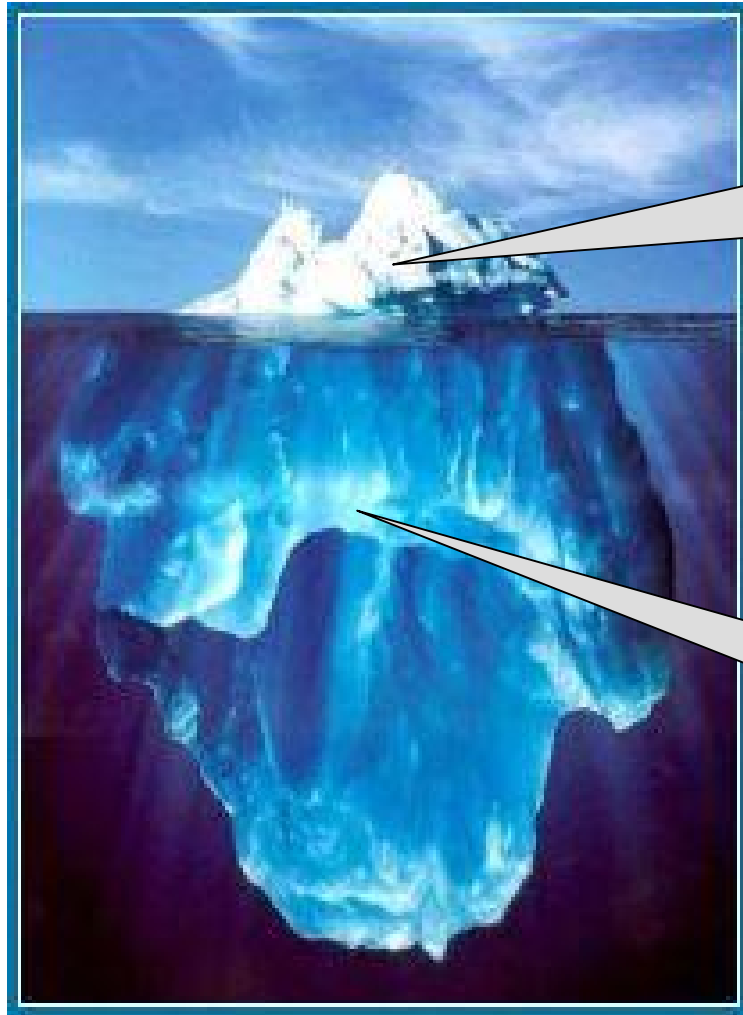
## Agenda

- **Motivation**
- Konzeptionelle Modellierung des Reporting Layers
- Zeitabhängigkeit im Reporting Layer am Beispiel Star Schema
- Staging und Core Data Warehouse Modellierung
- Architektur für Enterprise Data Warehousing

## Typische Data Warehouse Probleme

- “We Can’t Meet our Batch Windows”
  - KPI und Berichtsbereitstellung
  - Backup der Datenbank
  - Neuaufbau im Data Warehouse
- “Our Costs are Spiraling”
  - Storage / Hardware / Replikation
  - Rechenzentrumsbetrieb (Raum / Energie / Klimatechnik)
  - Datenadministration
- “The Targets Keep Changing”
  - Neue Geschäftsanforderungen
  - Spezifische Projektanforderungen
  - Externe Anforderungen / Audit / Innenrevision

# Information as Corporate Asset – We Do not Know What we not Know...



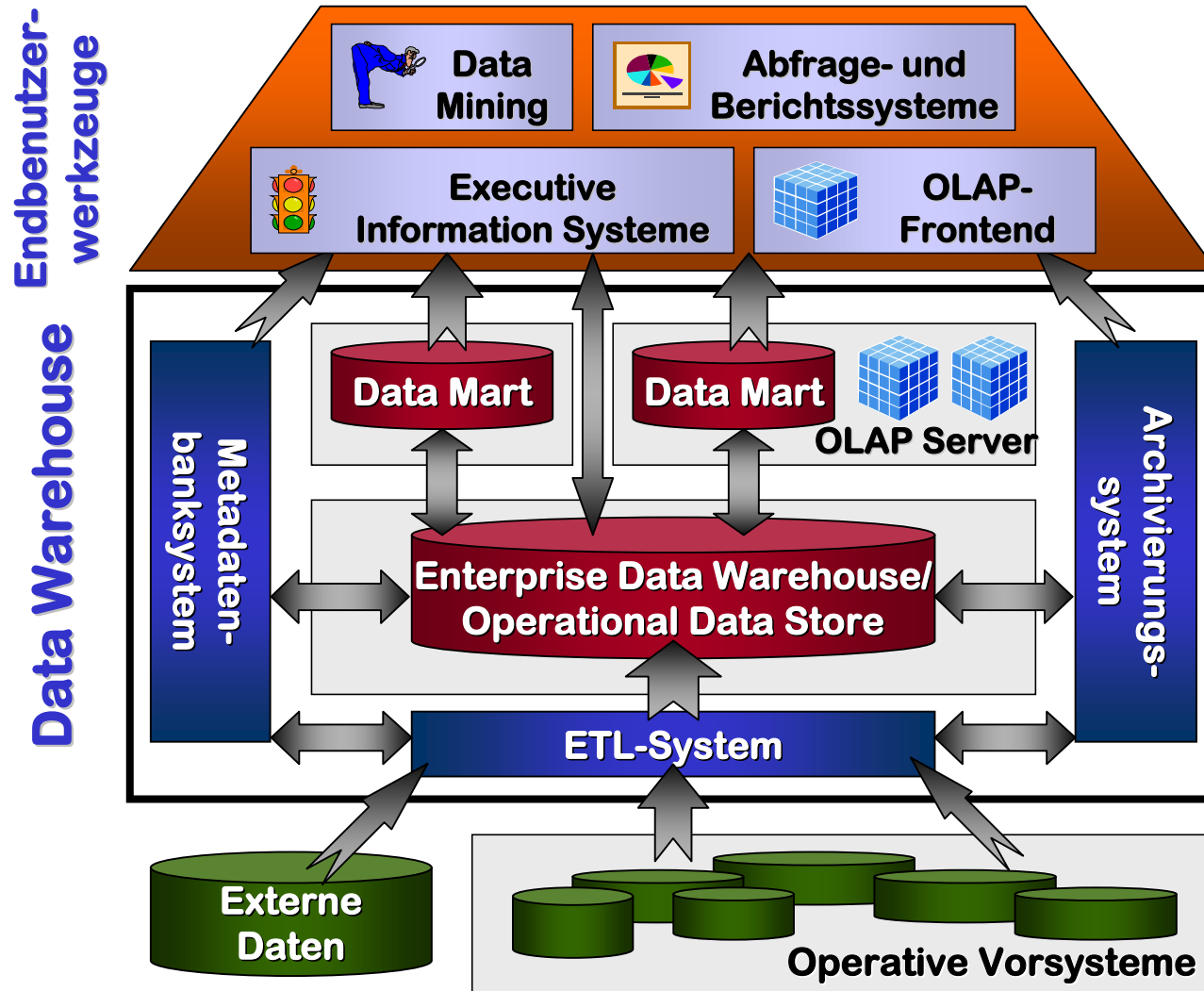
## **The Known**

Current BI implementations are set up to answer known requirements

## **The Unknown**

Little or nothing is done to be prepared for unpredictable future information needs

# Aktuelles DW-Architektur-Konzept



## Agenda

- Motivation
- **Konzeptionelle Modellierung des Reporting Layers**
- Zeitabhängigkeit im Reporting Layer am Beispiel Star Schema
- Staging und Core Data Warehouse Modellierung
- Architektur für Enterprise Data Warehousing

# Semantische Data Mart Modellierung mit ADAPT


## *A*pplication *D*esign for *A*nalytical *P*rocessing *T*echnologies


## ADAPT Modelliermethode

- Entwickelt von Dan Bulos, Symmetry Corporation
- Unabhängig von Technologie und Architektur auf semantischer Modellierebene zur Modellierung von OLAP Systemen
- Zielsetzung
  - Kommunikation (mit Fachanwender)
  - Konzeptionelles Framework
  - Frühe Problemidentifikation
  - Verbesserte Qualität
  - Grundlage für Entwicklung
- Folgt dem OLAP Paradigma
- Einfach und leicht verständlich
- Kein Case Tool
- Umfasst Konstrukte zur Abbildung von Cubes, Dimensionen und Beziehungen

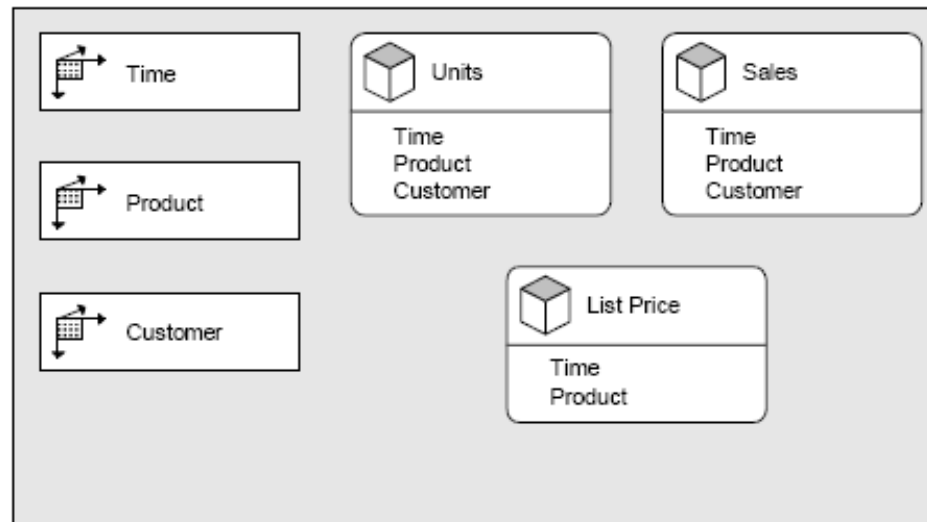
# ADAPT: Grundobjekte

## Core objects

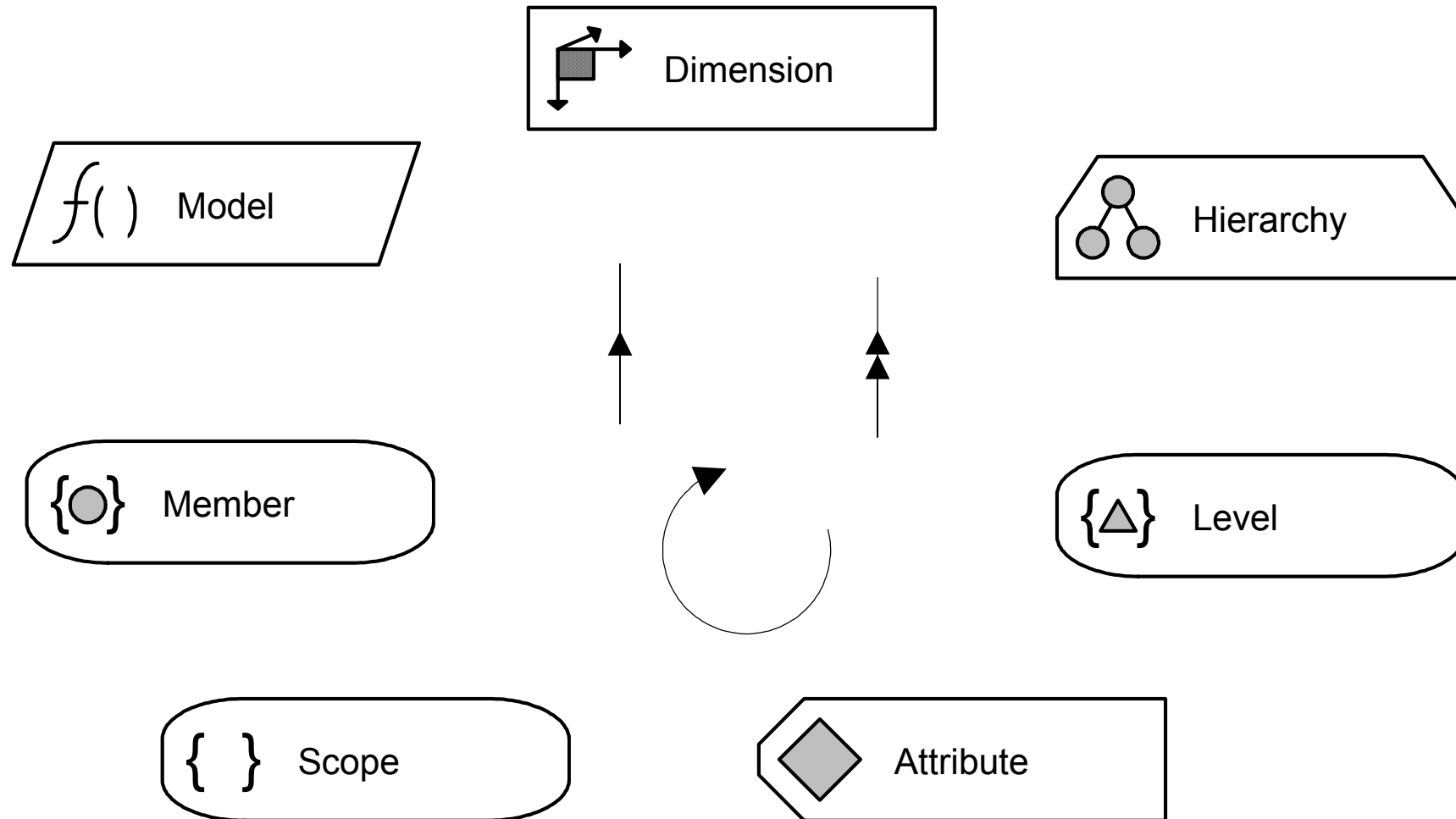
 **Dimension**  
Axes or index of Hypercubes

 **Hypercube**  
Basic data storage unit, an n-dimensional array

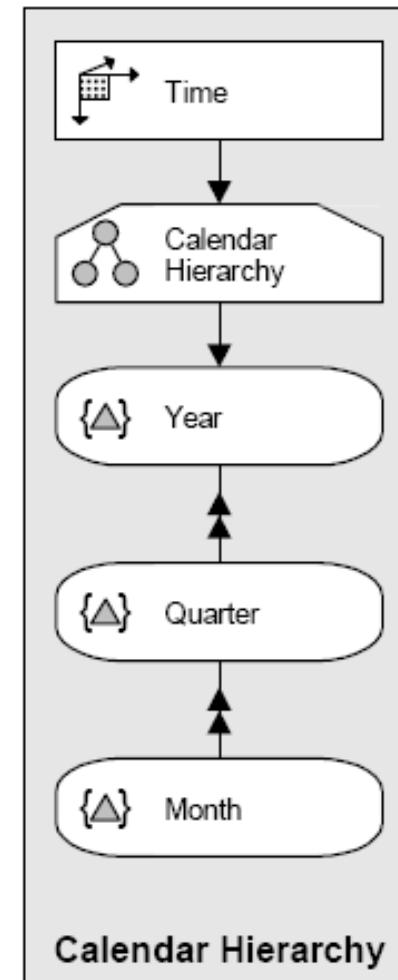
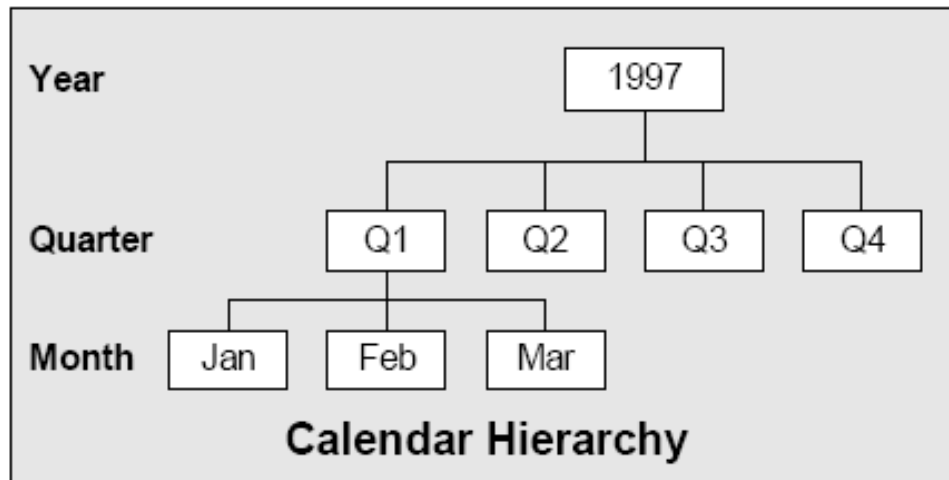
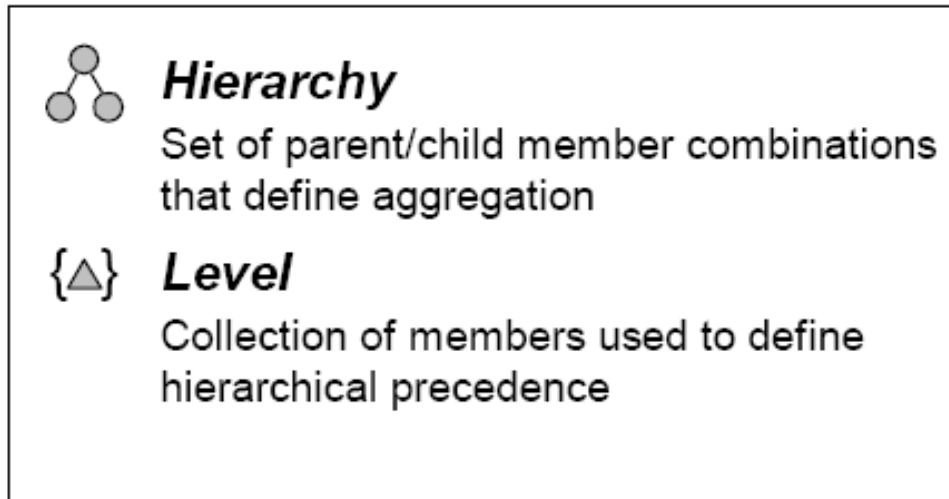
**Core Objects**



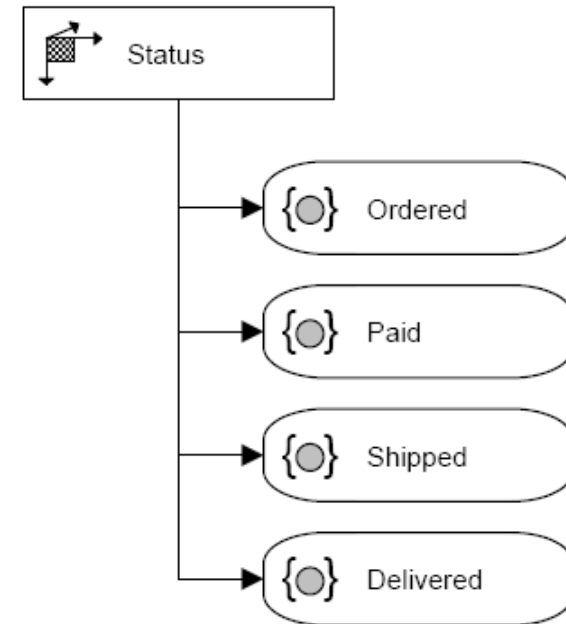
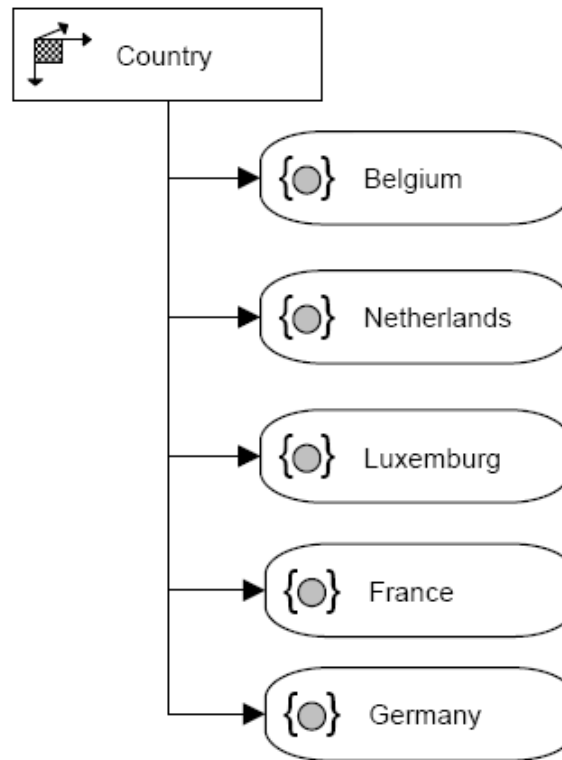
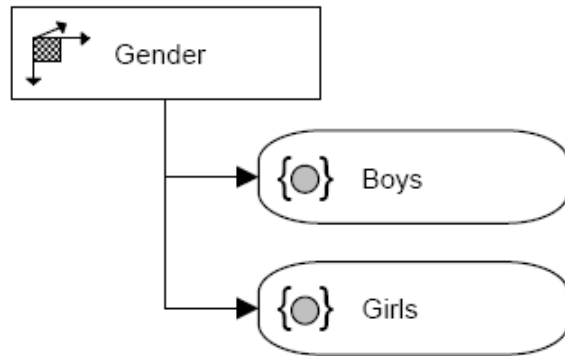
# Objekte der Dimensionsmodellierung in ADAPT



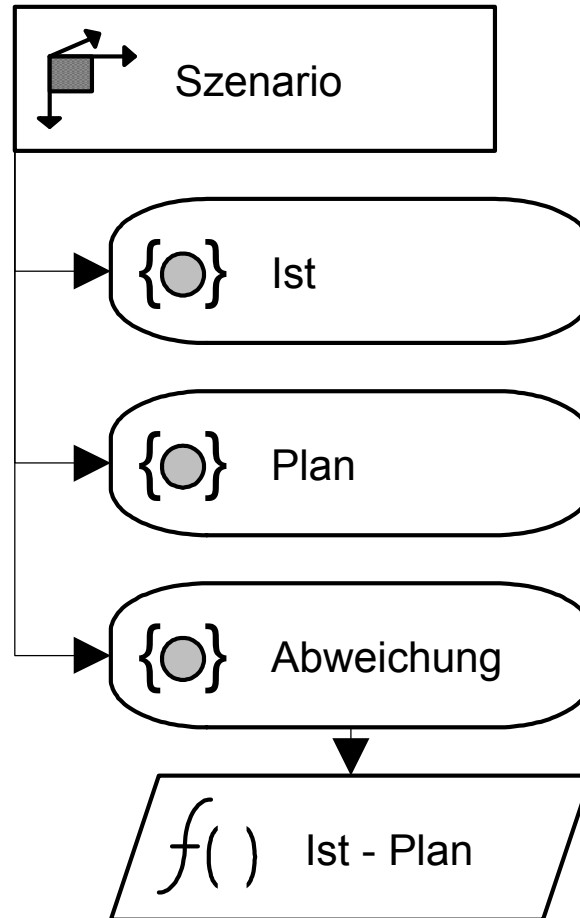
# ADAPT: Dimensionsobjekte in der Übersicht



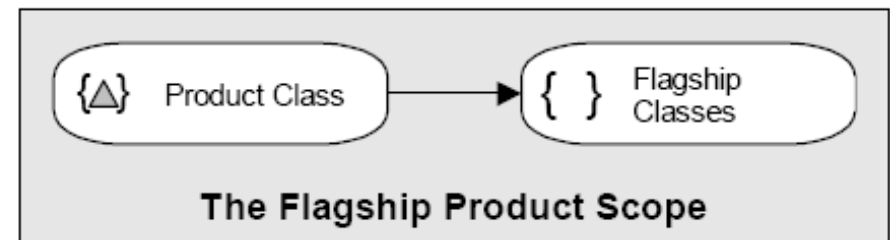
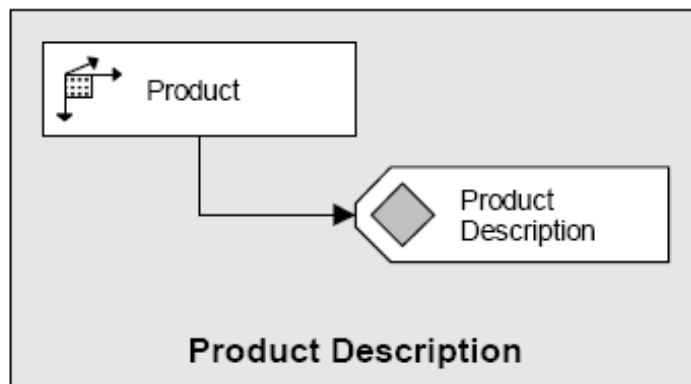
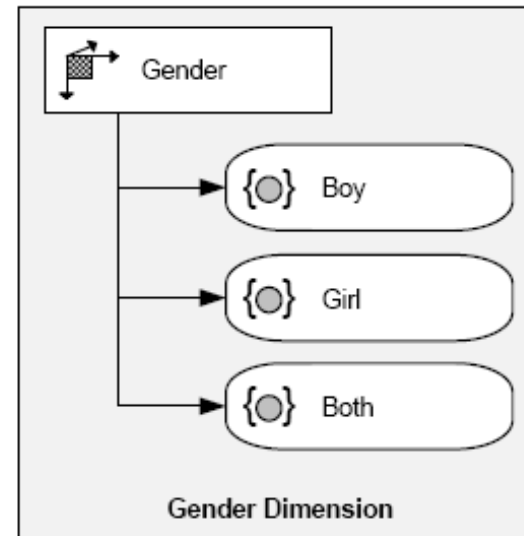
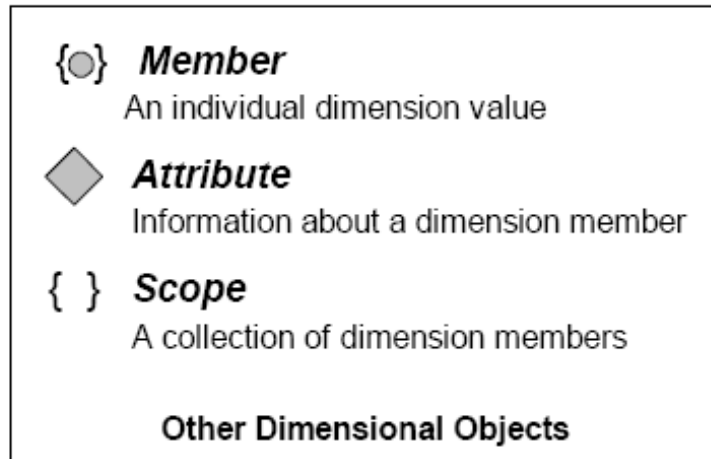
# Einfache Dimensionselemente



# Elementbestimmte Dimension

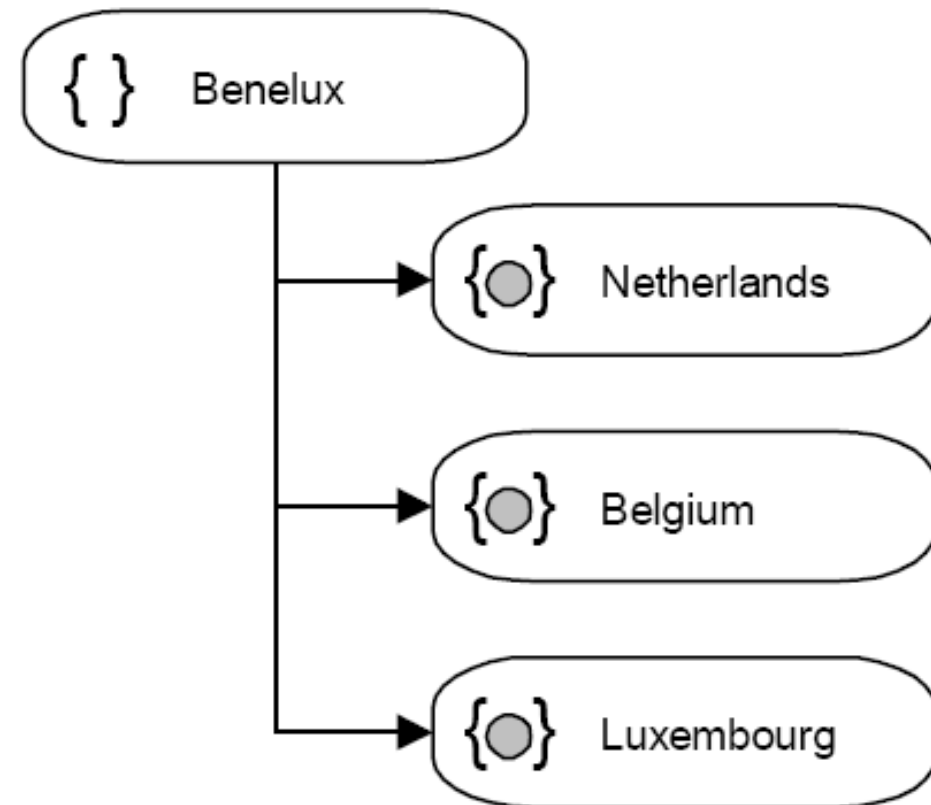


# ADAPT: Erweiterte Dimensionsobjekte



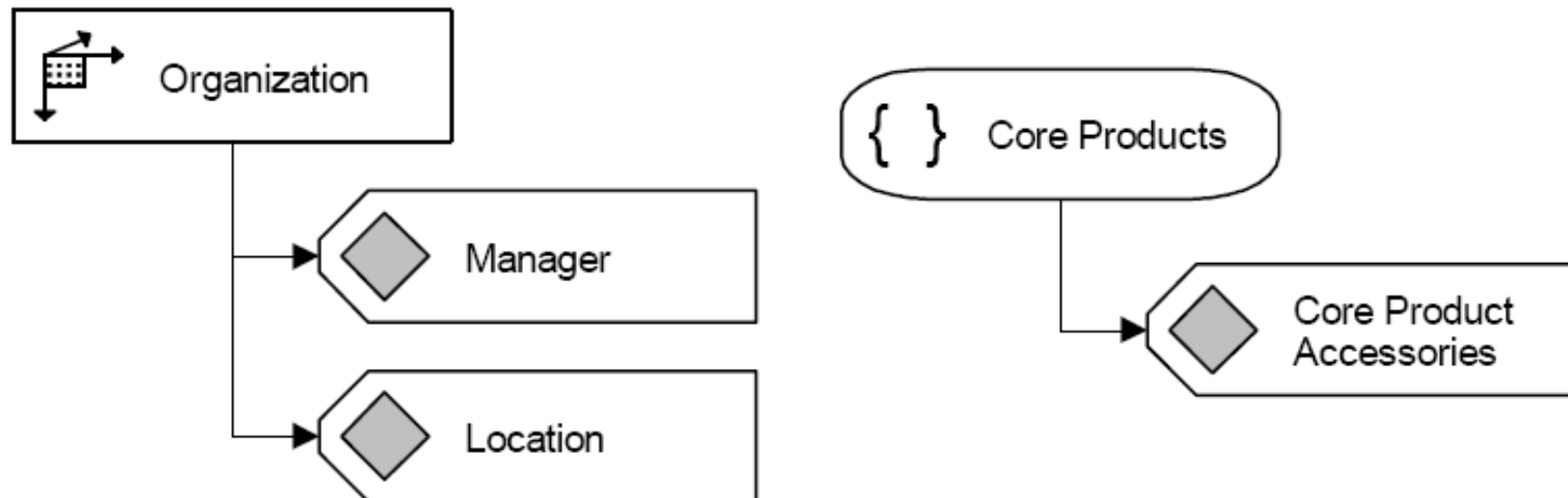
## Dimension Scope

- Definiert eine Menge von Dimensionselementen
- Ist eine Teilmenge der übergeordneten Menge von Elementen (z.B. Dimension oder ein anderer Scope)
- Sind hilfreich zur Modellierung von Ausnahmen

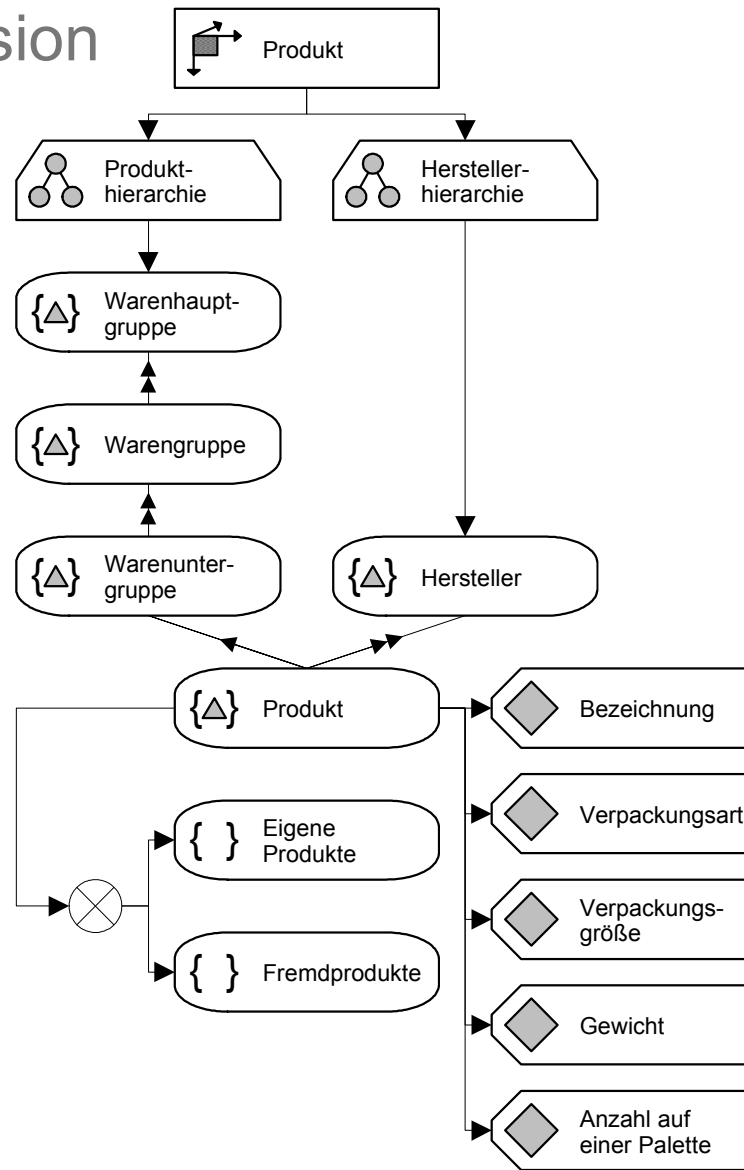


## Beispiele für Attribute

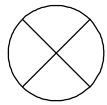
- Attribute können an der Dimension, an einer Ebene oder an einem Scope hängen



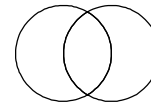
# Ebenenbestimmte Dimension



# Beziehungstypen in ADAPT (Domain Relationships)



Fully exclusive



Fully overlapping



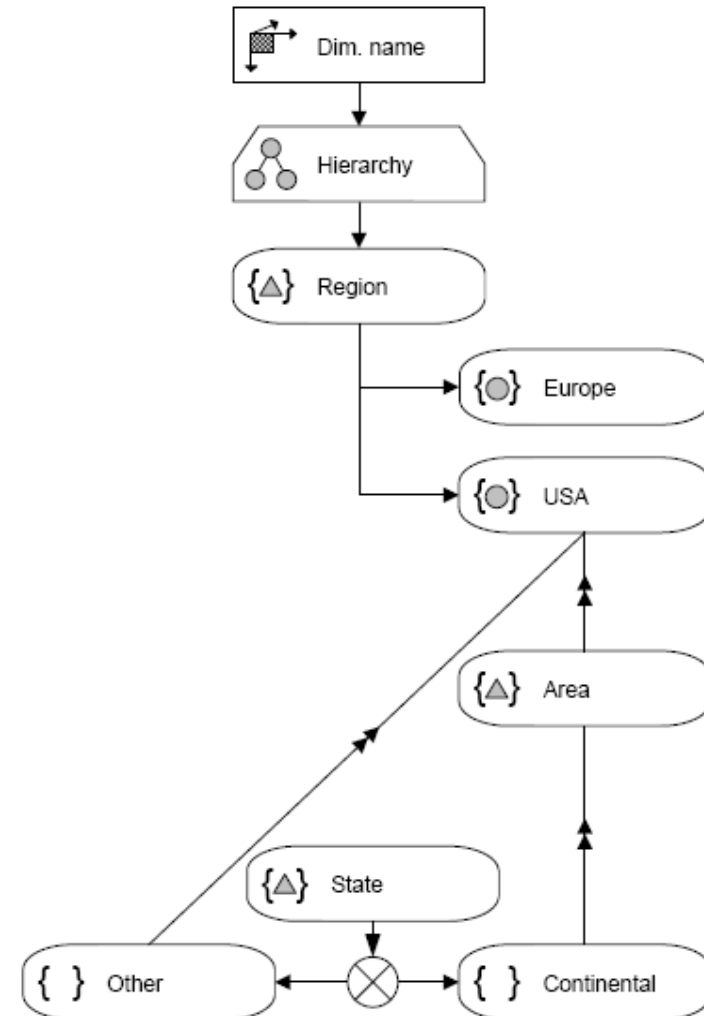
Partially exclusive



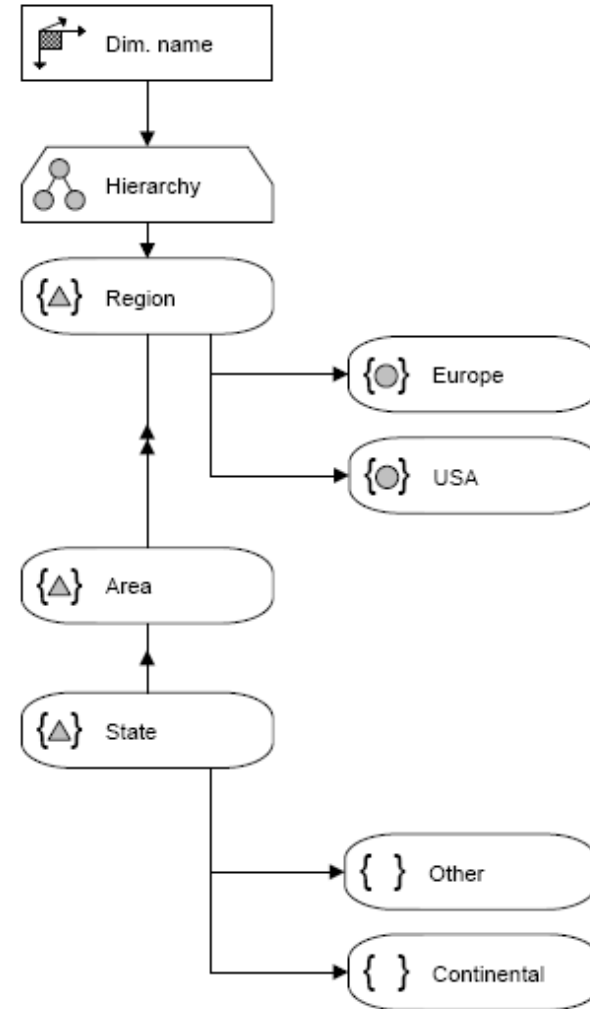
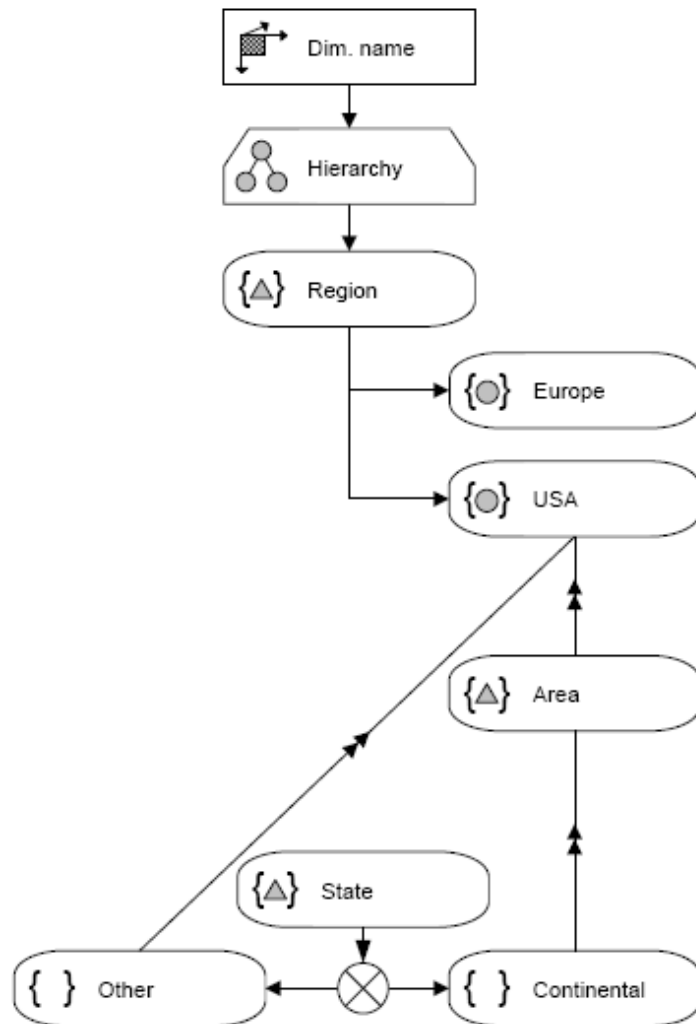
Partially overlapping

# Weitere Einsatzmöglichkeiten von Beziehungen

- Ebene Region hat USA und Europe
- Alle States sind entweder Continental oder Other
- Continental States verdichten sich zu Area
- Other States verdichten sich zu Region USA

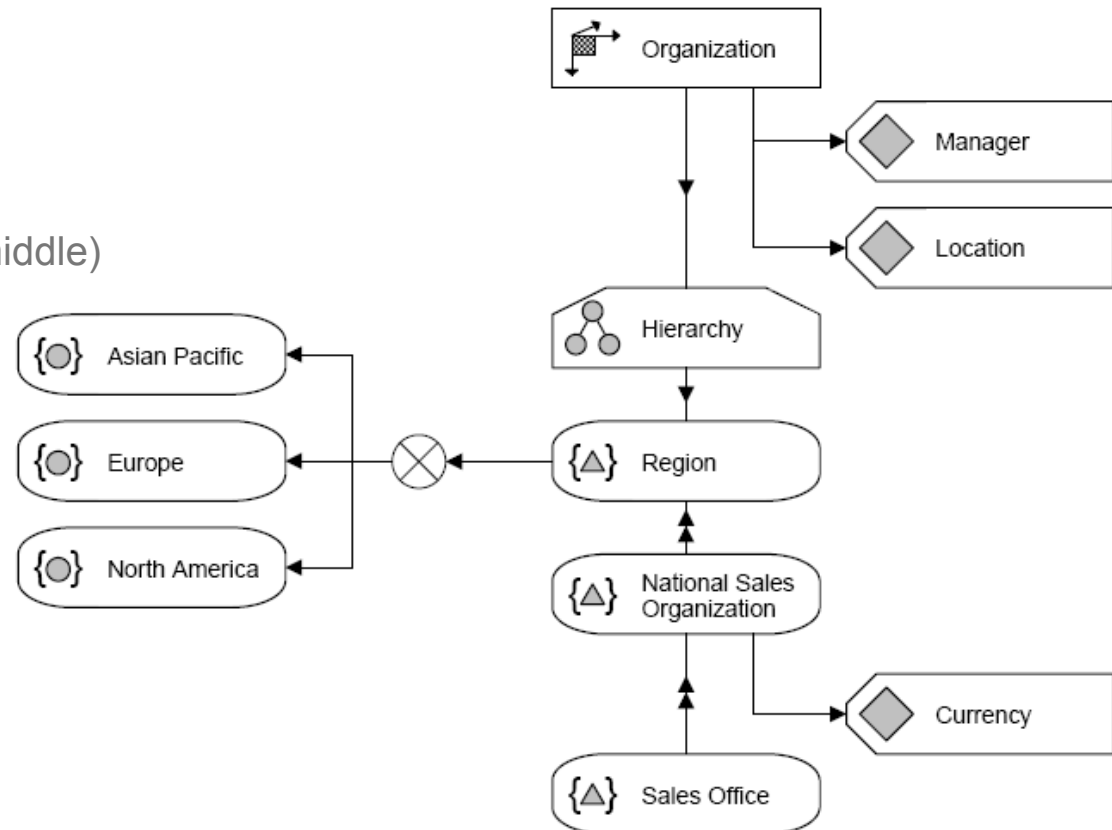


# Mit und ohne Domain Relationships



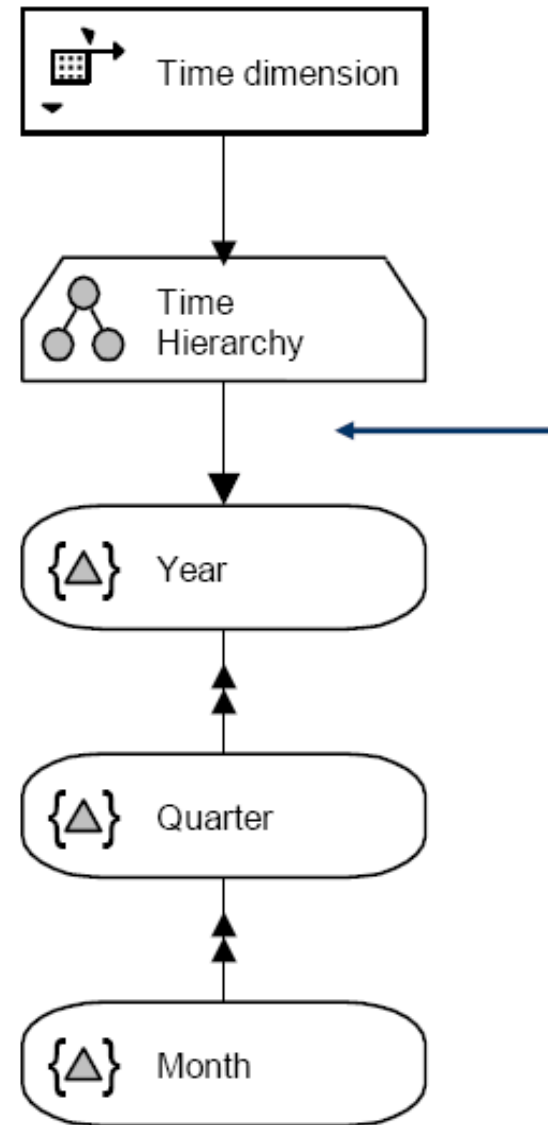
# Attribute auf unterschiedlichen Ebenen

- Eine Organisationsdimension
- mit einer Hierarchie
- mit drei Ebenen:
  - Region (top level)
  - National Sales Organization (middle)
  - Sales office (bottom)
- Ebene Region hat drei Elemente:
  - APAC
  - Europe
  - NA
- Alle Elemente aller Ebenen haben
  - Einen Manager
  - Eine Adresse
- Jede national Sales Org hat
  - Eine Currency



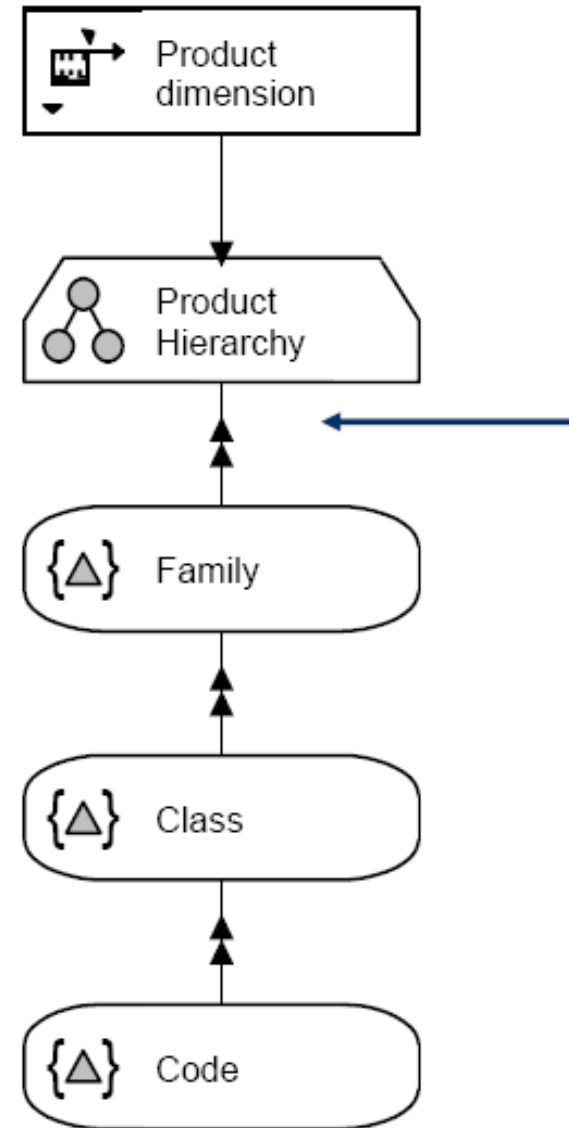
# Totale Verdichtung

- In diesem Beispiel gibt es keinen „All Years“ Knoten
- Wird in ADAPT durch den Pfeil vom Hierarchiesymbol zu der obersten Ebene verdeutlicht



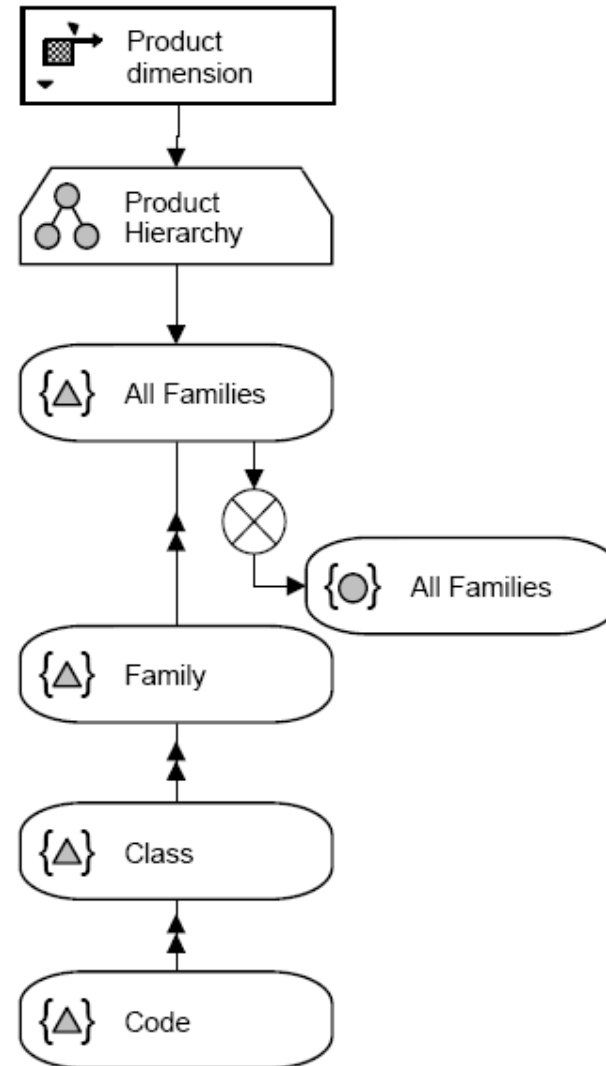
# Implizite Totale Verdichtung

- In diesem Fall gibt es implizit definiertes Element „All Families“ als totale Verdichtung
- Dies wird in ADAPT durch den Doppelpfeil von der obersten Ebene zum Hierarchiesymbol dargestellt

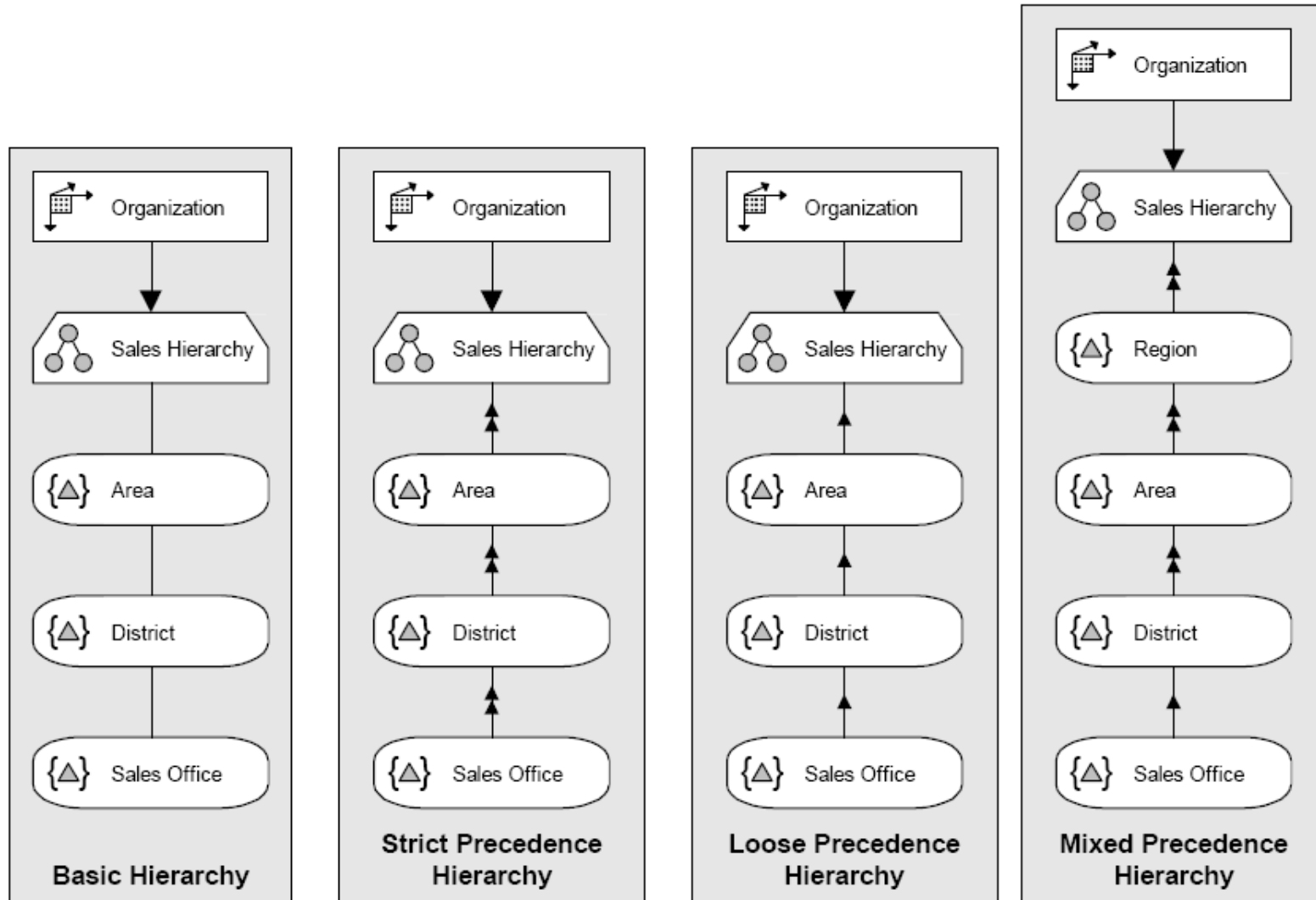


# Explizite Totale Verdichtung

- „All“ als Präfix für die Top-Ebene weist auf die totale Verdichtung hin
- Optional kann in ADAPT das Top-Element auch über eine Domain Relationship modelliert werden (exclusive full set)

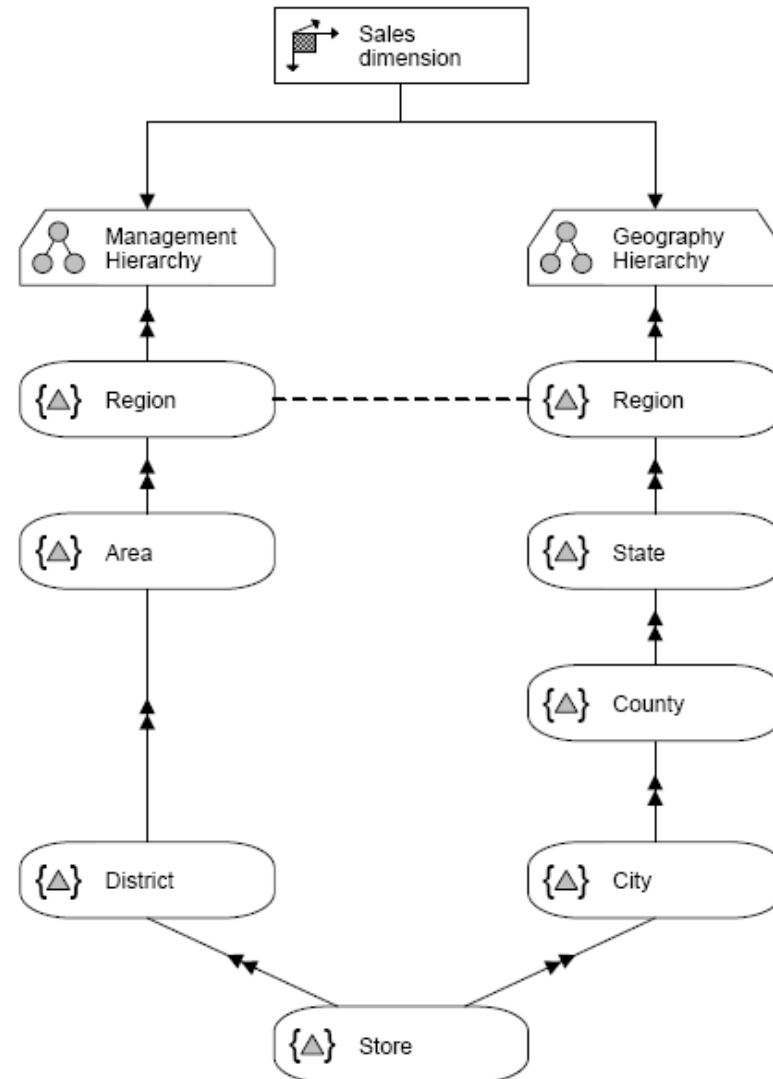


# Unbalancierte Hierarchien



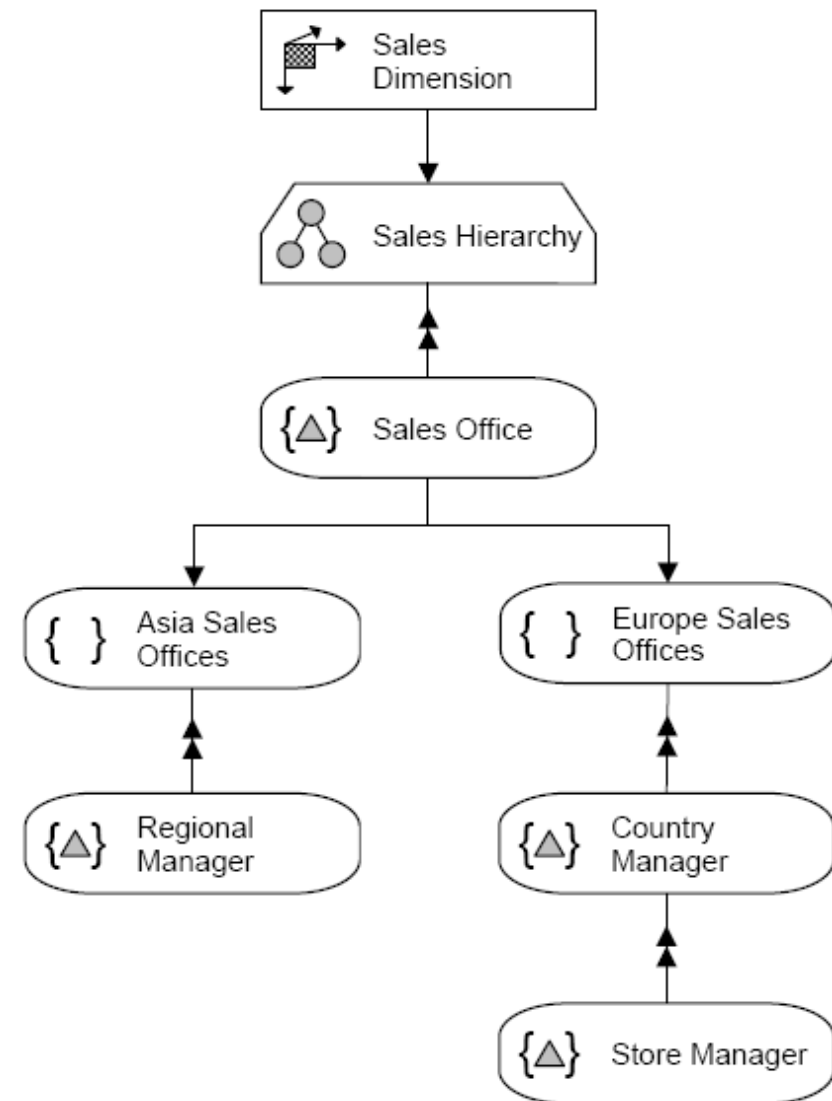
# Parallele Hierarchien mit gemeinsamer Ebene

- Gemeinsam genutzte Ebenen sind meistens die Top-Ebene
- Werden in verschiedenen Hierarchien der gleichen Dimension genutzt



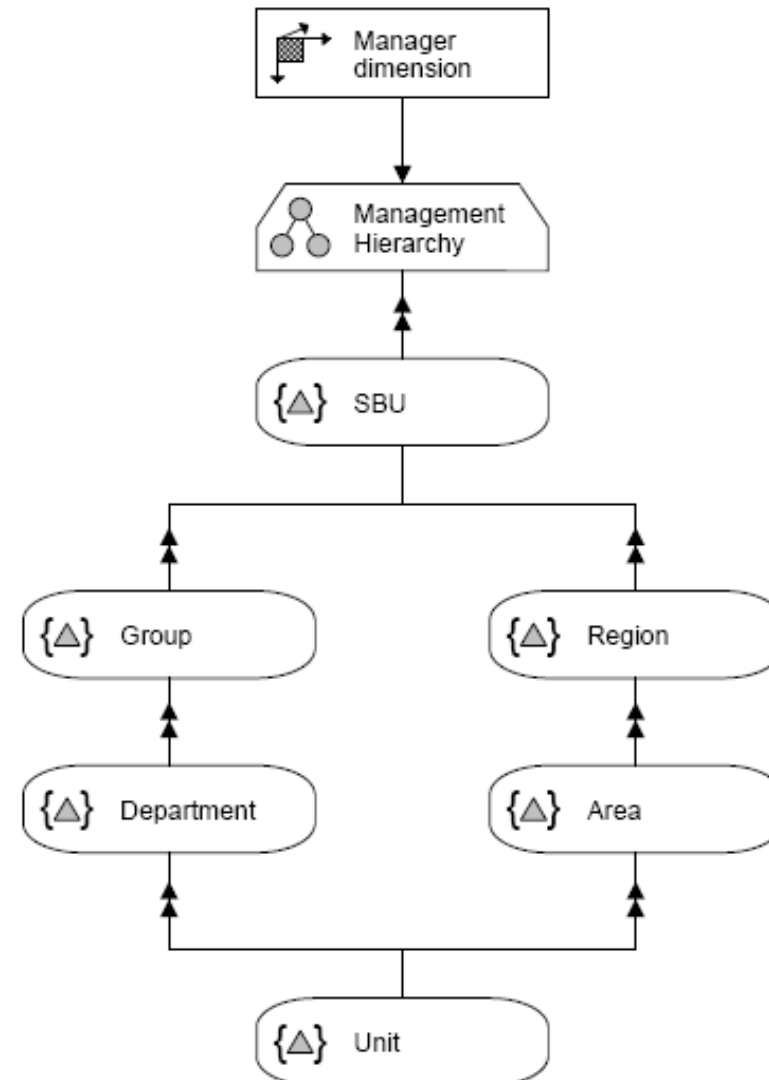
# Hierarchien mit Verzweigungen

- Ein Hierarchiezweig ist ein Teil der Hierarchie, der von einem Knoten bis zur granularen Ebene geht
- Hilfreich ist, das Element oder den Scope am Eingang in den Zweig explizit zu modellieren



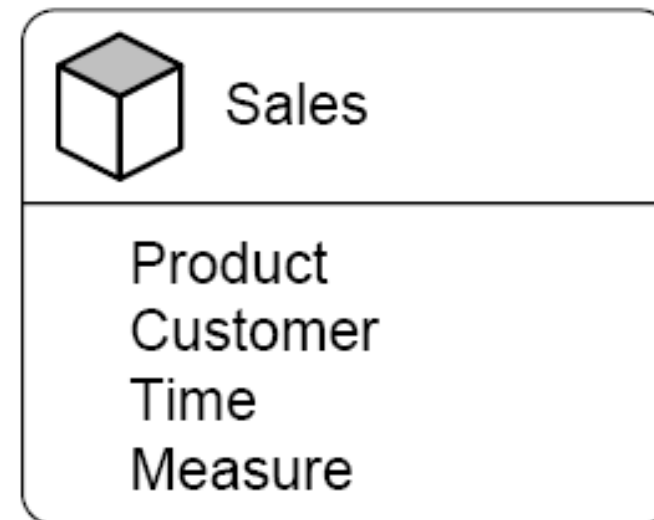
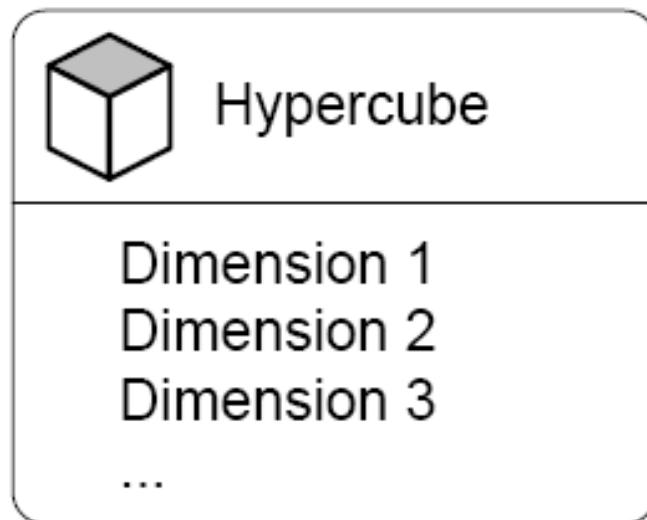
## Spezielle Form der Verzweigung

- Eine Unit berichtet entweder an eine Area oder an ein Department, aber nicht an beides
- Ausgedrückt durch die einzelne gemeinsame Linie von Unit zur nächst höheren Ebene

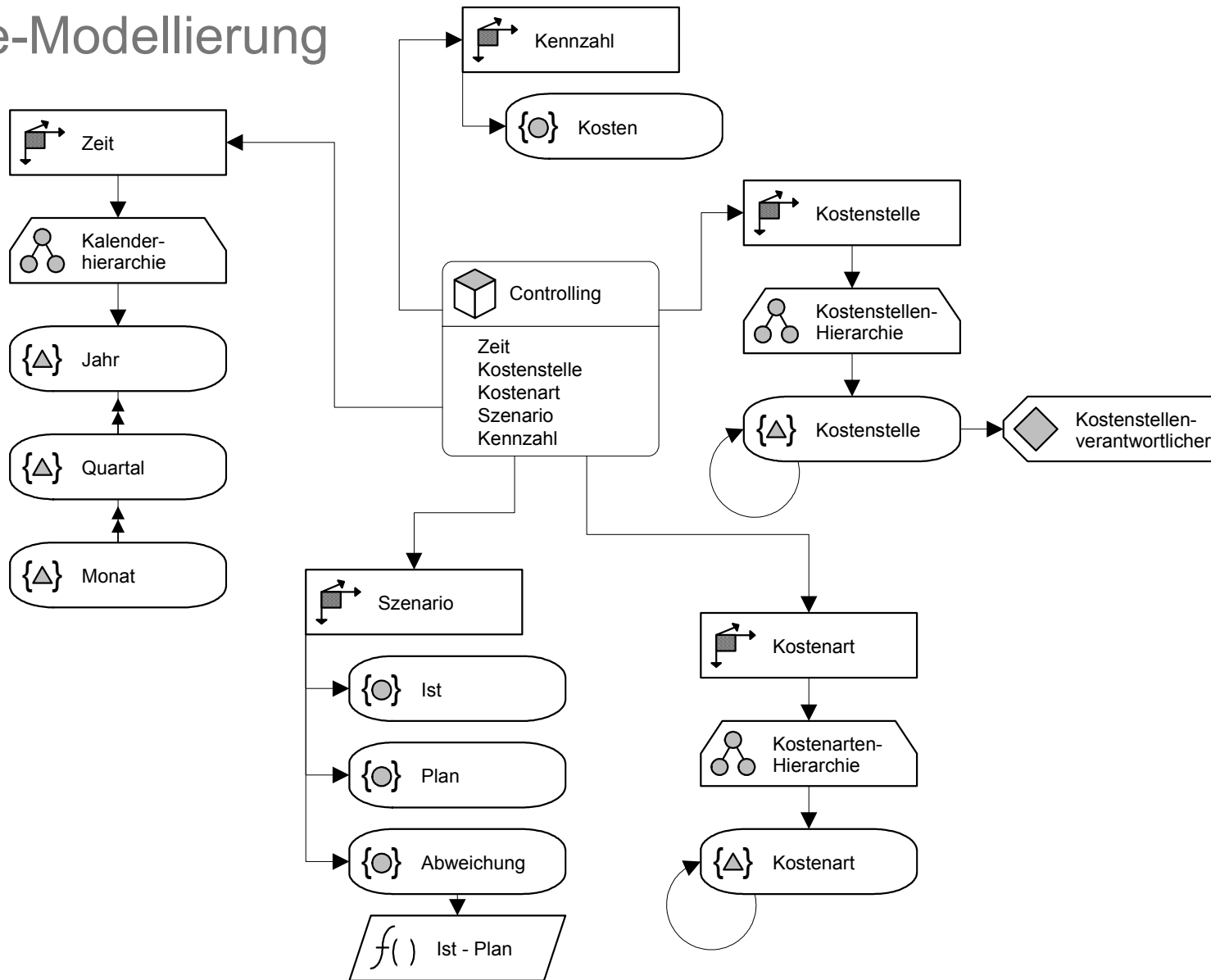


## Cube Shape in ADAPT

- Ein Cube hat immer eine Kennzahlen-Dimension
- Die meisten haben auch eine Zeitdimension
- Cubes gruppieren Kennzahlen gleicher Dimensionierung und gleicher Granularität



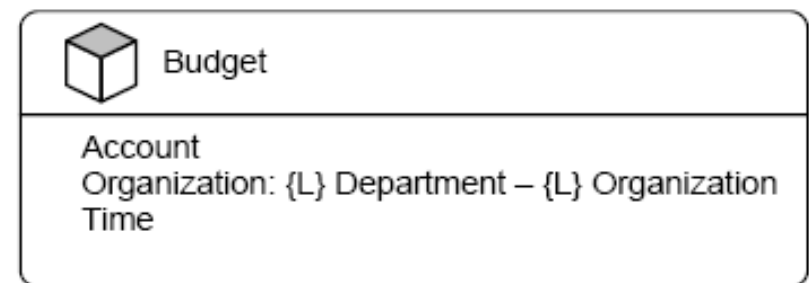
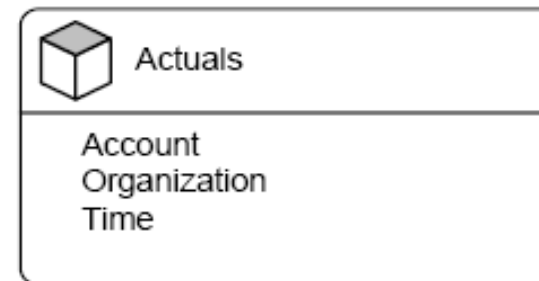
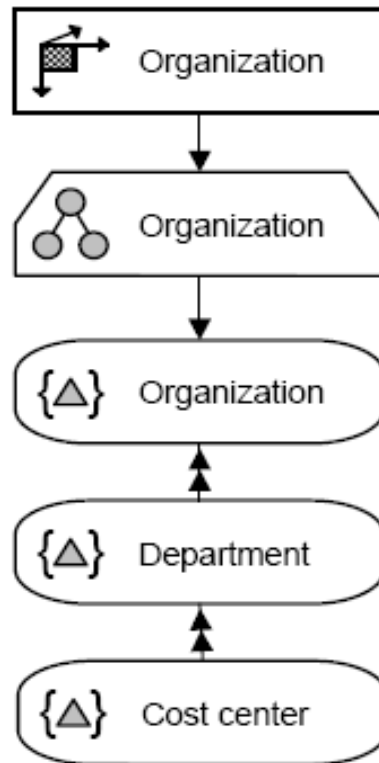
# Cube-Modellierung



# Cube Modifiers

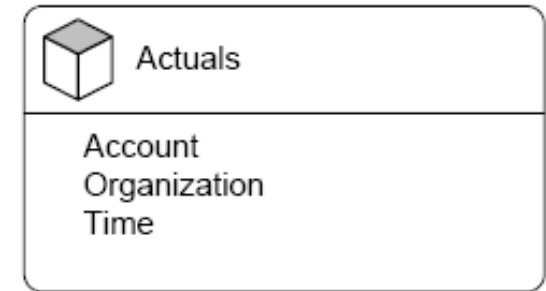
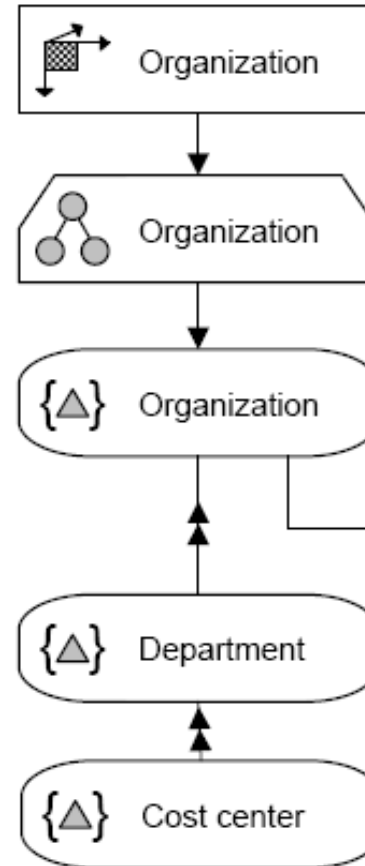
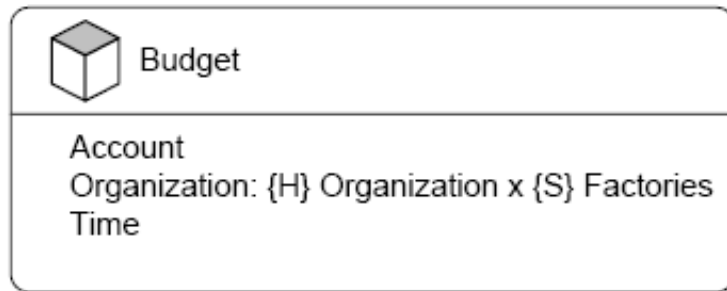
- Ein Cube nutzt nicht immer die vollständige „conformed“ dimension
- Mit „Modifiers“ kann die Granularität und Dimensionsteilmenge des Cubes festgelegt werden

- {M} filtert Member
- {L} filtert Level
- {H} filtert Hierarchy
- {S} filtert Scope
- - Bereich
- x inkl. der Parents und Childs
- ; mehrere Subsets

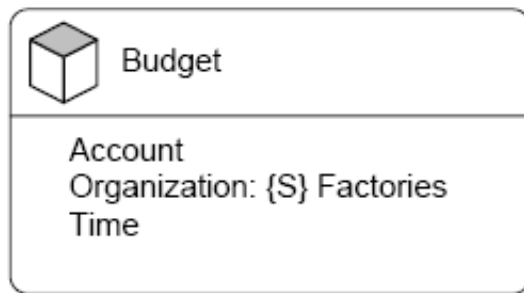


Cube Modifiers: Erweiterung von ADAPT, die Nippur (NL) entwickelt hat

# Beispiel für Cube Modifiers

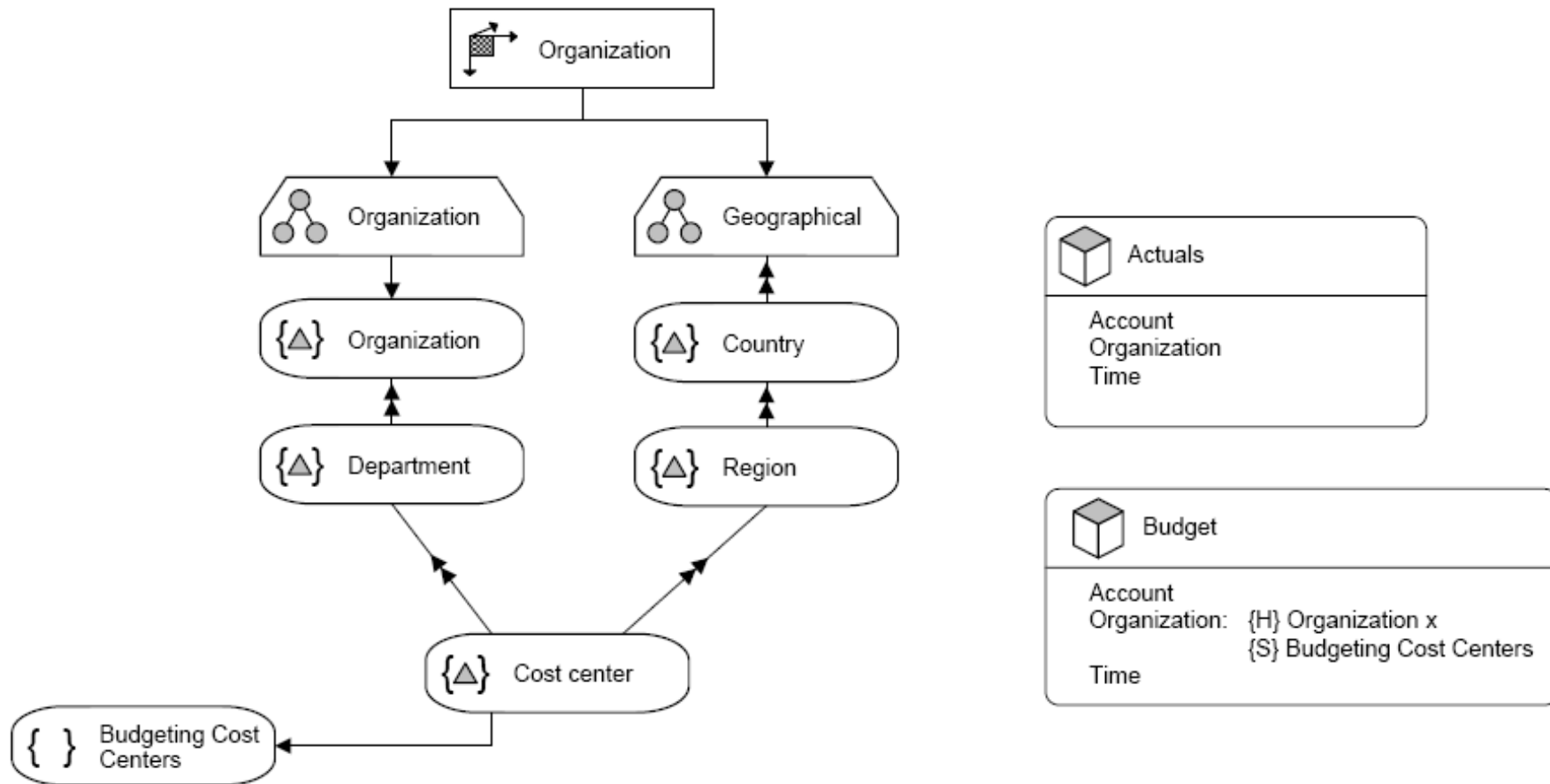


Alle Elemente der Organisationshierarchie deren Organisation in der Menge Factories ist

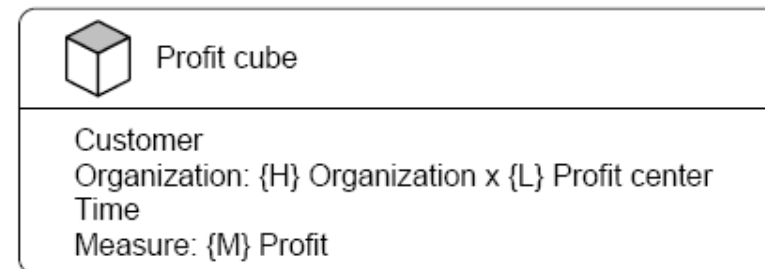
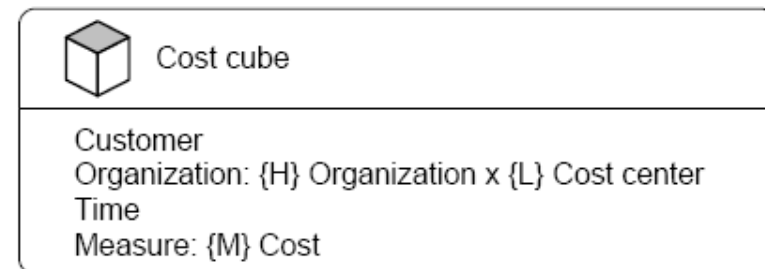
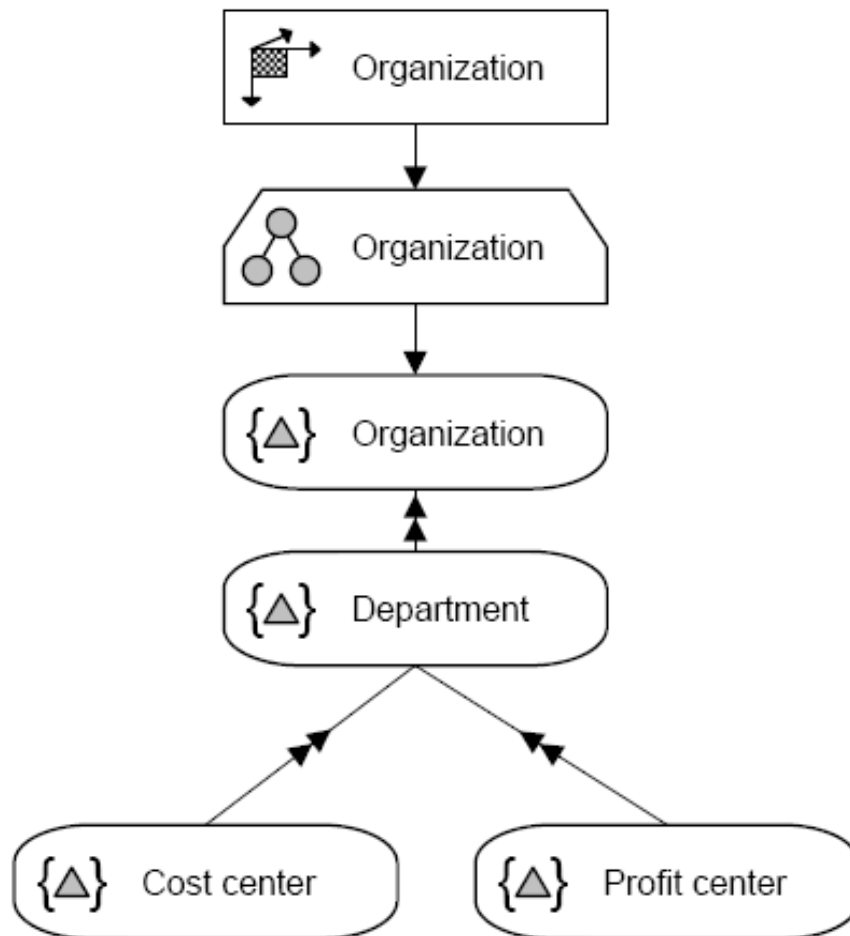


Nur Elemente des Scope Factories

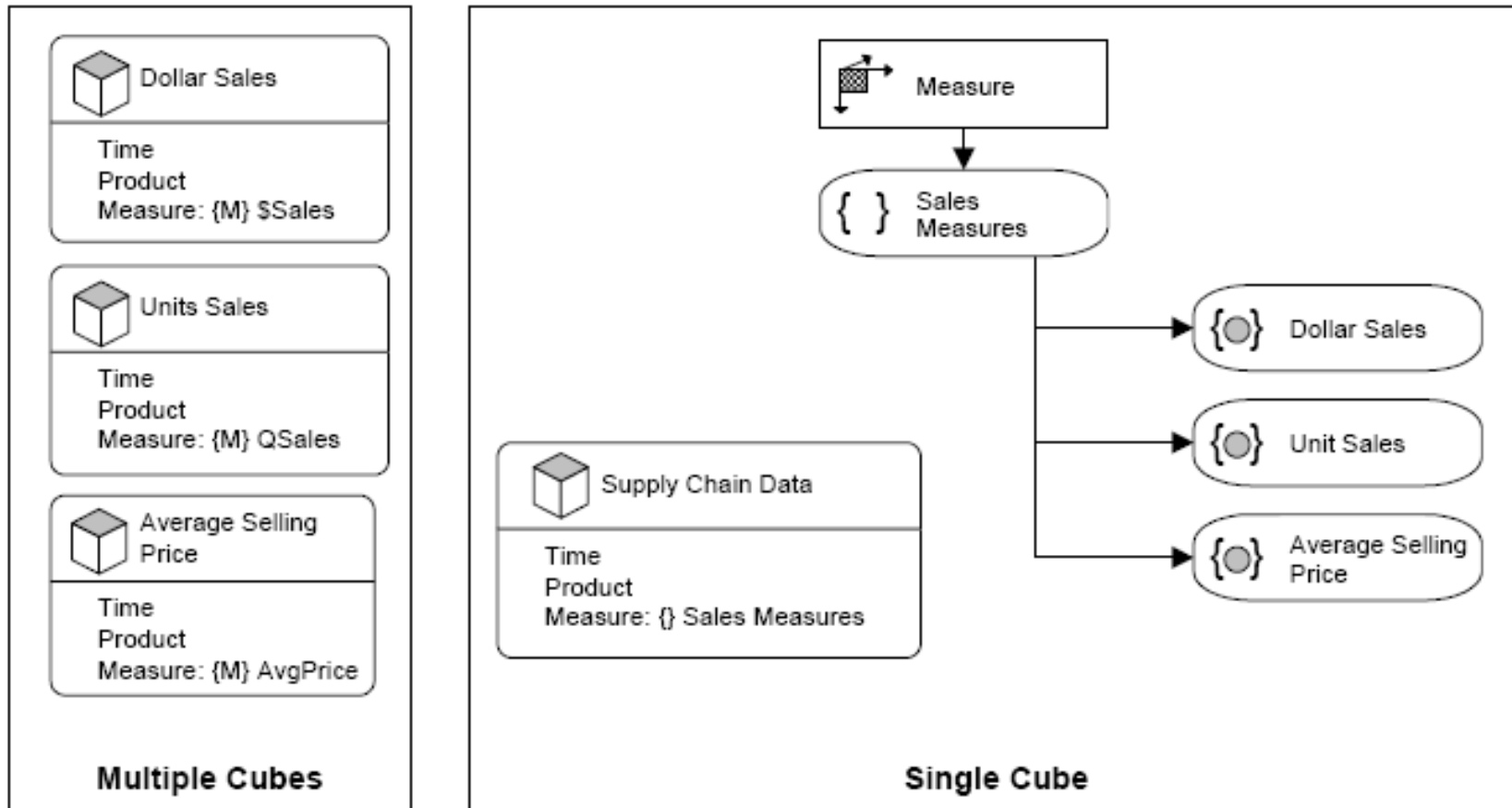
# Beispiel für Plandaten



# Weiteres Beispiel

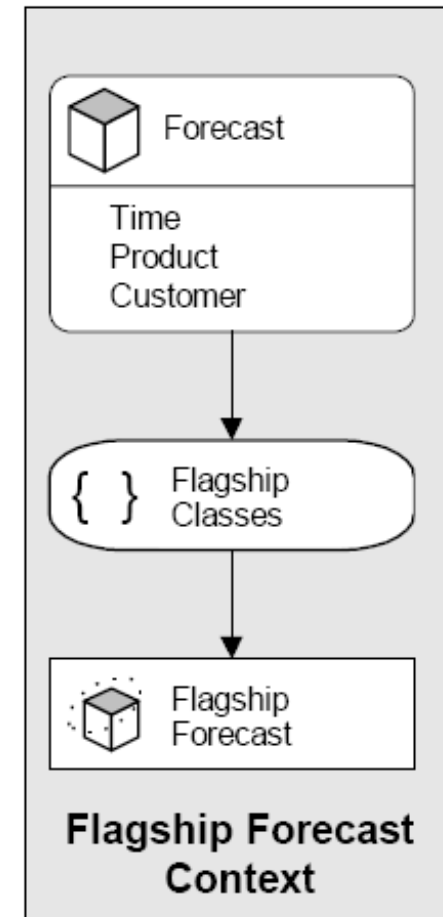
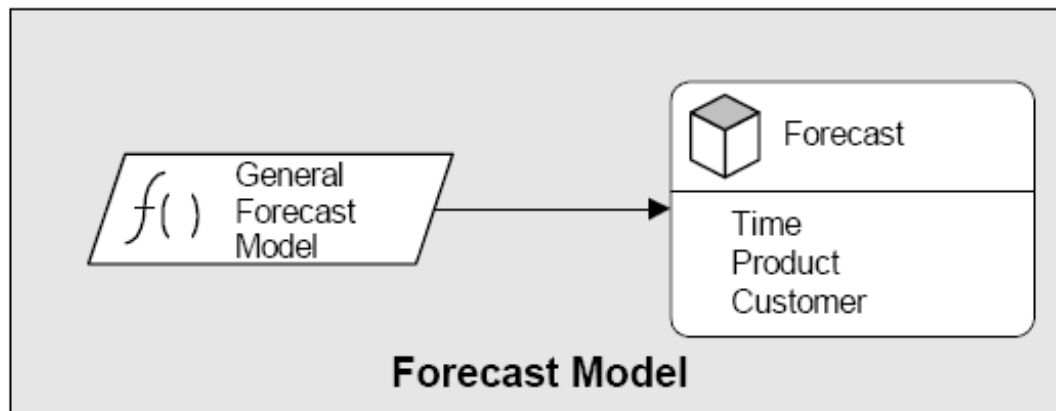
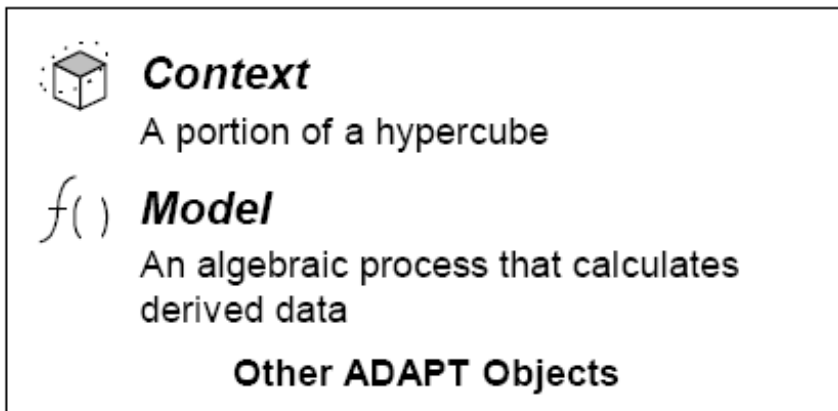


# Mehrere Cubes oder einen Cube?



- Trade-Off: Viele Cube-Übergreifende Operationen vs. sparse Cubes

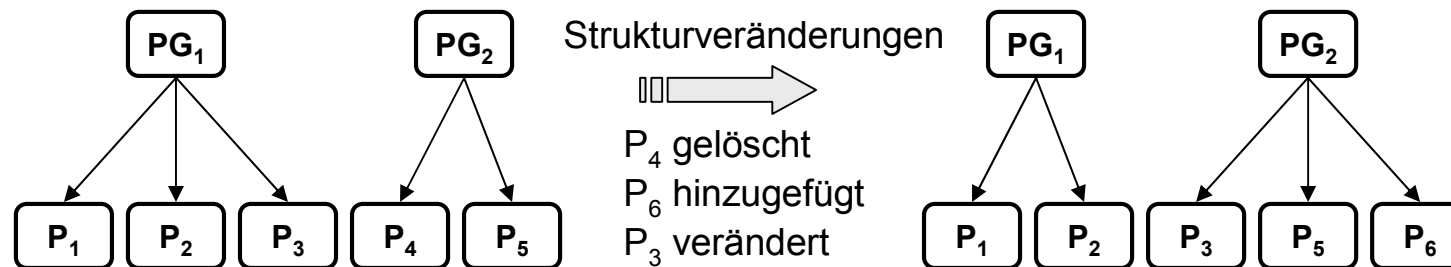
# ADAPT: Weitere Objekte der Modellierung



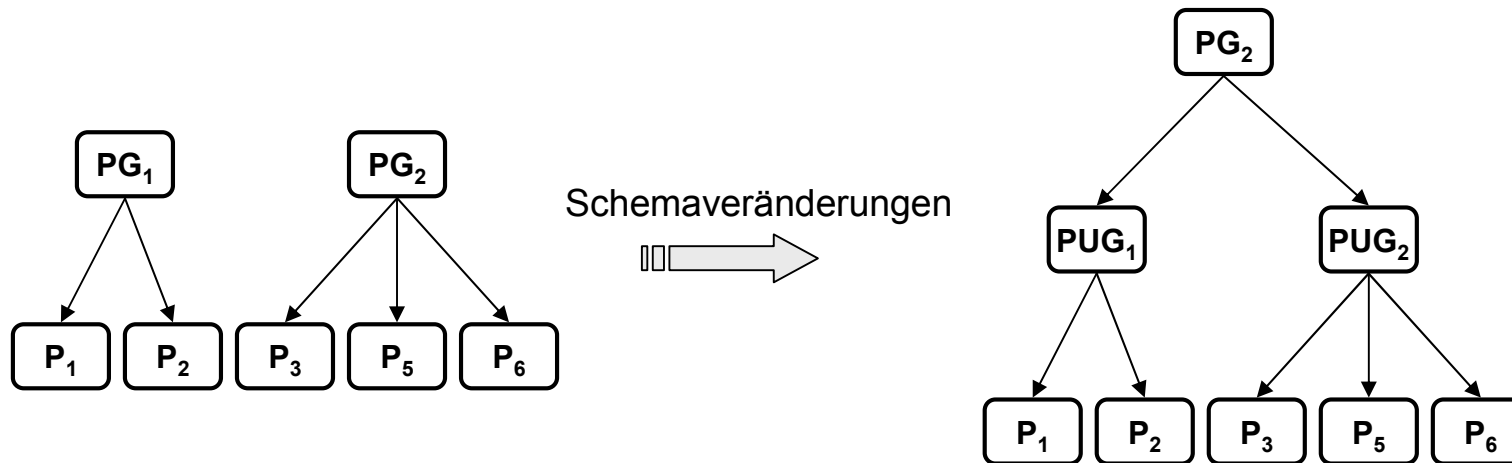
## Agenda

- Motivation
- Konzeptionelle Modellierung des Reporting Layers
- **Zeitabhängigkeit im Reporting Layer am Beispiel Star Schema**
- Staging und Core Data Warehouse Modellierung
- Architektur für Enterprise Data Warehousing

# Zeitabhängigkeit: Strukturelle Veränderungen in Dimensionen



# Zeitabhängigkeit: Implizierte Schemaveränderung



## Anforderungen des Berichtswesens

### Mögliche Anforderungen:

- Berichte nach aktueller Struktur
- Berichte nach definierter historischer Struktur (Zeitpunkt)
- Berichte gemäß historischer Wahrheit (Transaktionsorientierte Sicht der Bewegungsdaten)
- Bericht vergleichbarer Resultate (Zeitintervall)

# Beispiel für "Slowly Changing Dimensions"

## Produktdimension in 2010-04

Produkt	Produktgruppe
P A	PG X
P B	PG X
P C	PG Y
P D	PG Y

## Produktdimension in 2010-05

Produkt	Produktgruppe
P A	PG X
P B	PG Y (geändert)
P C	PG Y
P D	PG Y
P E	PG Y (neu)

## Fakten-Tabelle

Produkt	Periode	Umsatz
P A	2010-04	100
P B	2010-04	100
P C	2010-04	100
P D	2010-04	100
P A	2010-05	100
P B	2010-05	100
P C	2010-05	100
P D	2010-05	100
P E	2010-05	100

## Berichts-anforderungen - Szenarien

### Berichtsszenario aktuelle Struktur

Produktgruppe	Umsatz 2010-04	Umsatz 2010-05
PG X	100	100
PG Y	300	400

### Berichtsszenario alte Struktur

Produktgruppe	Umsatz 2010-04	Umsatz 2010-05
PG X	200	200
PG Y	200	200

### Berichtsszenario historische Wahrheit

Produktgruppe	Umsatz 2010-04	Umsatz 2010-05
PG X	200	100
PG Y	200	400

### Berichtsszenario vergleichbare Resultate

Produktgruppe	Umsatz 2010-04	Umsatz 2010-05
PG X	100	100
PG Y	200	200

# Szenario I : Bericht mit aktueller Struktur

**Produktdimension in 2010-05**

Produkt	Produktgruppe
P A	PG X
P B	PG Y (geändert)
P C	PG Y
P D	PG Y
P E	PG Y (neu)

**Fakten-Tabelle**

Produkt	Periode	Umsatz
P A	2010-04	100
P B	2010-04	100
P C	2010-04	100
P D	2010-04	100
P A	2010-05	100
P B	2010-05	100
P C	2010-05	100
P D	2010-05	100
P E	2010-05	100



Produktgruppe	Umsatz 2010-04	Umsatz 2010-05
PG X	100	100
PG Y	300	400

# Szenario II : Bericht mit alter Struktur

## Fakten-Tabelle

### Produktdimension in 2010-04

Produkt	Produktgruppe
P A	PG X
P B	PG X
P C	PG Y
P D	PG Y

Produkt	Periode	Umsatz
P A	2010-04	100
P B	2010-04	100
P C	2010-04	100
P D	2010-04	100
P A	2010-05	100
P B	2010-05	100
P C	2010-05	100
P D	2010-05	100
P E	2010-05	100



Produktgruppe	Umsatz 2010-04	Umsatz 2010-05
PG X	200	200
PG Y	200	200

# Szenario III : Bericht nach historischer Wahrheit

## Produktdimension in 2010-04

Produkt	Produktgruppe
P A	PG X
P B	PG X
P C	PG Y
P D	PG Y

## Produktdimension in 2010-05

Produkt	Produktgruppe
P A	PG X
P B	PG Y (geändert)
P C	PG Y
P D	PG Y
P E	PG Y (neu)

## Fakten-Tabelle

Produkt	Periode	Umsatz
P A	2010-04	100
P B	2010-04	100
P C	2010-04	100
P D	2010-04	100
P A	2010-05	100
P B	2010-05	100
P C	2010-05	100
P D	2010-05	100
P E	2010-05	100

Produktgruppe	Umsatz 2010-04	Umsatz 2010-05
PG X	200	100
PG Y	200	400

# Szenario IV : Bericht mit vergleichbaren Resultaten

## Produktdimension in 2010-04

Produkt	Produktgruppe
P A	PG X
P B	PG X
P C	PG Y
P D	PG Y

## Produktdimension in 2010-05

Produkt	Produktgruppe
P A	PG X
P B	PG Y (geändert)
P C	PG Y
P D	PG Y
P E	PG Y (neu)

## Fakten-Tabelle

Produkt	Periode	Umsatz
P A	2010-04	100
P B	2010-04	100
P C	2010-04	100
P D	2010-04	100
P A	2010-05	100
P B	2010-05	100
P C	2010-05	100
P D	2010-05	100
P E	2010-05	100

Produktgruppe	Umsatz 2010-04	Umsatz 2010-05
PG X	100	100
PG Y	200	200

## Lösungsalternativen bei Strukturbrüchen

- **Anpassung des historischen Datenmaterials an neue Strukturen**  
**Vorteile:** Keine Aufblähung des Datenbestandes; Datenstrukturen bleiben überschaubar  
**Nachteile:** Alte Strukturen sind verloren; Anwender wollen u. U. Berichte mit alten Strukturen
- **Separate Speicherung des historischen Datenbestandes zusätzlich zum Komplettbestand mit neuen Strukturen**  
**Vorteile:** Alte Auswertungen können abgerufen werden  
**Nachteile:** Datenvolumen; aufwendige Aktualisierung, wenn Anwender auch neue Zahlen in den alten Strukturen sehen wollen; verwirrend für Endbenutzer
- **Aufbau paralleler Hierarchien mit alten/neuen Strukturen**  
**Vorteile:** Alle Zahlen können mit Stammdaten basierten Strukturen angezeigt werden  
**Nachteile:** Dimensionsstruktur kaum überschaubar
- **Temporale Datenbanken - Gültigkeitsstempel**  
**Vorteile:** Alle Zahlen können mit beliebigen Strukturen angezeigt werden  
**Nachteile:** Performance-Impact

## Varianten der Berücksichtigung von Strukturänderungen

- „as is“ Szenario
- „as of“ & „as is“ Szenario
- „as posted“ Szenario
- „as posted“ & „as is“ Szenario
- „as posted“ & „as of“ & „as is“ Szenario

## „as is“ Szenario: Update-Verfahren

- Einfach zu implementieren
- Stellt keine Historie von Attributsausprägungen zur Verfügung
- Einfachster Fall
- Unterstützt nur „as is“ Szenarien
- Meistens ausreichend
- **Empfehlung für nur „as is“**

Produkt_Id	SKU	Produkt	Gruppe
101	AB6	PA600	High Fidelity
102	AX3	PAX300	High Fidelity
103	AX4	PAX450	Consumer Electronic

Produkt_Id	SKU	Produkt	Gruppe
101	AB6	PA600	High Fidelity
<del>102</del>	<del>AX3</del>	<del>PAX300</del>	<del>High Fidelity</del>
102	AX3	PAX300	Consumer Electronic
103	AX4	PAX450	Consumer Electronic

## „as is“ & quasi „as of“: Mehrere Attributspalten

- Geringfügig größerer Speicherbedarf in Dimensionstabellen
- Stellt historische Version ohne Zeitzuordnung zur Verfügung
- Unterstützt Auswertungen aktueller Daten mit historischer Gruppierung und umgekehrt
- Entspricht paralleler Hierarchie

Produkt_Id	SKU	Produkt	Gruppe	Alte Gruppe
101	AB6	PA600	High Fidelity	High Fidelity
102	AX3	PAX300	High Fidelity	High Fidelity
103	AX4	PAX450	Consumer Electronic	Consumer Electronic

Produkt_Id	SKU	Produkt	Gruppe	Alte Gruppe
101	AB6	PA600	High Fidelity	High Fidelity
102	AX3	PAX300	Consumer Electronic	High Fidelity
103	AX4	PAX450	Consumer Electronic	Consumer Electronic

## „as is“ & „as of“ Szenario: Versionierung ohne Datum

- Etwas größerer Speicherbedarf in Dimensionstabellen
- Benötigt unbedingt künstliche Dimensionsschlüssel
- Stellt historische Versionen ohne Zeitzuordnung zur Verfügung
- Die neue Gruppierung führt zu einer neuen Ausprägung in der Dimensionstabelle
- Unterstützt keine Auswertungen aktueller Daten mit historischer Gruppierung und umgekehrt

Produkt_Id	Version	SKU	Produkt	Gruppe
101	1	AB6	PA600	High Fidelity
102	1	AX3	PAX300	High Fidelity
103	1	AX4	PAX450	Consumer Electronic

Produkt_Id	Version	SKU	Produkt	Gruppe
101	1	AB6	PA600	High Fidelity
102	1	AX3	PAX300	High Fidelity
137	2	AX3	PAX300	Consumer Electronic
103	1	AX4	PAX450	Consumer Electronic

## „as is“ & „as of“ Szenario: Gültigkeitsstempel

- Etwas größerer Speicherbedarf in Dimensionstabellen
- Benötigt zusätzlich künstlichen oder zusammen gesetzten Dimensionsschlüssel
- Die neue Gruppierung führt zu einer neuen Ausprägung in der Dimensionstabelle
- Unterstützt Auswertungen aktueller Daten mit historischer Gruppierung u. umgekehrt
- **Empfehlung für „as is“ mit „as of“**

Produkt_Id	...	Fakten
101	...	...
102	...	...
102	...	...
103	...	...

Produkt_Ald	Produkt_Id	GültigVon	GültigBis	SKU	Produkt	Gruppe
101	101	20100101	99991231	AB6	PA600	High Fidelity
102	102	20100101	20100131	AX3	PAX300	High Fidelity
137	102	20100201	99991231	AX3	PAX300	Consumer Electronic
103	103	20100101	99991231	AX4	PAX450	Consumer Electronic

# Implementierung des „as posted“ Szenarios

- Etwas größerer Speicherbedarf in Dimensionstabellen
- Benötigt unbedingt künstliche Dimensionsschlüssel
- Stellt historische Versionen zum Transaktionszeitpunkt zur Verfügung
- Die neue Gruppierung führt zu einer neuen Ausprägung in der Dimensionstabelle
- Unterstützt keine Auswertungen aktueller Daten mit historischer Gruppierung oder umgekehrt
- **Empfehlung für nur „as posted“**

Produkt_Ald	...	Fakten
101	...	...
102	...	...
137	...	...
103	...	...

Produkt_Ald	Produkt_Id	Version	SKU	Produkt	Gruppe
101	101	1	AB6	PA600	High Fidelity
102	102	1	AX3	PAX300	High Fidelity
137	102	2	AX3	PAX300	Consumer Electronic
103	103	1	AX4	PAX450	Consumer Electronic

# „as posted“ & „as of“ & „as is“ Szenario: Snapshot-Verfahren

Faktentabelle

Produkt_Id	DimX	Fact	Loaddate
101	xx	100	20100101
102	xx	130	20100101
103	xx	120	20100101
102	xx	90	20100102

- Deutlich größerer Speicherbedarf in Faktentabelle und in Dimensionstabellen
- Jede Aktualisierung der Dimensionstabelle fügt sämtliche Datensätze an die Tabelle an
- Ermöglicht vollständige Historisierung:  
 „as is“ über join Produkt\_Id und Filter Curr  
 „as of“ über join Produkt\_Id und Filter Loaddate  
 „as posted“ über join Loaddate und Produkt\_Id
- Current-Flag in Dimensionstabelle definiert aktuelle Zuordnung
- Ladedatum bzw. Zeitstempel muss in Dimensions- und Faktentabelle in Schlüssel aufgenommen werden
  - Schlechte Performance (join!)
- **Bei kleinen Dimensionen und für „as \*“ empfehlenswert**

Dimensionstabelle

Produkt_Id	Curr	SKU	Produkt	Gruppe	Loaddate
101	0	AB6	PA600	High Fidelity	20100101
102	0	AX3	PAX300	High Fidelity	20100101
103	0	AX4	PAX450	Consumer Electronic	20100101
101	1	AB6	PA600	High Fidelity	20100102
102	1	AX3	PAX300	Consumer Electronic	20100102
103	1	AX4	PAX450	Consumer Electronic	20100102

# „as posted“ & „as of“ & „as is“ Szenario: Snapshot-Variante Delta-Historisierung

Faktentabelle

Produkt_Id	DimX	Fact	Loaddate
101	xx	100	20100115
102	xx	130	20100115
103	xx	120	20100115
102	xx	90	20100215

- Größerer Speicherbedarf in Dimensionstabellen
- Jede Aktualisierung der Dimensionstabelle fügt verwendete Datensätze an die Tabelle an
- Current-Flag in Dimensionstabelle definiert aktuelle Zuordnung
- Ladedatum bzw. Zeitstempel muss in Dimensions- und Faktentabelle in Schlüssel aufgenommen werden → Schlechtere Performance (join!)
- **Empfehlenswert für „alle“ Szenarien, wenn „as posted“ Attribute unbekannt**

Dimensionstabelle

Produkt_Id	Curr	SKU	Produkt	Gruppe	Loaddate	GueltigVon	GueltigBis
101	1	AB6	PA600	High Fidelity	20100115	20100101	99991231
102	0	AX3	PAX300	High Fidelity	20100115	20100101	20100131
102	1	AX3	PAX300	Consumer Electronic	20100215	20100201	20100228
103	1	AX4	PAX450	Consumer Electronic	20100115	20100101	99991231

# „as posted“ & „as of“ & „as is“ Szenario: Mehrere Fremdschlüssel und Gültigkeitsstempel

- Etwas größerer Speicherbedarf in Dimensionstabellen
- Benötigt zusätzlich künstliche Dimensionsschlüssel
- Stellt historische Versionen auch zum Transaktionszeitpunkt zur Verfügung
- Die neue Gruppierung führt zu einer neuen Ausprägung in der Dimensionstabelle
- Unterstützt Auswertungen aktueller Daten mit historischer Gruppierung u. umgekehrt
- **Empfehlung für „alle“ Szenarien, wenn „as posted“ Attribute bekannt**

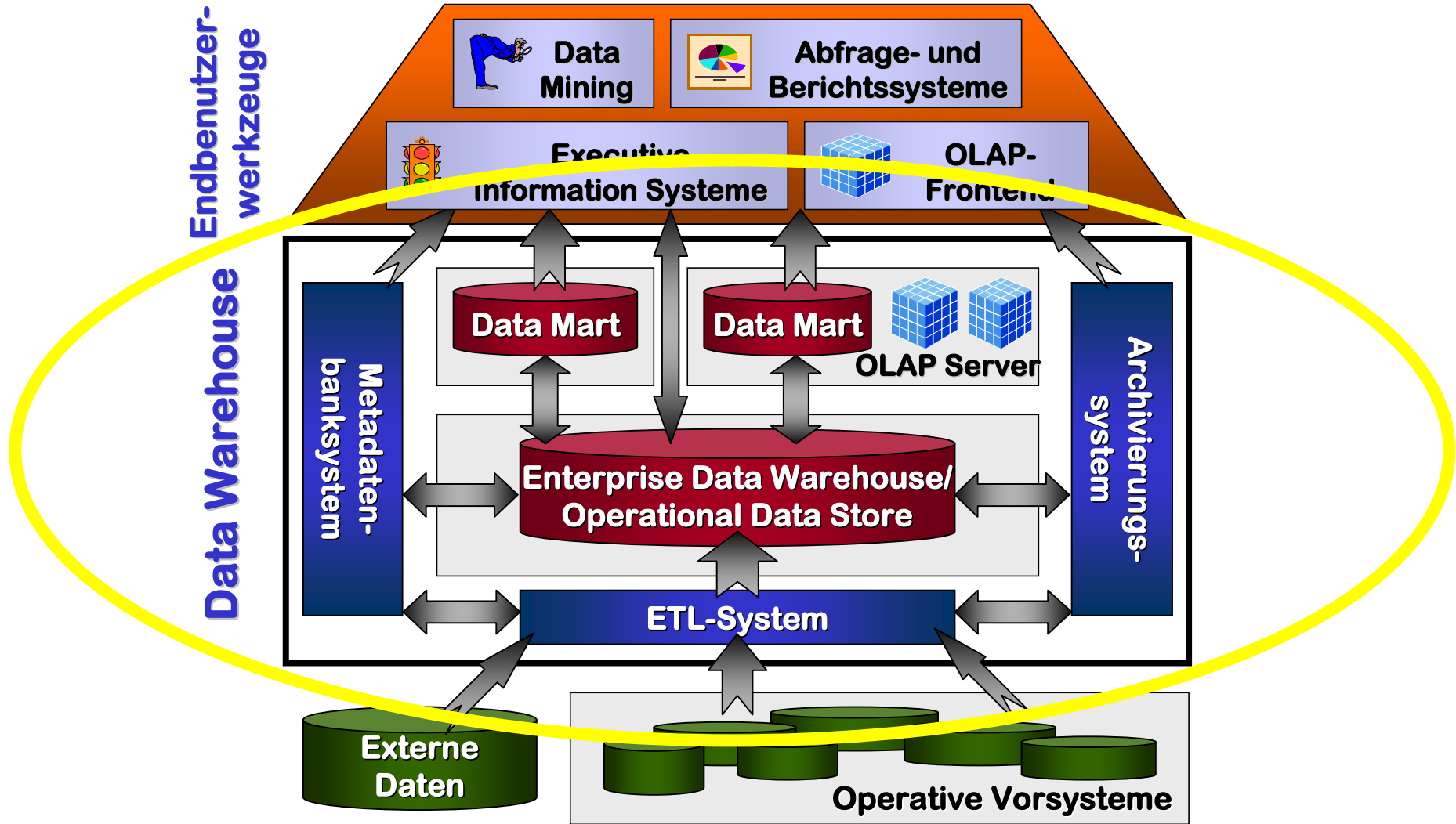
Produkt_Ald	Produkt_Id	...	Fakten
101	101	...	...
102	102	...	...
137	102	...	...
103	103	...	...

Produkt_Ald	Produkt_Id	GültigVon	GültigBis	SKU	Produkt	Gruppe
101	101	20100101	99991231	AB6	PA600	High Fidelity
102	102	20100101	20100131	AX3	PAX300	High Fidelity
137	102	20100201	99991231	AX3	PAX300	Consumer Electronic
103	103	20100101	99991231	AX4	PAX450	Consumer Electronic

## Agenda

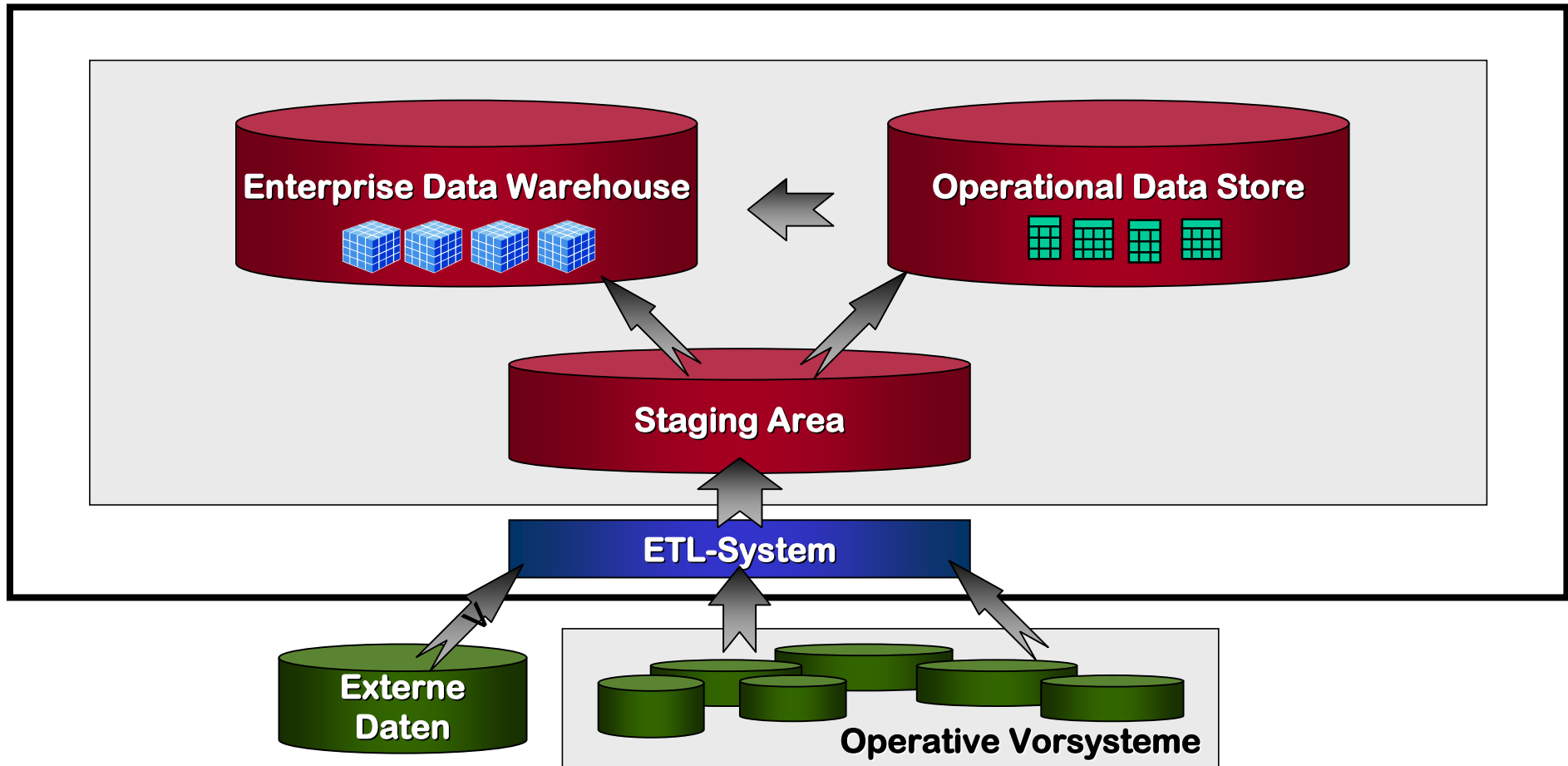
- Motivation
- Konzeptionelle Modellierung des Reporting Layers
- Zeitabhängigkeit im Reporting Layer am Beispiel Star Schema
- **Staging und Core Data Warehouse Modellierung**
- Architektur für Enterprise Data Warehousing

# Enterprise Data Warehousing: Aktuelles Architektur-Konzept



# Speichersegmente des Data Warehouse

## Data Warehouse

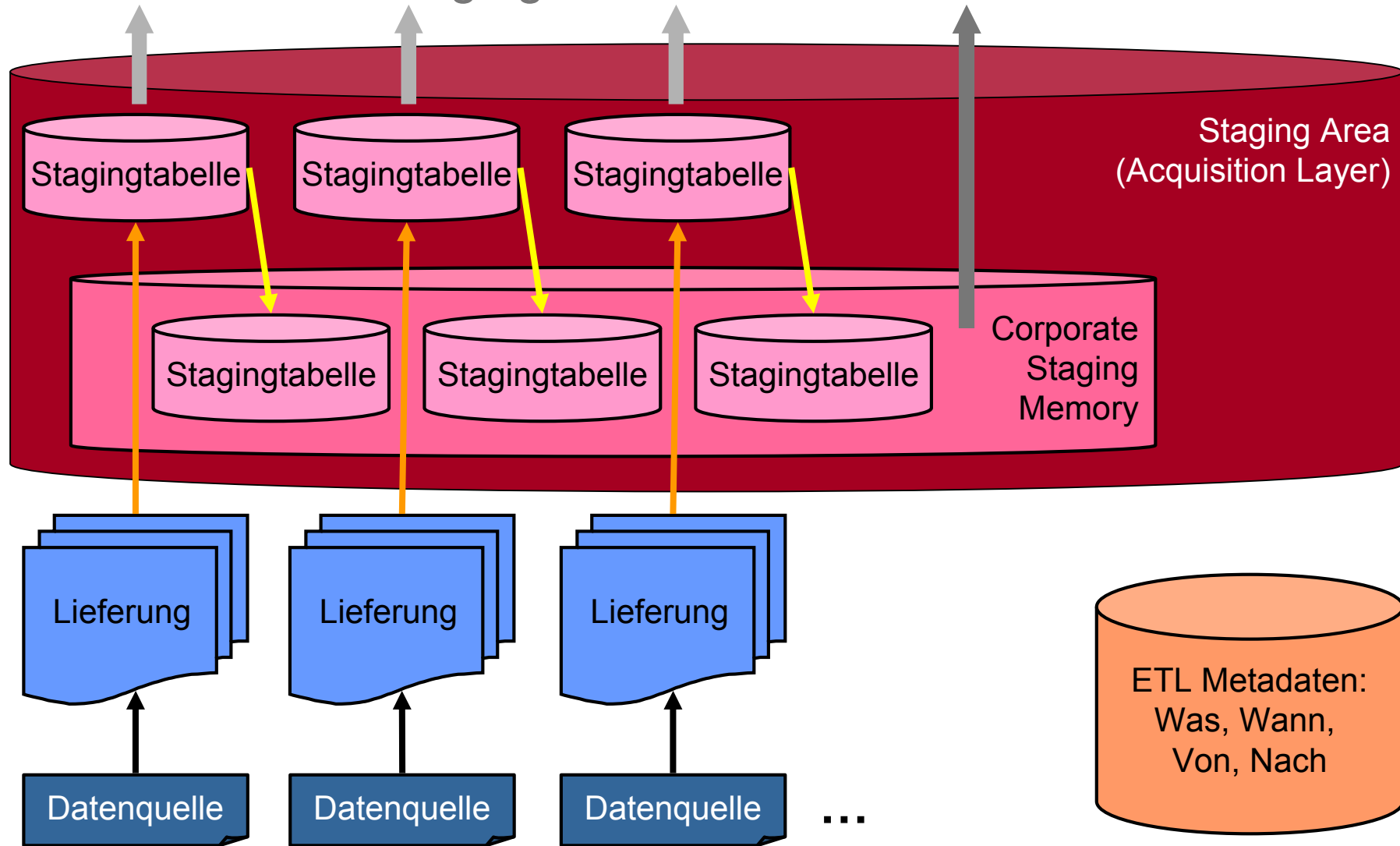


## Staging Area

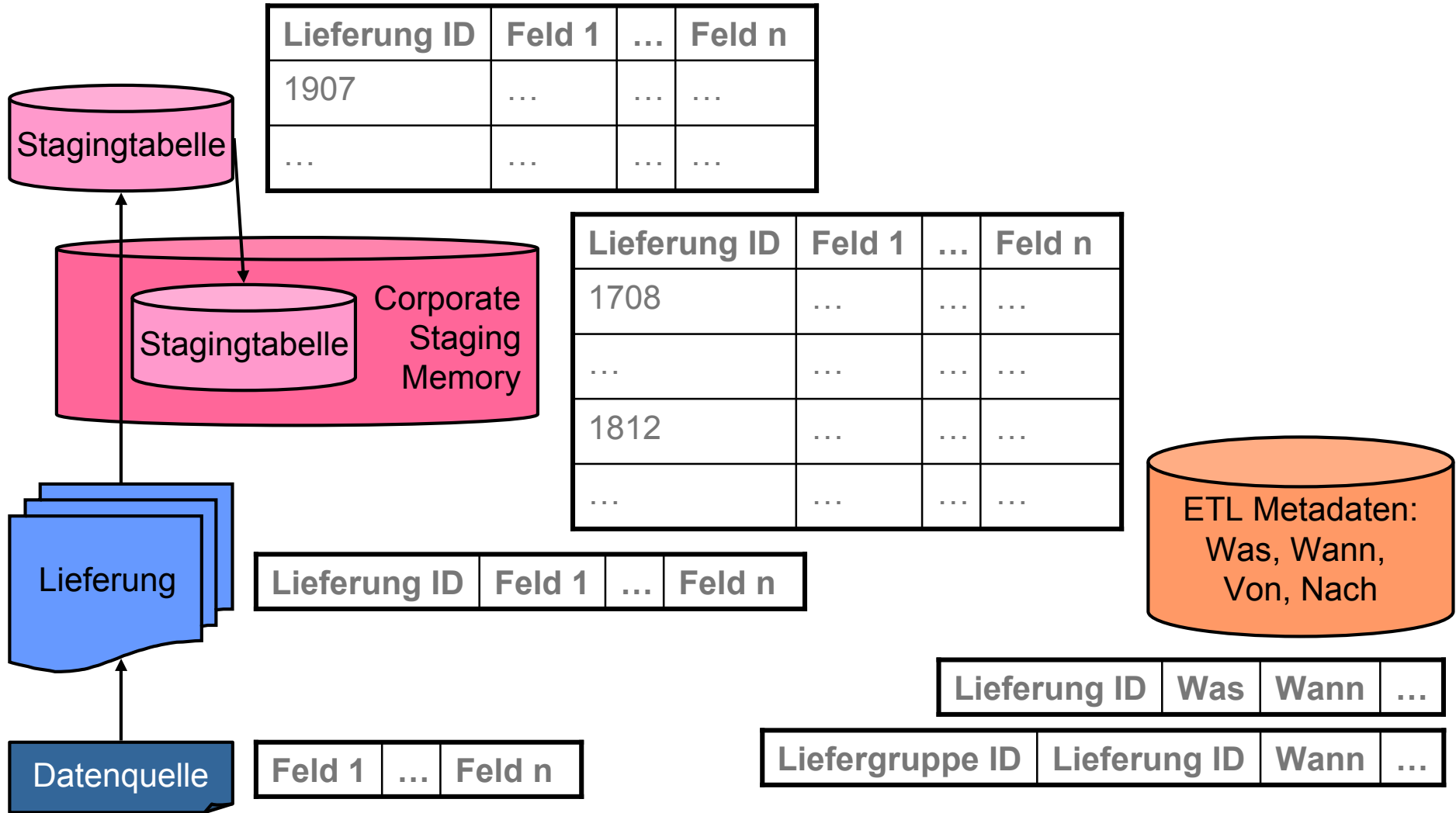
- Temporäre Zwischenspeicherung extrahierter Rohdaten vor der Integration
- Dauerhafte Speicherung der Extraktionshistorie im Corporate Staging Memory
- Kein direkter Zugriff der Endbenutzer für Analysen
- Plattform für Datentransformationen zur Entlastung von Quelldatenbanken und Data Warehouse



# Datenfluss im Staging Prozess



# Staging Area im relationalen DBMS



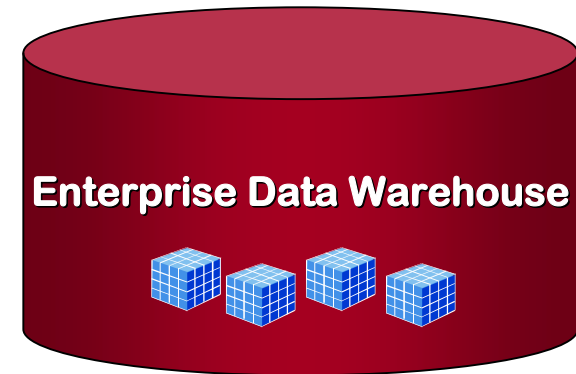
## Operational Data Store

- Integrierte Detaildaten ggf. bis auf Belegebene für operatives Reporting
- I. d. R. keine allzu lange Historie
- Ggf. Datenquelle für das Enterprise Data Warehouse
- Klassische relationale Datenablage (3. Normalform)

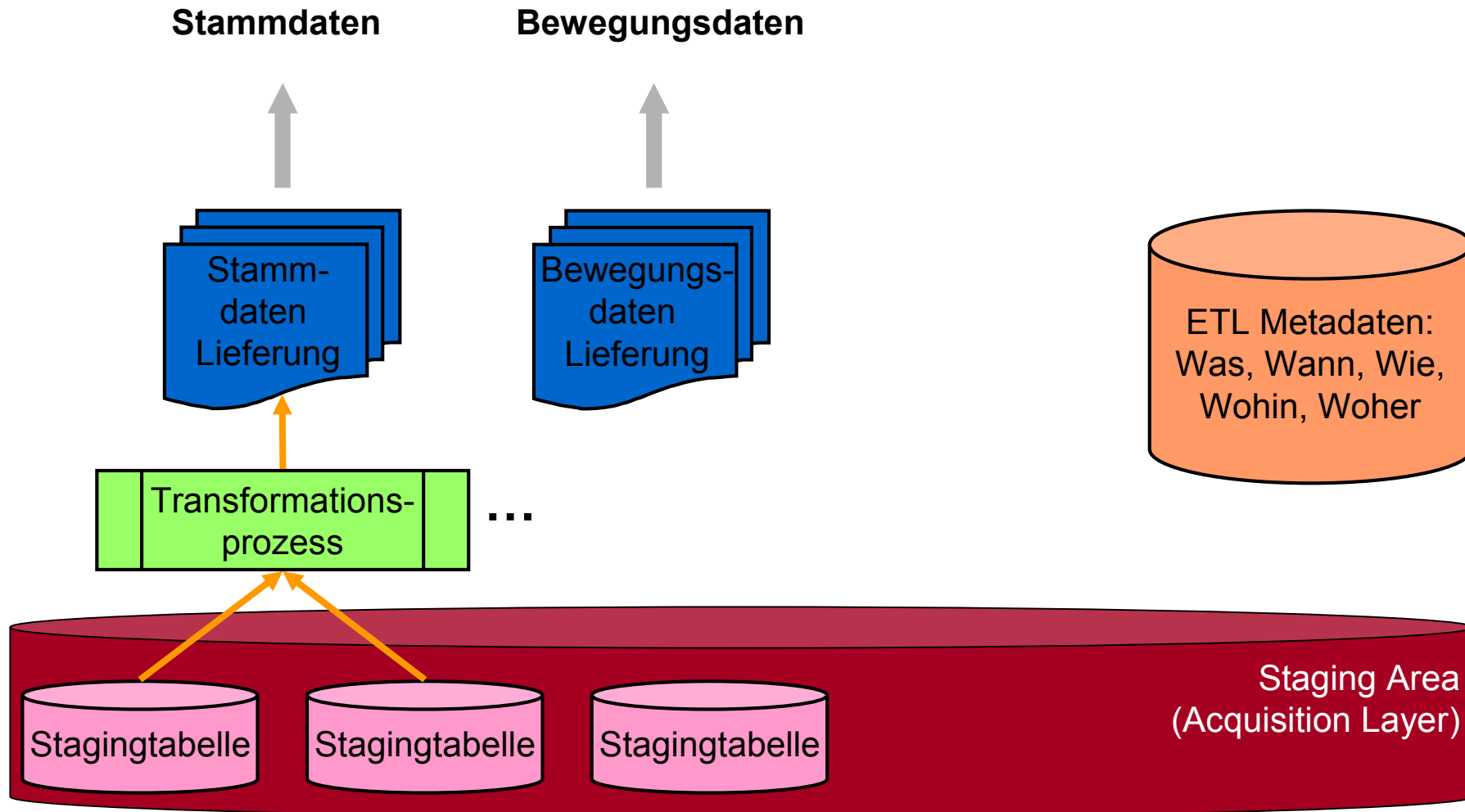


## Enterprise Data Warehouse

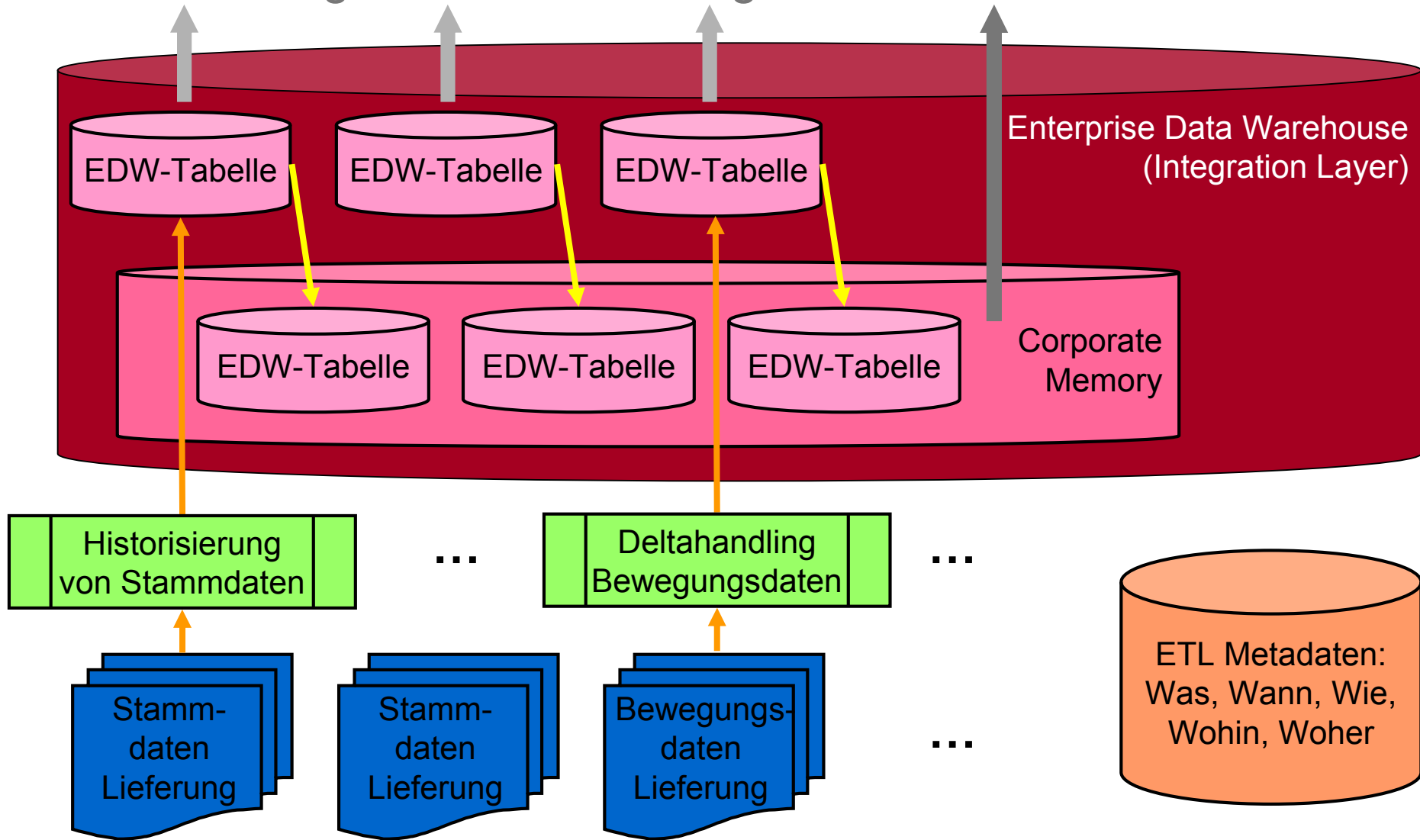
- Verdichtete Daten (z. B. Tagesbasis)
- Lange Historie (> 5 Jahre)
- Großes Datenvolumen (u. U. > 100 TB)
- Unternehmens- bzw. konzernweit ausgerichtet
- Nicht unbedingt denormalisiert, oftmals rein relational (Normalform)
- SPOT: „Single Point of the Truth“
- Entspricht Core Data Warehouse



# Transformation und Integration im EDW



# Historisierung und Deltahandling für das EDW



# Stammdaten und Bewegungsdaten im Core Warehouse

Stammdatentabelle im EDW (FG = Fachlich gültig; TG = Technisch gültig):

Lief ID	Dim ID	Blatt ID	FG Von	FG Bis	TG Von	TG Bis	SKU	Produkt	...
514	101	101	20100101	99991231	20100101	99991231	AB6	PA600	...
514	102	102	20100101	99991231	20100101	<b>20100131</b>	AX3	PAX300	...
<b>637</b>	102	102	20100101	<b>20100131</b>	<b>20100201</b>	99991231	AX3	PAX300	...
<b>637</b>	<b>137</b>	102	<b>20100201</b>	99991231	20100201	99991231	AX3	PAX300	...
514	103	103	20100101	99991231	20100101	99991231	AX4	PAX450	...

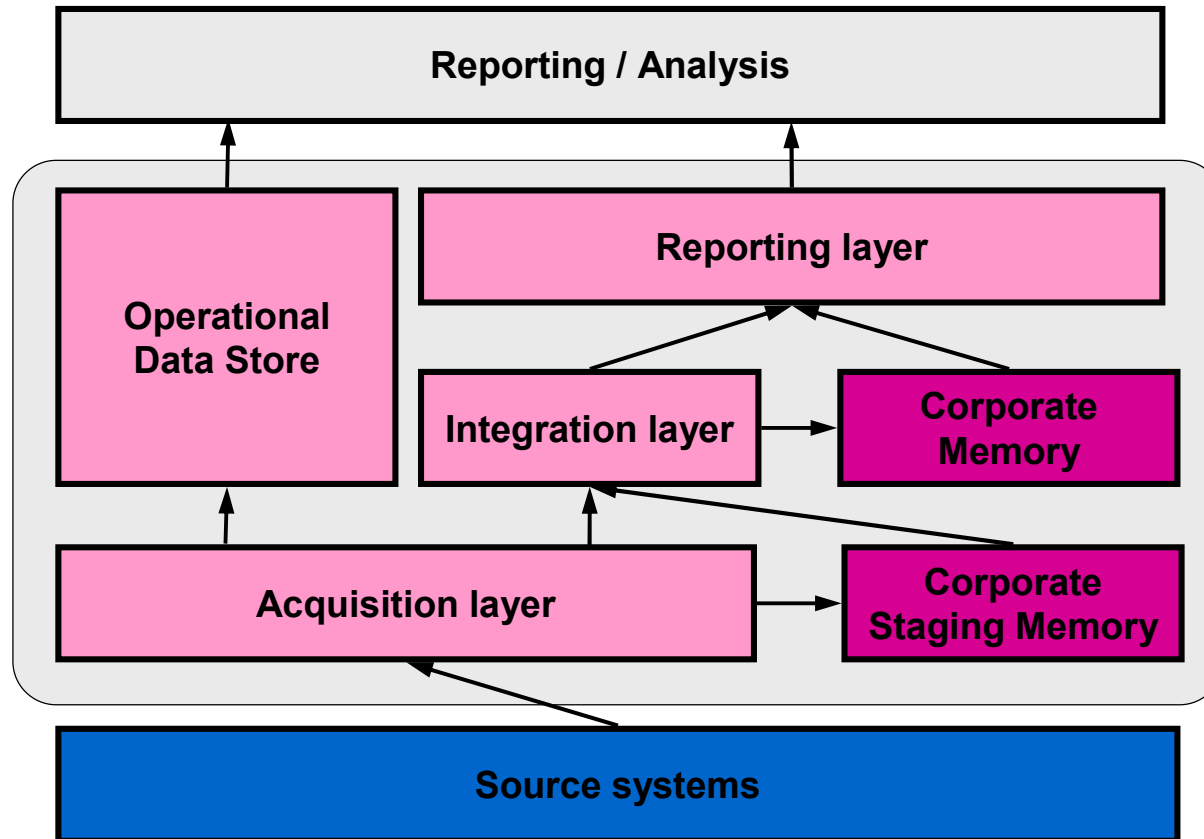
Bewegungsdatentabelle im EDW:

Lief ID	Feld 1	...	Fakten	...
547	101	...	...	...
547	102	...	...	...
...	...	...	...	...
592	101	...	...	...

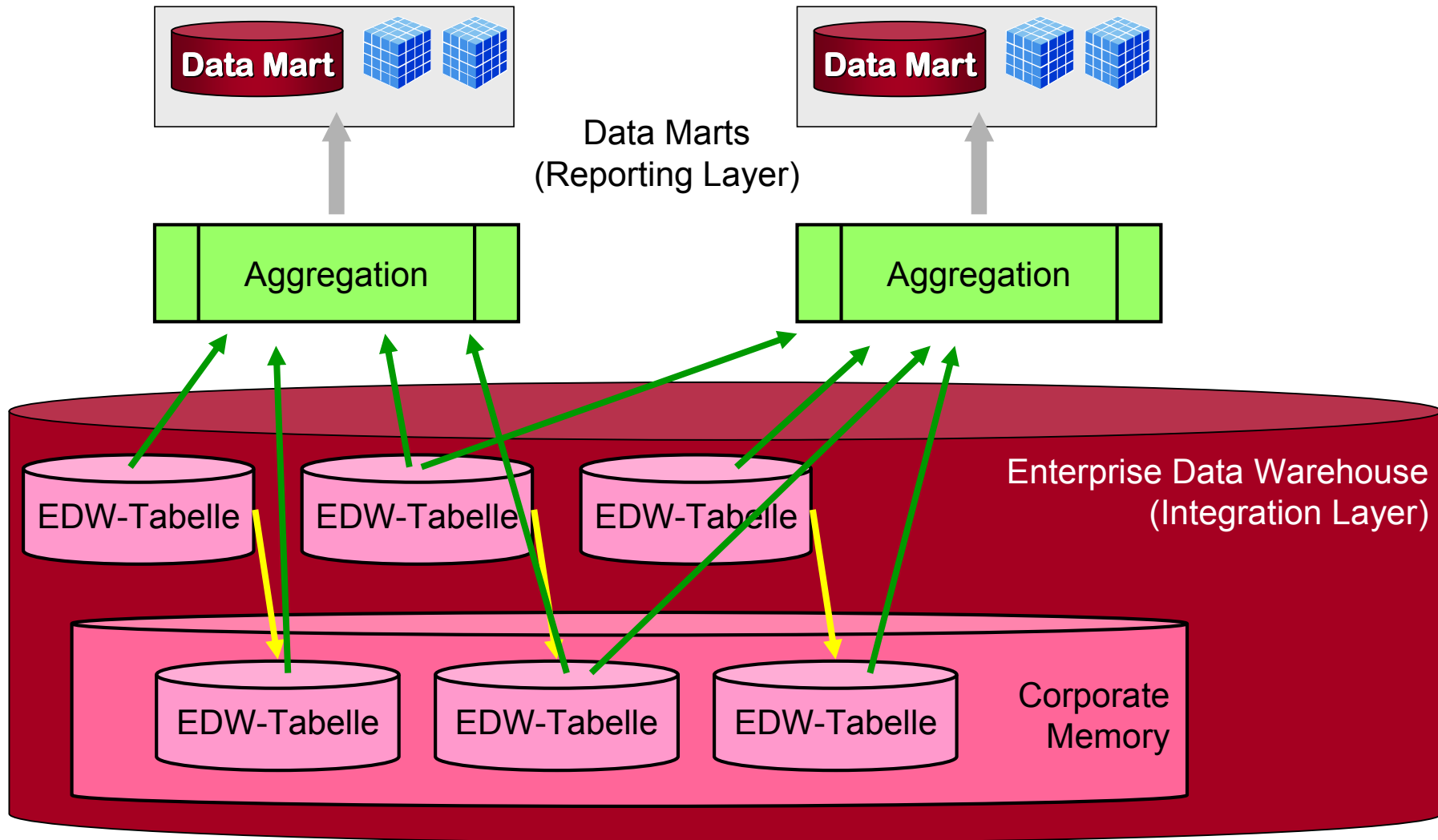
Zum Beispiel:

- Auftrag Kopfdaten
- Auftrag Positionsdaten
- Faktura Kopfdaten
- Faktura Positionsdaten
- Lieferung Kopfdaten
- Lieferung Positionsdaten

# Konzeptionelle Multi-Layer Architektur



# Nutzung des EDW: Data Marts on Demand

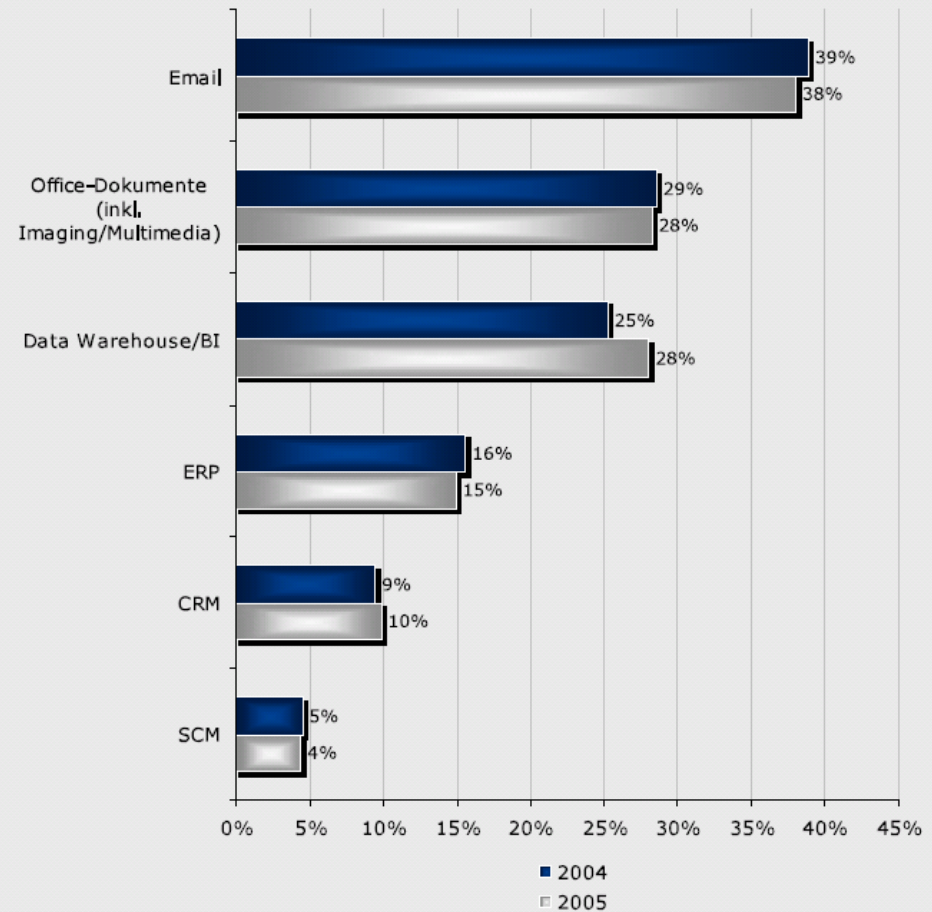


## Agenda

- Motivation
- Konzeptionelle Modellierung des Reporting Layers
- Zeitabhängigkeit im Reporting Layer am Beispiel Star Schema
- Staging und Core Data Warehouse Modellierung
- **Architektur für Enterprise Data Warehousing**

# Datenwachstum

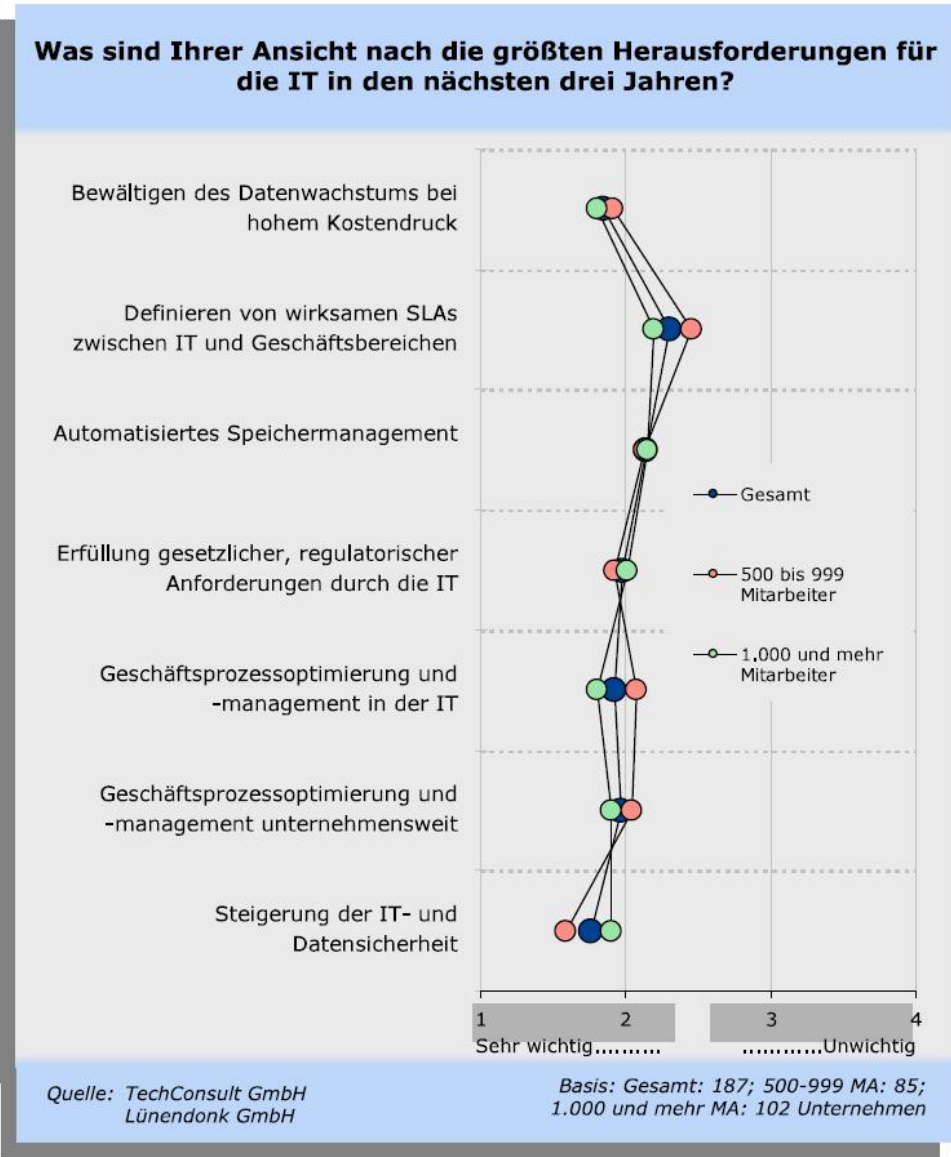
Bitte nennen Sie uns die voraussichtlichen Wachstumsraten Ihres Datenvolumens für die nächsten zwei Jahre innerhalb der folgenden Bereiche  
(Gesamt)



Quelle: TechConsult GmbH  
Lünendonk GmbH

Mehrfachnennungen möglich  
Basis: 153 Unternehmen

# Herausforderungen IT

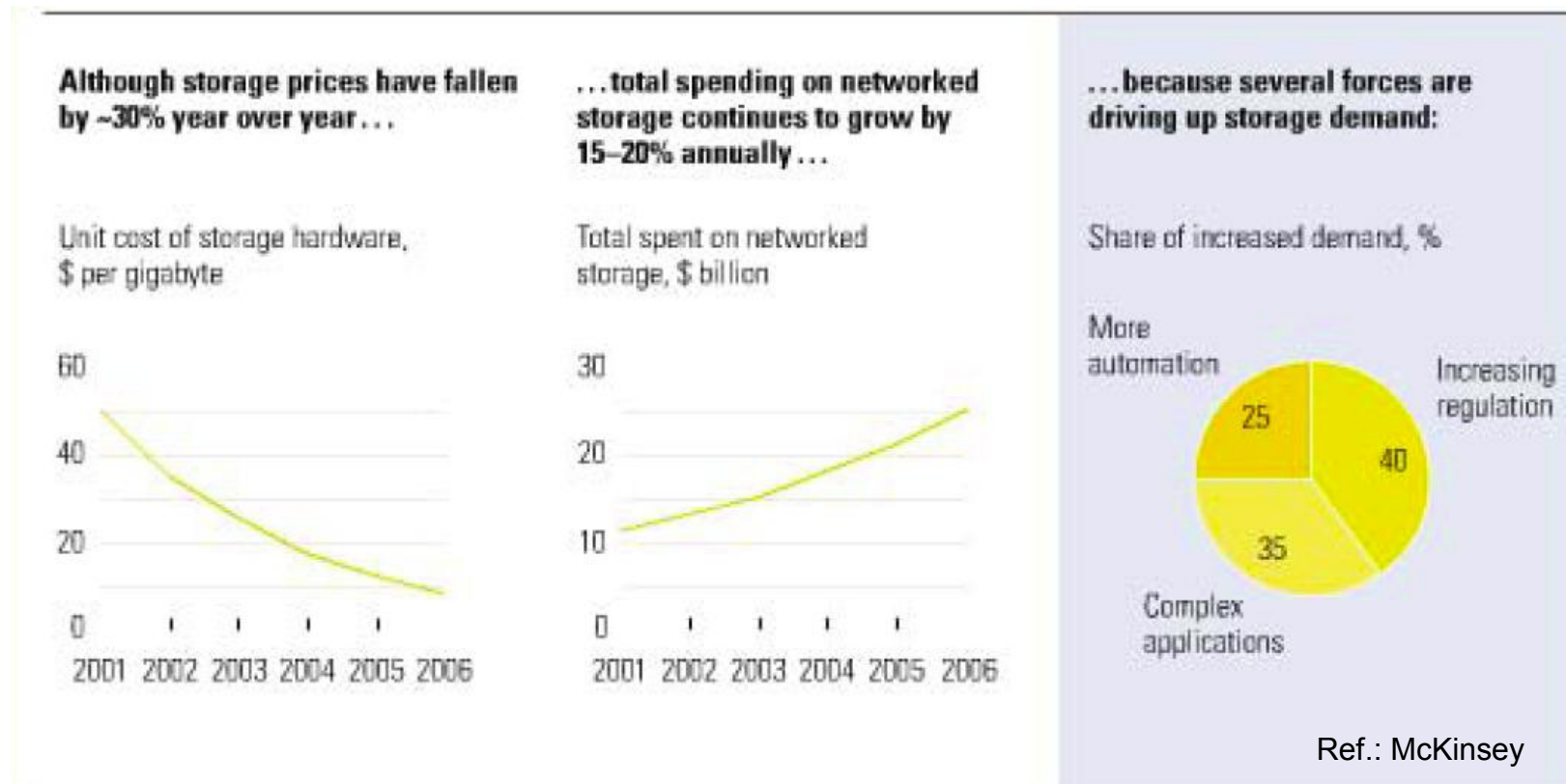


## Typische Data Warehouse Probleme

- “We Can’t Meet our Batch Windows”
  - KPI und Berichtsbereitstellung
  - Backup der Datenbank
  - Neuaufbau im Data Warehouse
- “Our Costs are Spiraling”
  - Storage / Hardware / Replikation
  - Rechenzentrumsbetrieb (Raum / Energie / Klimatechnik)
  - Datenadministration
- “The Targets Keep Changing”
  - Neue Geschäftsanforderungen
  - Spezifische Projektanforderungen
  - Externe Anforderungen / Audit / Innenrevision

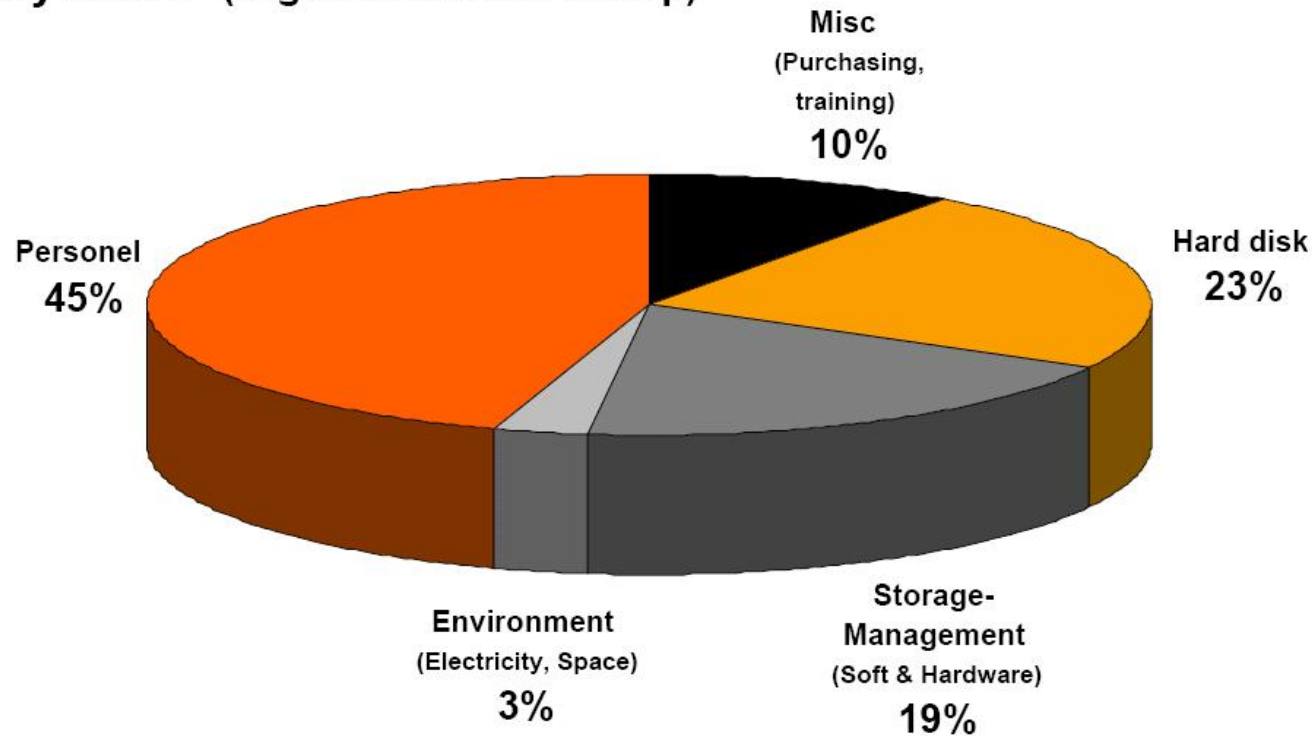
# Ausgaben für Storage ...

... (Platten, Bandsysteme, Netzwerk, Administration) wachsen jährlich um 15% bis 20%, obwohl die Speicherkosten um 30% jährlich sinken



# Speicherkosten

“The costs for data media don’t even make up a quarter of memory costs” (Giga Information Group)



“The administrative expense for 1 terabyte of memory is five to seven times as high as the memory cost itself” (Dataquest/Gartner)

## Herausforderung fürs Data Warehouse

“With projected compounded annual growth rates for databases exceeding 125%, organizations face two basic options:

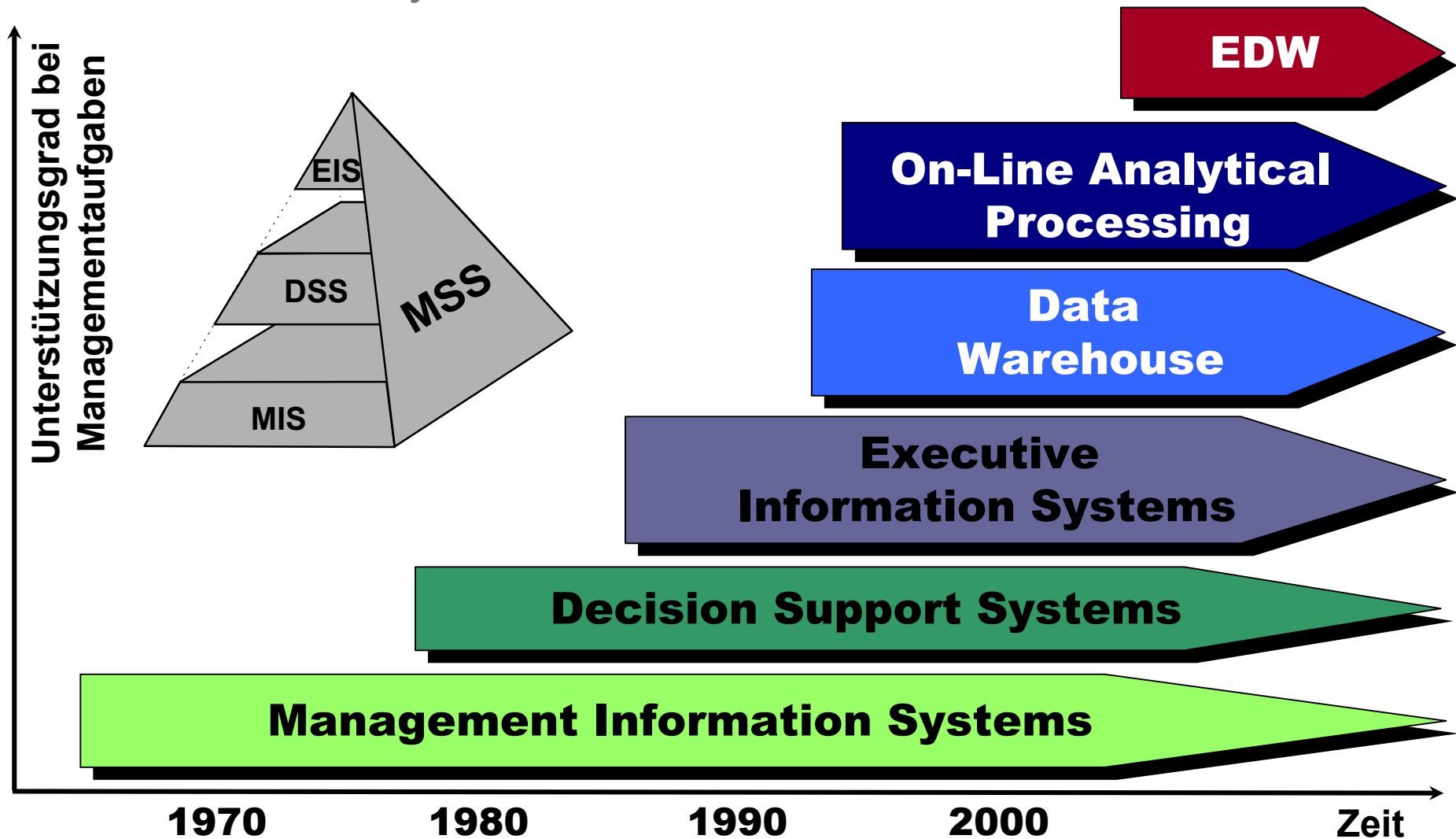
1) Continue to grow the infrastructure (e.g., server size, storage capacity)

OR

2) *Develop processes [and architectures] to separate dormant [archive-ready] data from active data.”*

Meta Group Report  
Databases on a Diet

# Historie der BI-Systeme



## Data Warehouse-Konzept

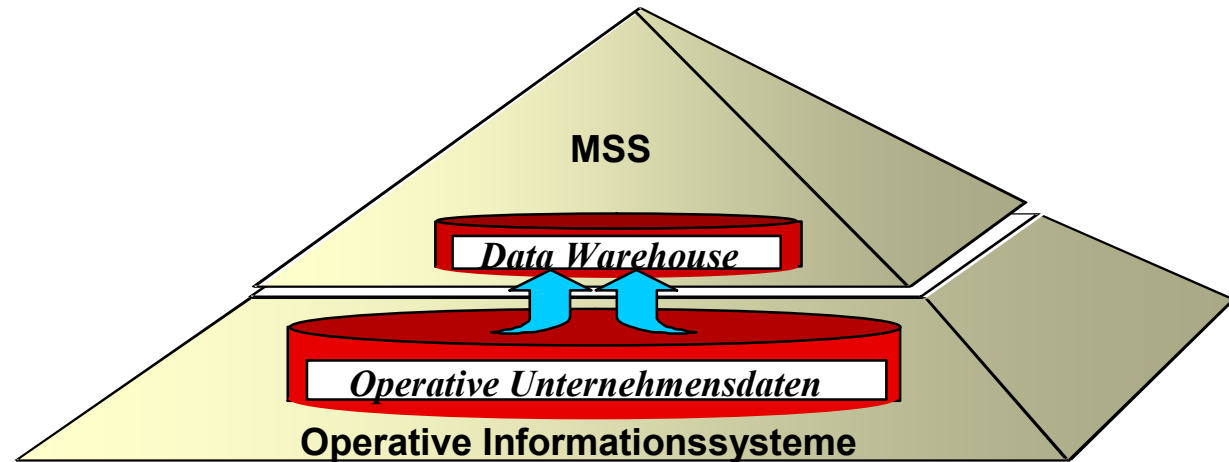
### Data Warehouse (DW) =

*“A subject oriented, integrated, nonvolatile, time variant collection of data organized to support management needs”*

W. H. Inmon, *Building the Data Warehouse*, New York u. a. 1993, S. 29.

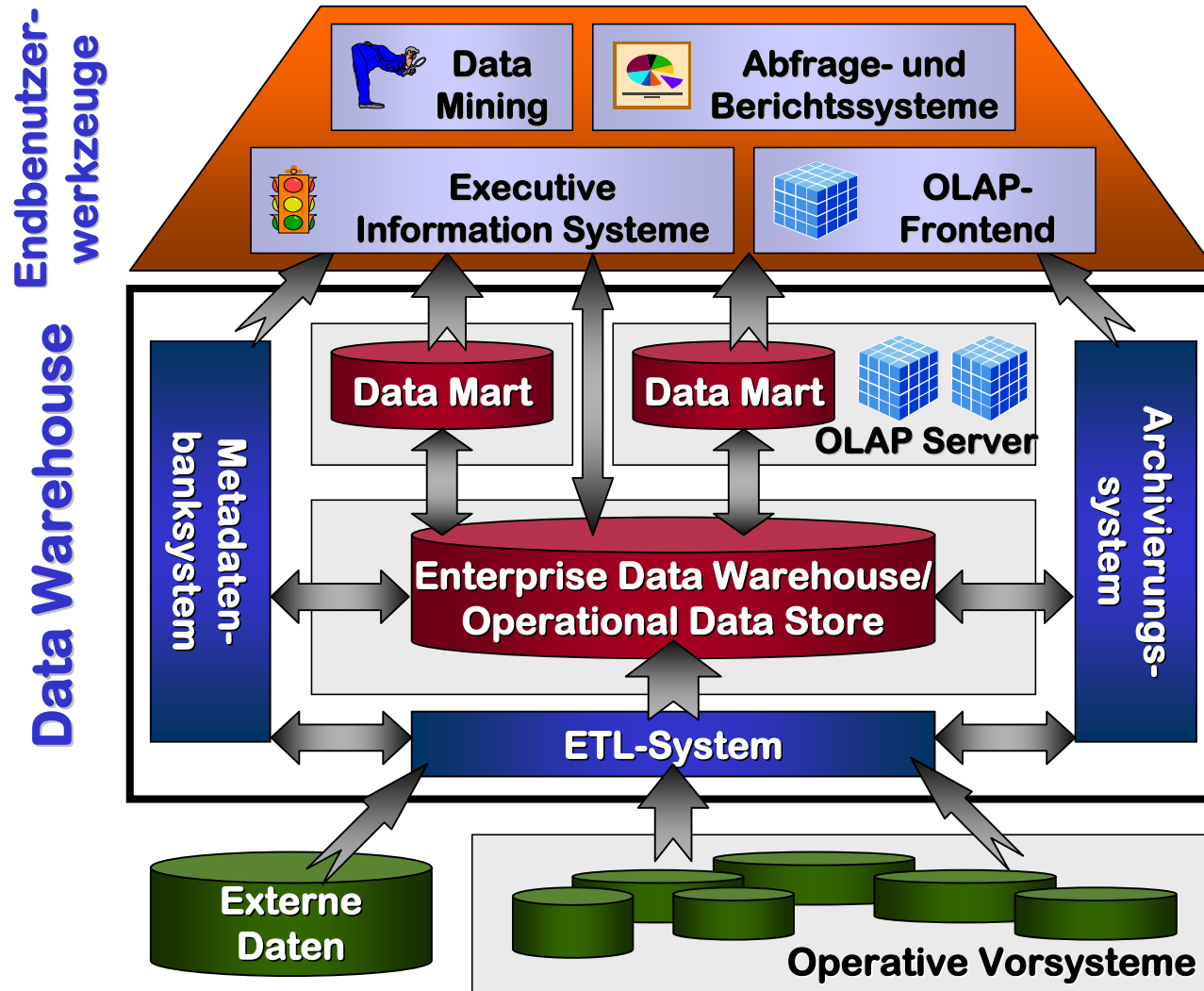
- subject oriented → Themenorientierung
- integrated → Vereinheitlichung
- nonvolatile → Dauerhaftigkeit, Stabilität
- time variant  
Informationen → Zeitorientierung der
- management needs → Analyse und  
Entscheidungsunterstützung

# Data Warehouse

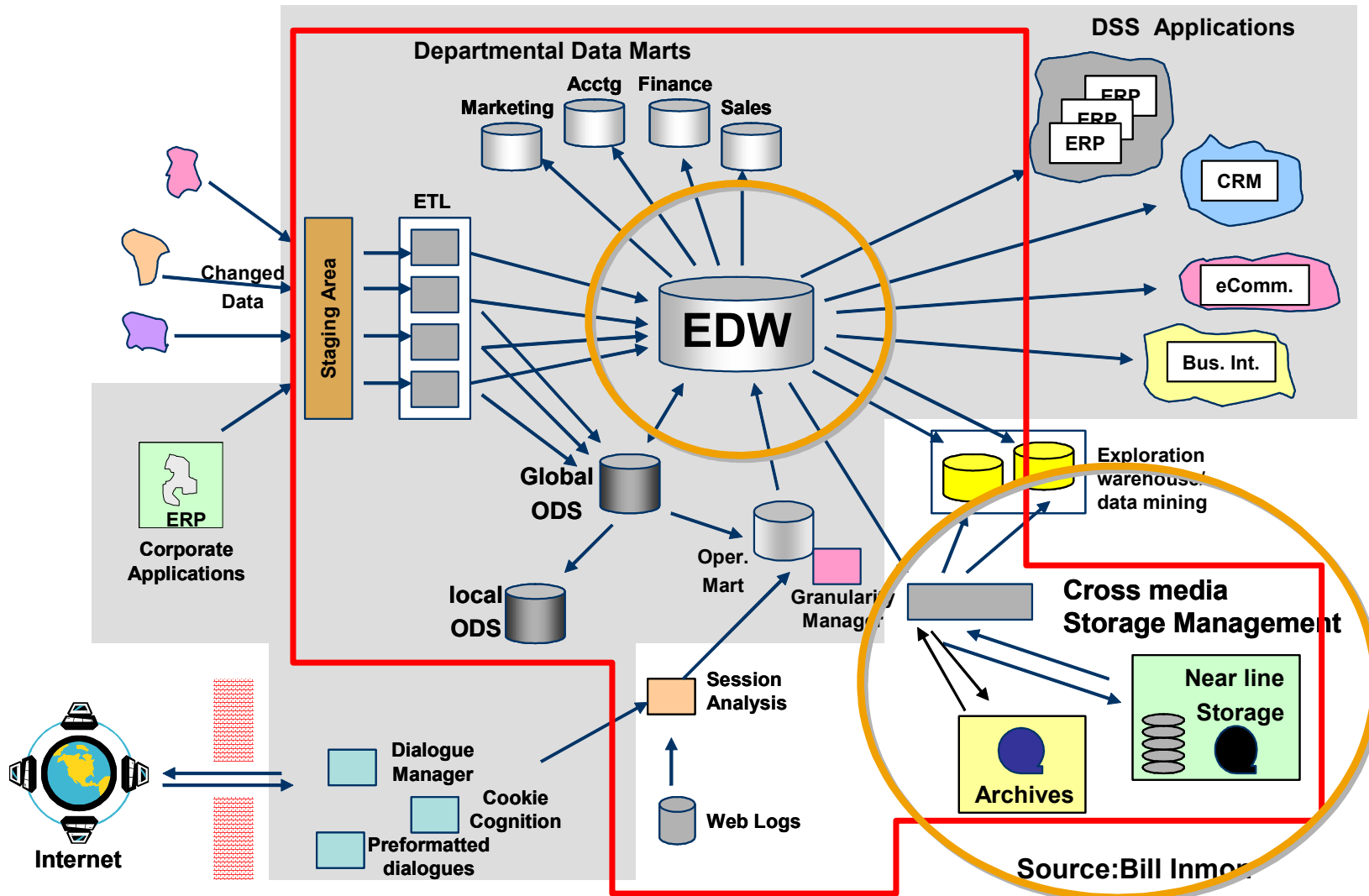


- umfasst die **Serverkomponenten** einer Systemlösung, die für die **unternehmensweite Datenversorgung** der Front-End-Systeme zur **Informationsbereitstellung und Entscheidungsunterstützung** betrieblicher Fach- und Führungskräfte zuständig sind,
- ist **physikalisch** von den operativen Vorsystemen **getrennt** und
- baut lediglich zum Zweck der **periodischen Datenaktualisierung** bzw. -ergänzung Verbindungen zu den operativen DV-Systemen auf.

# Aktuelles DW-Architektur-Konzept

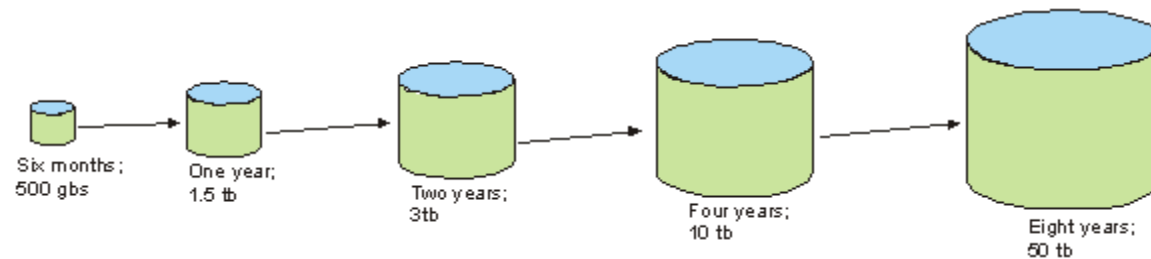


# Bill Inmon's Enterprise Data Warehousing Concept

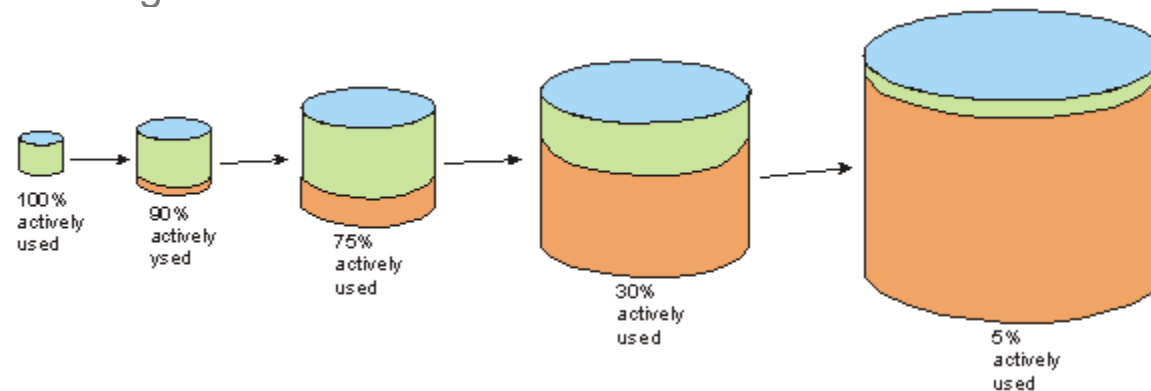


# Datenzugriff vs. Datenwachstum

- Typisches Datenwachstum



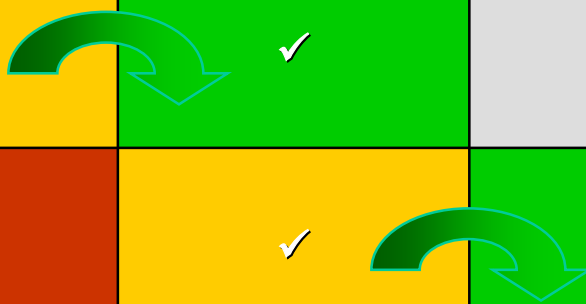
- Typischer Datenzugriff vs. Datenwachstum



- Mit zunehmendem Datenwachstum ändert sich die Zugriffswahrscheinlichkeit signifikant

# Information Lifecycle Management mit Near-Line Storage

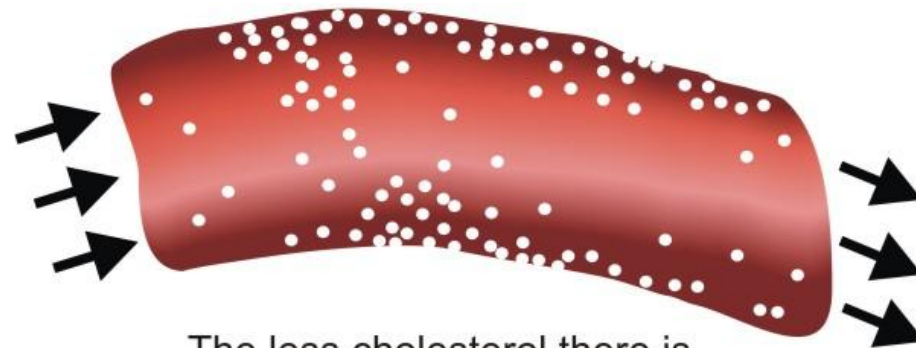
	Online Datenbank	Near Line Speicherung	Daten Archivierung
Häufiger Datenzugriff "Updates"	✓		
Wenig Datenzugriff	✓	✓	
Kaum Datenzugriff	✓	✓	✓



## Bill Inmon's Meinung zu Performance Problemen im Kontext Near-Line Storage

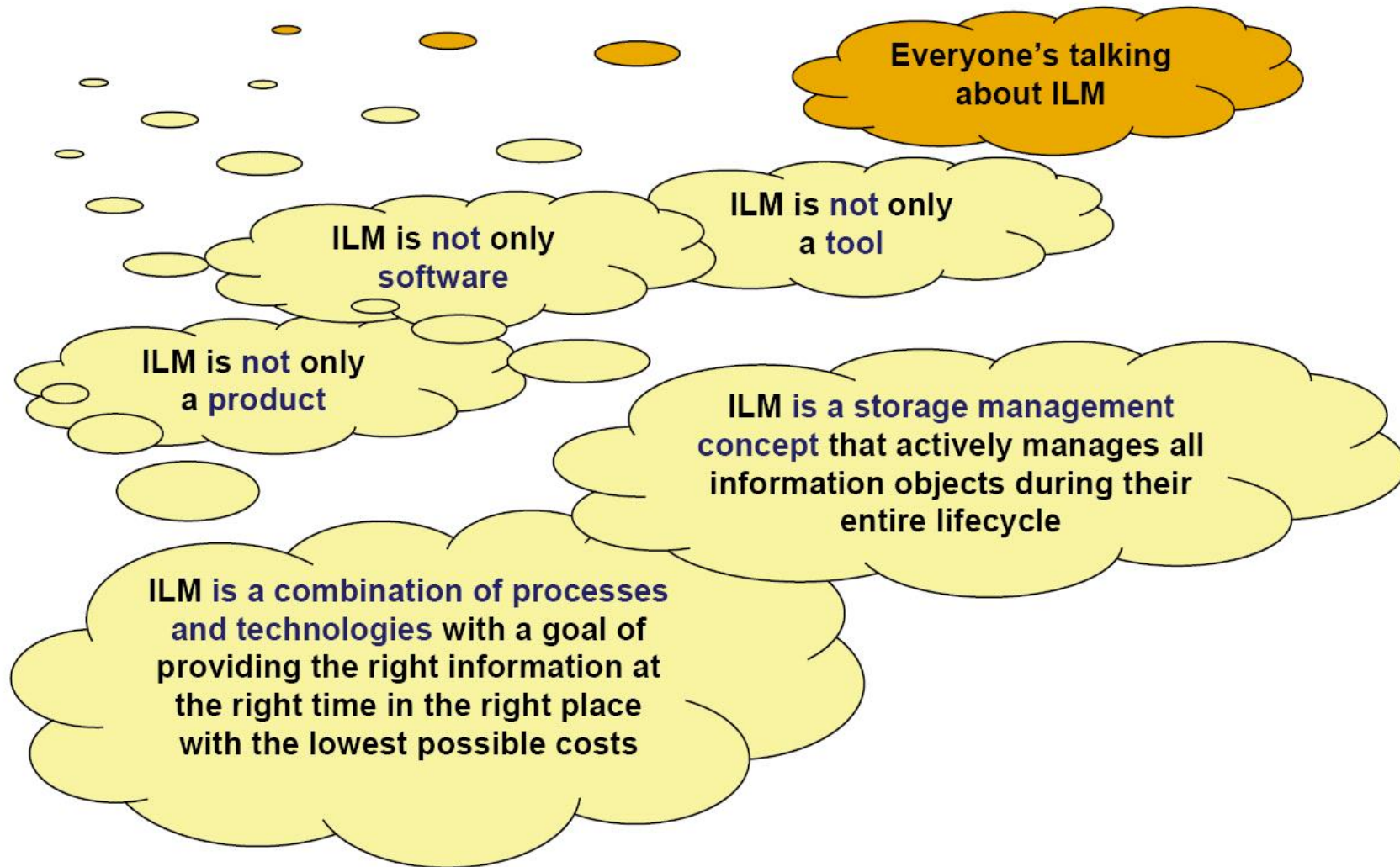
**“Indeed, leaving infrequently accessed data on disk storage greatly **HURTS** performance [...] because **mixing infrequently used data with actively used data is like adding lots of cholesterol into the blood stream.**”**

Information Lifecycle Management for Data Warehousing:  
Matching Technology to Reality  
By W.H. Inmon (2005)



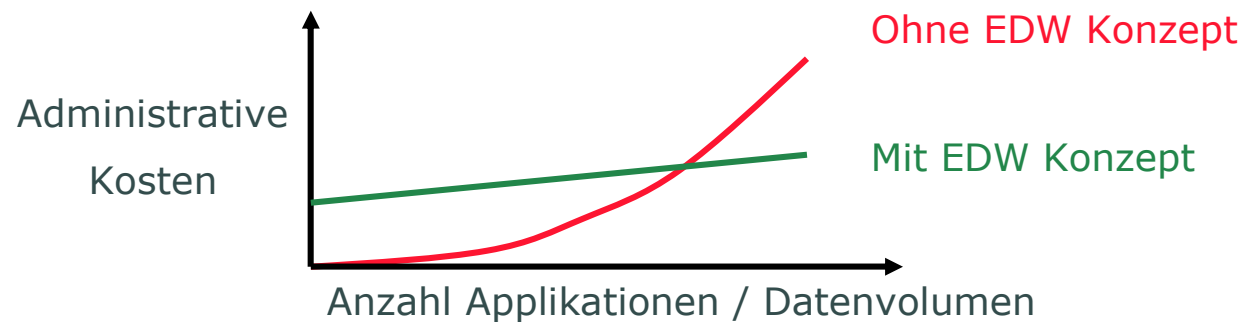
The less cholesterol there is,  
the more efficient the flow of blood

# Information Lifecycle Management (ILM)

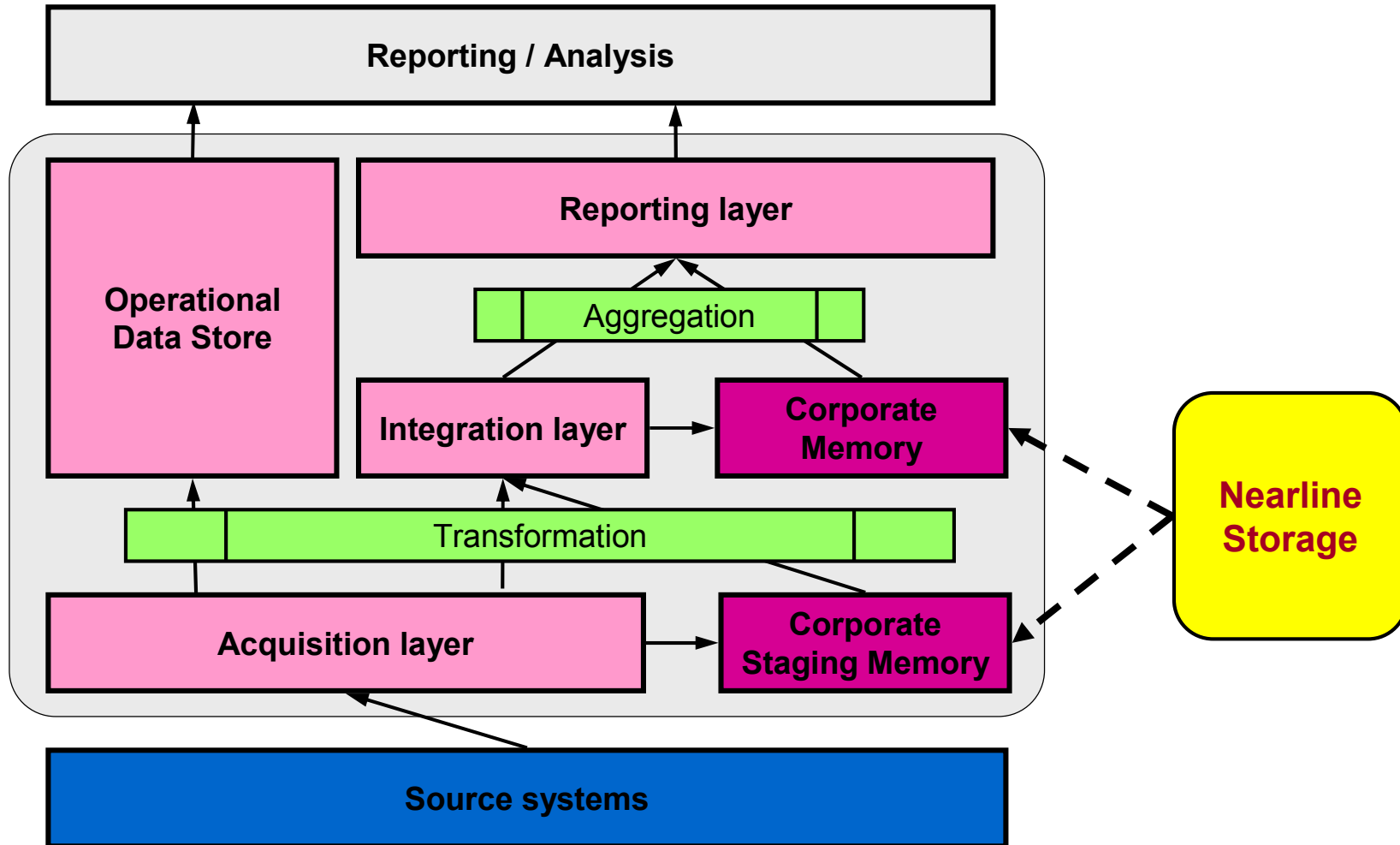


## Motivation für ein EDW Konzept – *Anticipating the unknown*

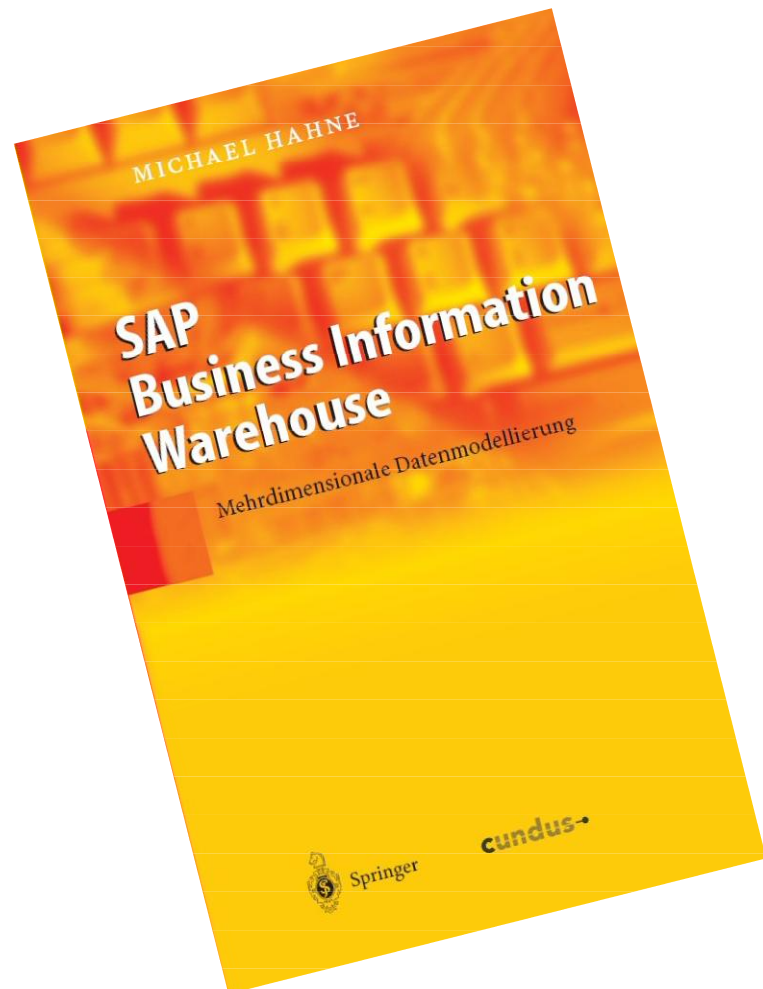
- Datenwachstum
- Zunehmende Anzahl an Applikationen
- Dies führt zu
  - Steigenden administrativen Kosten
  - Höheres Ausfallrisiko von Anwendungen
  - Gefahr eines Totalausfalls



# Konzeptionelle Multi-Layer Architektur



Weitere Informationen...



**Die zweite Auflage  
erscheint in Kürze!**

Sowie unter:  
[www.hahneconsulting.de](http://www.hahneconsulting.de)  
[www.hahneonline.de](http://www.hahneonline.de)  
[www.t-adapt.de](http://www.t-adapt.de)

## Kontakt

# hahne consulting gmbh

IT solutions for your business

Hahne Consulting GmbH  
Dr. Michael Hahne  
Geschäftsführer / Managing Director

fon: 0049 671 20 27 180

fax: 0049 671 92 03 661

email: [michael@hahneconsulting.de](mailto:michael@hahneconsulting.de)

[www.hahneconsulting.de](http://www.hahneconsulting.de)

[www.hahneonline.de](http://www.hahneonline.de)

[www.t-adapt.de](http://www.t-adapt.de)