Data Warehouse und ETL

Einführung und Überblick

Data Warehouse – Definition I

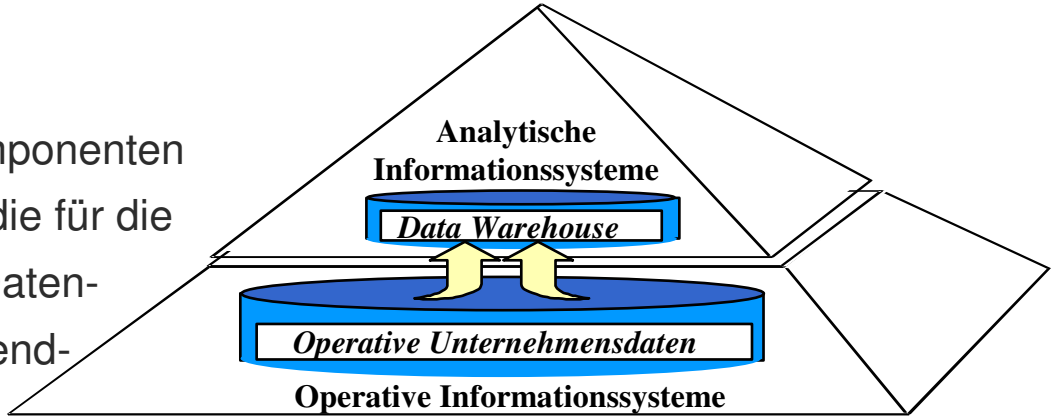
- “A subject-oriented, integrated, non-volatile, time-variant collection of data organized to support management needs”

  Inmon, Database Newsletter ‘92.

• subject-oriented	→	Themenorientierung, Quellenunabhängigkeit
• integrated	→	Struktur- und Formatvereinheitlichung
• non-volatile	→	Dauerhaftigkeit, Stabilität
• time-variant	→	Zeitraumbezug der Information
• management needs	→	Analyse und Entscheidungsunterstützung

*Data Warehouse – Definition II

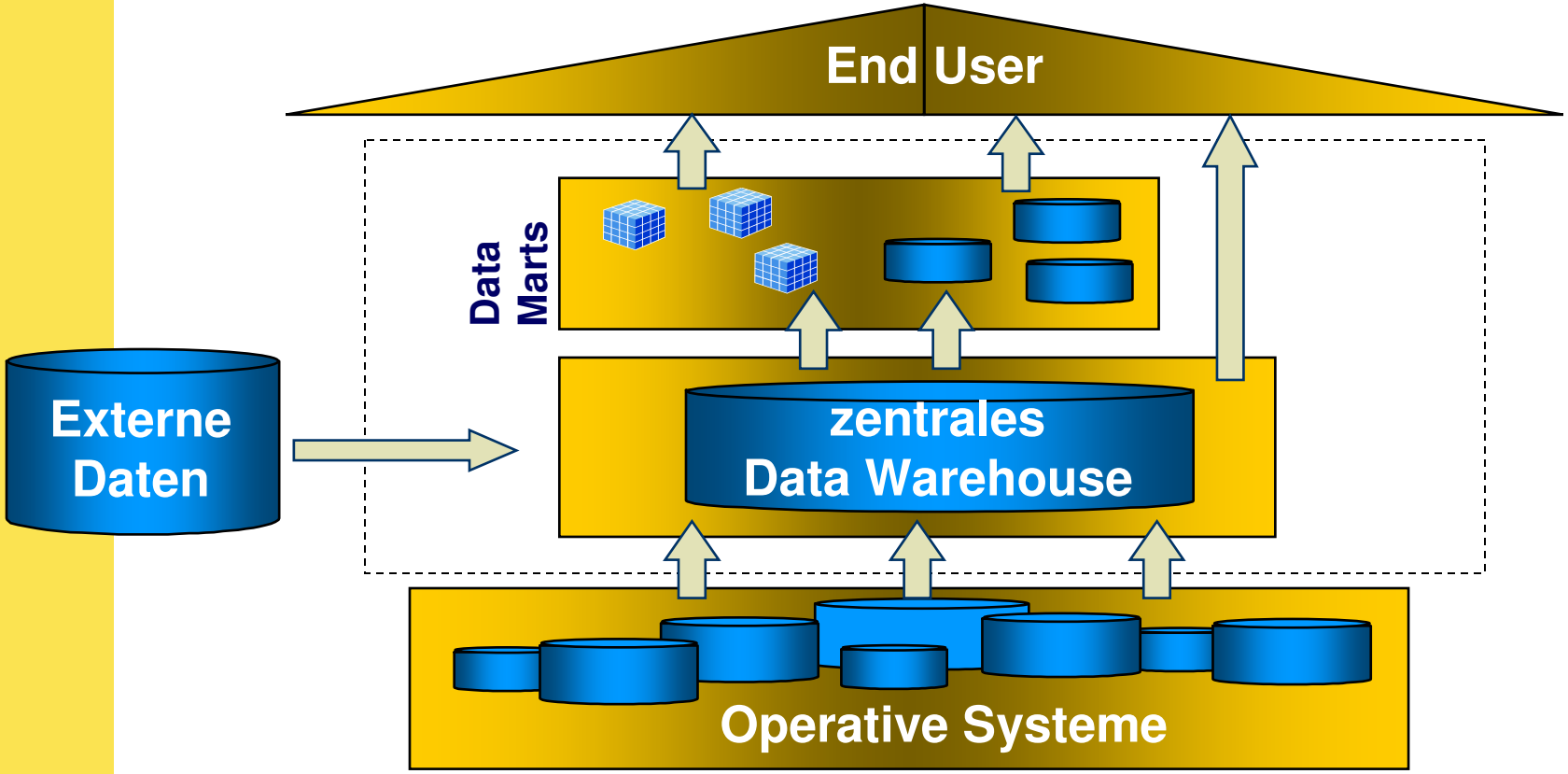
- Ein Data Warehouse
  - umfaßt die Serverkomponenten einer Systemlösung, die für die unternehmensweite Daten-versorgung der Frontend-Systeme zur Infor-mationsversorgung und Entscheidungsunterstützung betrieblicher Fach- und Führungskräfte zuständig sind,
  - ist physikalisch von den operativen Vorsystemen getrennt und
  - baut lediglich zum Zweck der periodischen Datenaktualisierung bzw. - ergänzung Verbindungen zu den operativen DV-Systemen auf.



Charakteristika von OLTP und Data Warehouse-Lösungen

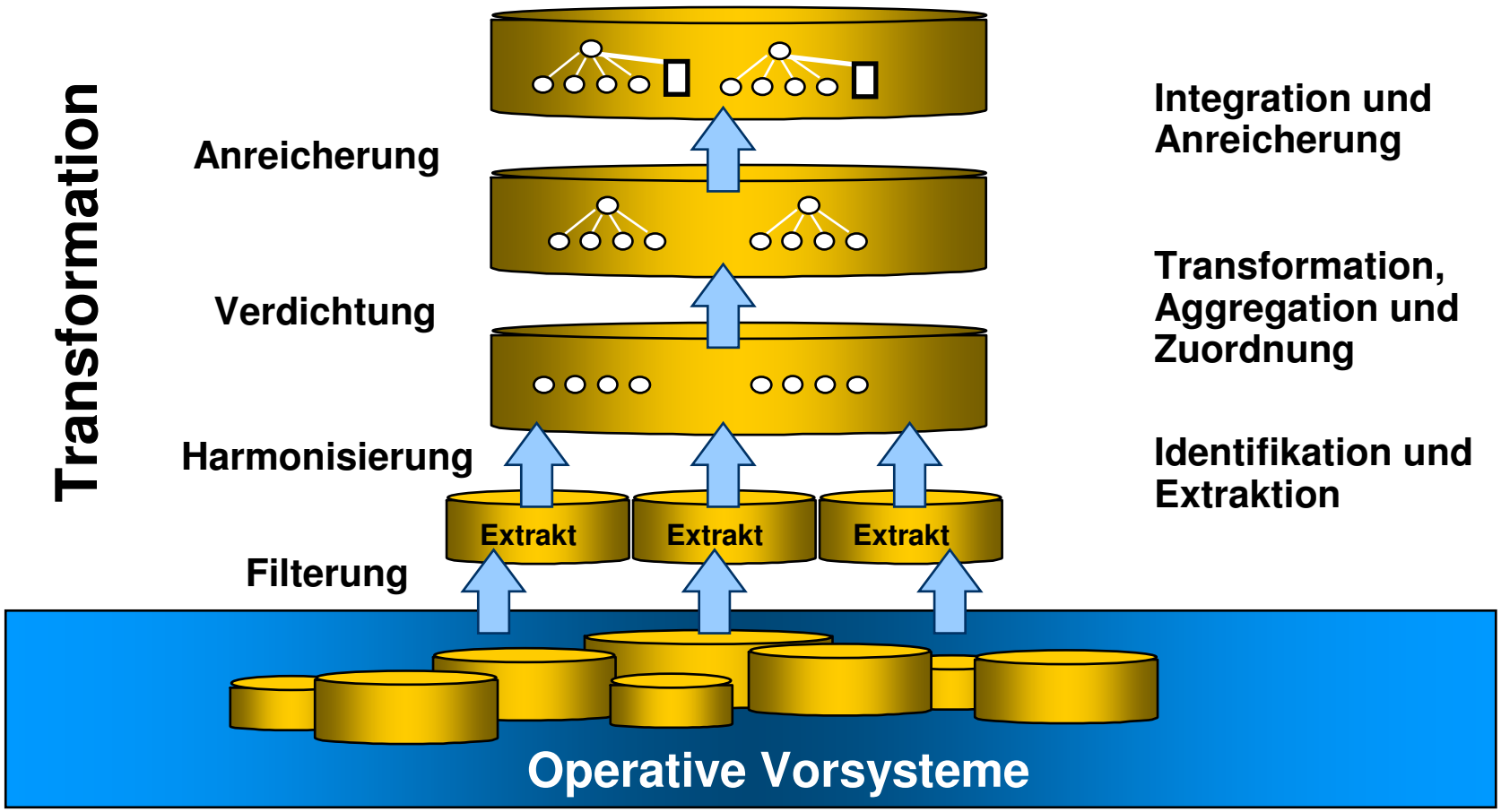
Charakteristika	Operative Systeme	Data Warehouse
Änderungen	sehr viele, kleine Transaktionen	nur durch Ladevorgänge
Zugriffsform	lesend, schreibend	lesend
DB-Größe	Gigabytes	Gigabytes bis Terabytes
Aktualität	jederzeit aktuell	historisch
Dateninhalte	Prozessorientiert	nach Aufgabenbereich
Datenstrukturen	redundanzfrei	mit Redundanzen
Nutzungsintensität	gleichbleibend	schwankend
Abfragen	statisch, vorhersehbar	dynamisch
Datenquellen	intern	intern und extern

Prinzipielle Architektur



ETL-Prozess

ETL = *E*xtraktion, *T*ransformation, *L*aden



Begriffserklärung

Metadaten

- sind *Daten über Daten*
- beschreibende und klassifizierende Angaben zum Problemdatenbestand
- Angaben über Datenstrukturen und Konsistenzbedingungen
- gehören zu den kritischen Erfolgsfaktoren eines DWH

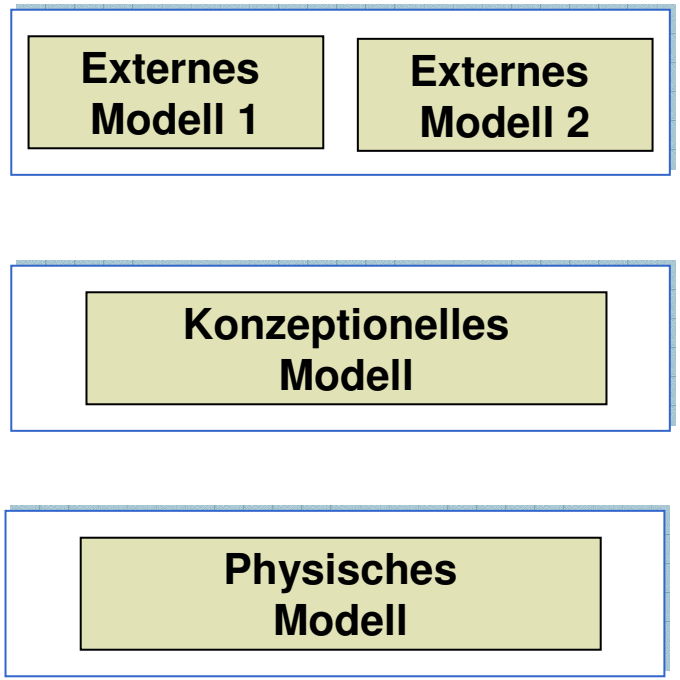
Beispiele im DWH Umfeld:

- Extraktionsregeln (Ziel- und Quelltabelle und –spalte, ...)
- Transformationsregeln (Filterung, Rundung, Typkonvertierung, ...)
- Ausnahme- und Alert-Regeln
- Verdichtungsregeln (Summierung, statistische Funktionen, ...)
- Mappingregeln
- Scheduling-Regeln etc.

*Datenintegrität

Im Bereich der Datenintegrität (lat.: Makellosigkeit) werden Fragen behandelt, die sich mit der Korrektheit der Daten befassen.

Architektur von DBS



Integritätsziele

Datenschutz

(Zugriff durch Befugte im Rahmen der definierten Befugnis)

Datenkonsistenz

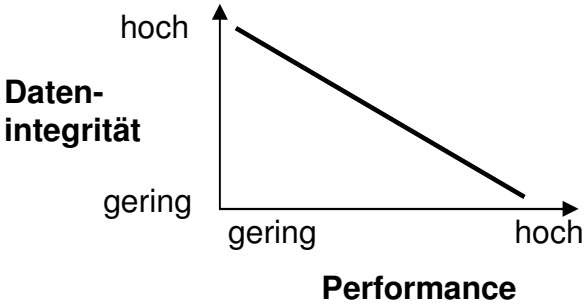
(Widerspruchsfreiheit der Daten zu sich selbst und den Vereinbarungen des konzeptionellen/der externen Datenmodelle)

Datensicherheit

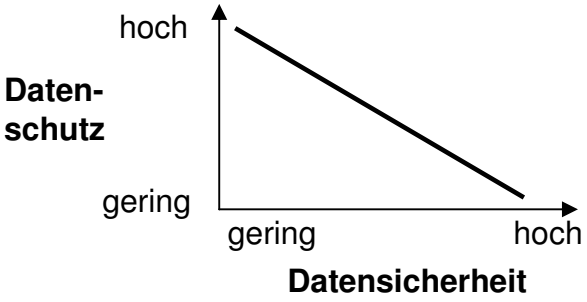
(Daten zu jedem Zeitpunkt in korrekter Form nutzbar)

*Datenintegrität

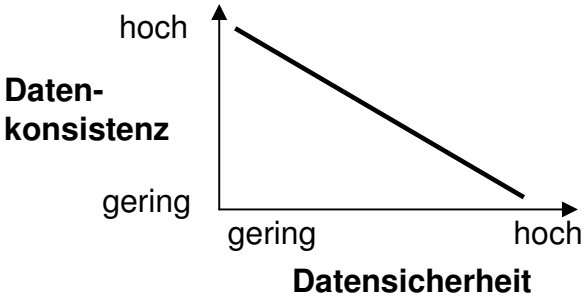
Problem der Zielkonflikte



Maßnahmen zur Datenintegrität gehen in der Regel zu Lasten der Performance des DB-Systems



Maßnahmen zur Datensicherheit gehen in der Regel zu Lasten des Datenschutzes



Maßnahmen zur Datensicherheit gehen in der Regel zu Lasten des Datenkonsistenz

Begriffserklärung

Dimensionen

sind die Kriterien, nach denen analysiert werden kann. Die häufigste Dimension ist die *Zeit*.

Kennzahlen

sind fachliche Variablen, die analysiert werden sollen. Diese existieren nur in Abhängigkeit von Dimensionen. In der Regel handelt es sich um quantitative Faktoren betriebswirtschaftlichen Charakters, wie *Umsatz, Gewinn, Anzahl Mitarbeiter* etc.

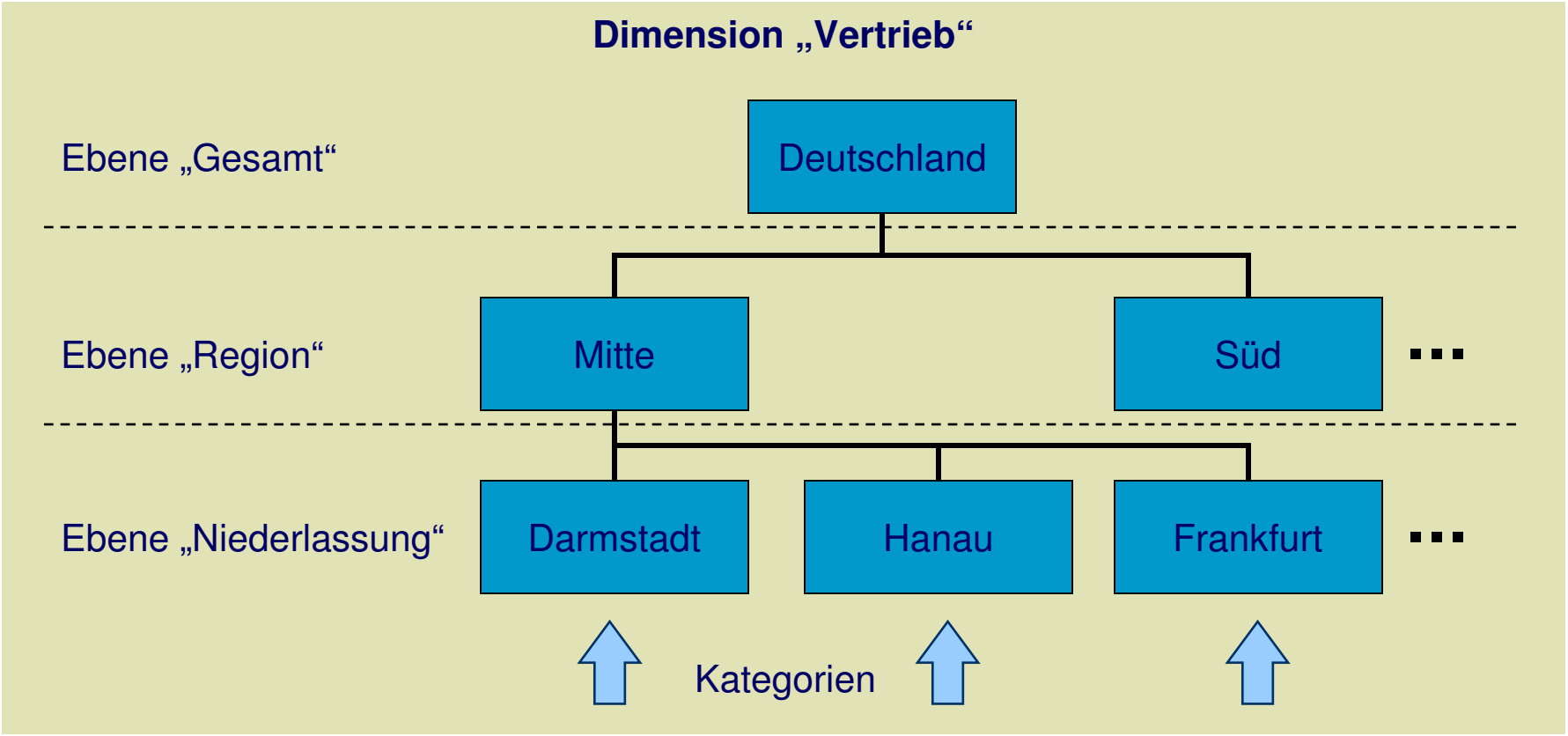
Numerische und additive Kennzahlen sind am besten darstellbar.

Dimensionen

- Dimensionen bestehen aus **Ebenen**
- Die Dimension *Vertrieb* kann z.B. folgende Ebenen haben:
  - Gesamt
  - Region
  - Niederlassung
  - Filiale
  - Mitarbeiter
- Eine Dimensionsebene besteht aus **Kategorien**
- Kategorien sind konkrete fachliche Begriffe
- Die Ebene *Region* kann aus folgenden Kategorien bestehen:
  - Nord, West, Mitte, Ost, Süd, ...
- Die Ebene *Niederlassung* kann aus folgenden Kategorien bestehen:
  - Frankfurt, Darmstadt, Hanau, ...
- Die Kategorien dienen der Beschriftung der Zeilen und Spalten in der OLAP Analyse.

Der Kategoriebaum

- Welche Kategorien gehören zu welcher „Vaterkategorie“ in der nächsthöheren Dimensionsebene?



*Dimensional Design Process

1. Geschäftsprozess auswählen

2. Die Granularität festlegen

3. Die Dimensionen auswählen

4. Die Kennzahlen identifizieren

*Schritte für die Umsetzung (nach Wieken)

(1) *Basismodell*

Klärung der DWH-Relevanz von Daten

(2) *Historisierungsmodell*

Einführung des Faktors Zeit und historischer Daten

(3) *Dimensionsmodell*

Aufbau von Strukturinformationen (Dimensionsdaten)

(4) *Aktualisierungsmodell*

Anpassung an Aktualisierungszeitpunkte

(5) *Qualitätsmodell*

Festlegung von Regeln für Konsistenz- und Plausibilitätschecks

(6) *Zugriffsmodell*

Aufbau von Zugriffsstrukturen (Umgruppierungen, Verdichtungen)

OLAP - Begriffserklärung

OLAP (Online Analytical Processing)

ist eine Ergänzung des Data Warehouse-Konzeptes zur analytischen multidimensionalen Datenauswertung, wobei es bei den Konzepten zu inhaltlichen Überschneidungen kommt. Der Begriff impliziert schnelle Abfrageergebnisse und eine intuitive Bedienung der Oberfläche.

OLAP- Gegenstand, Ziel und Philosophie

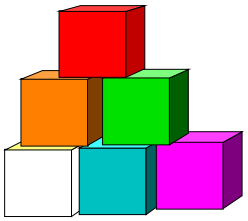
Gegenstand: Online Analytical Processing [OLAP] bietet einen endanwender-orientierten Gestaltungsrahmen für den Aufbau von Systemen zur Unterstützung dispositiver bzw. analytischer Aufgaben.

Philosophie: Multidimensionale Sichtweisen auf unternehmensinterne und -externe Datenbestände gewährleisten brauchbare Näherungen an das mentale Unternehmensbild des Managers.

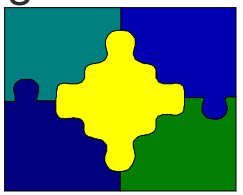
Ziel: Systeme, die es dem Endbenutzer erleichtern, selbständig, rasch und mit geringem Aufwand sowohl individuelle Ad-hoc-Auswertungen als auch komplexe betriebswirtschaftliche Analysen durchzuführen.

OLAP - Regeln I

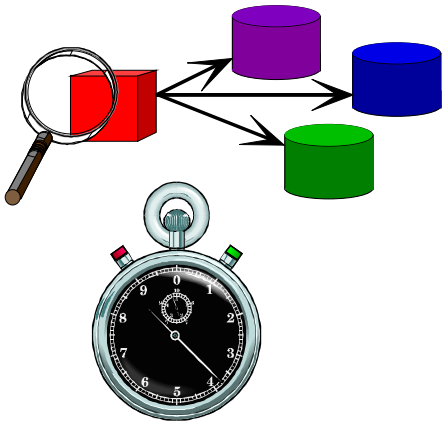
- Multidimensionale, konzeptionelle Sicht auf die Daten



- Transparenz (nahtlose Integration)
- Zugänglichkeit heterogener Datenbasen mit logischer Gesamtsicht



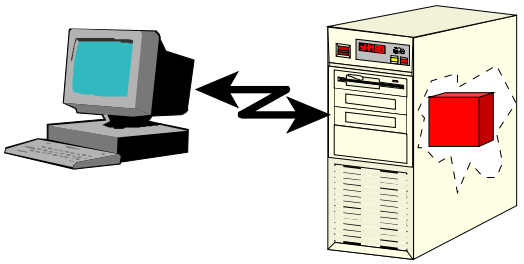
- Stabile, volumenunabhängige Antwortzeiten



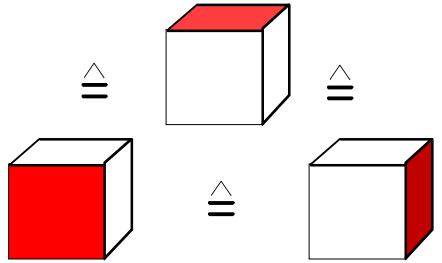
Quelle: E. Codd, S. Codd, C. Shelley: Providing OLAP to User-Analysts, White Paper, 1993

OLAP - Regeln II

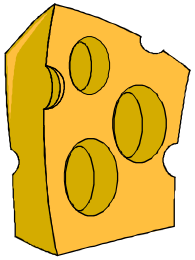
- Client-Server Architektur

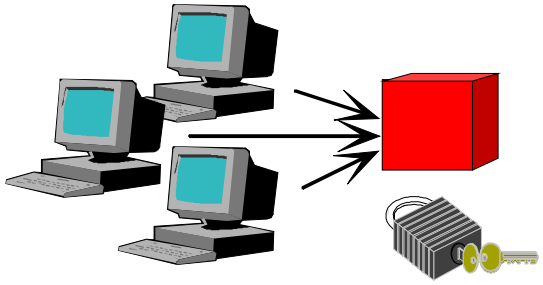


- Generische Dimensionalität
- Dynamisches Handling dünnbesetzter Datenmatrizen



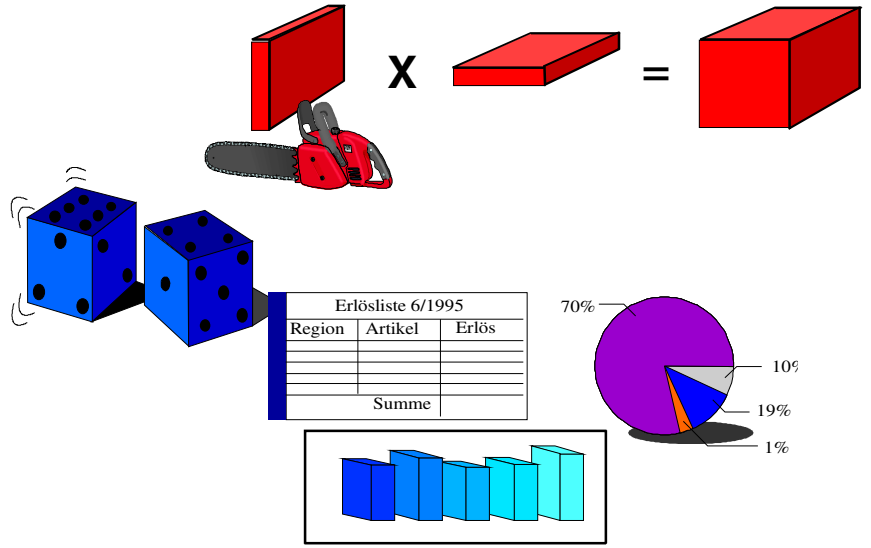
- Mehrbenutzerunterstützung

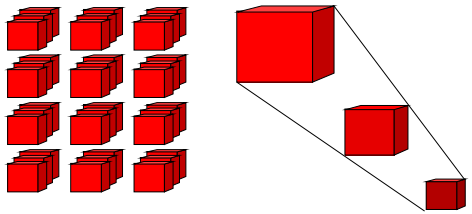




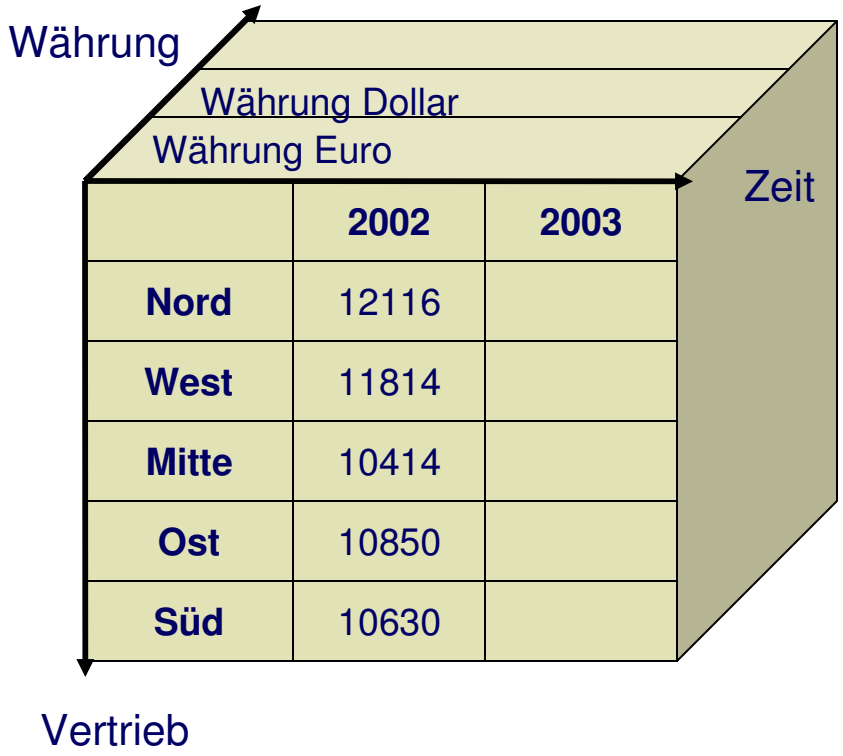
OLAP - Regeln III

- Unbeschränkte dimensionsübergreifende Operationen
- Intuitive Datenmanipulation
- Flexibles Berichtswesen
- Unbegrenzte Anzahl an Dimensionen und Aggregationsebenen

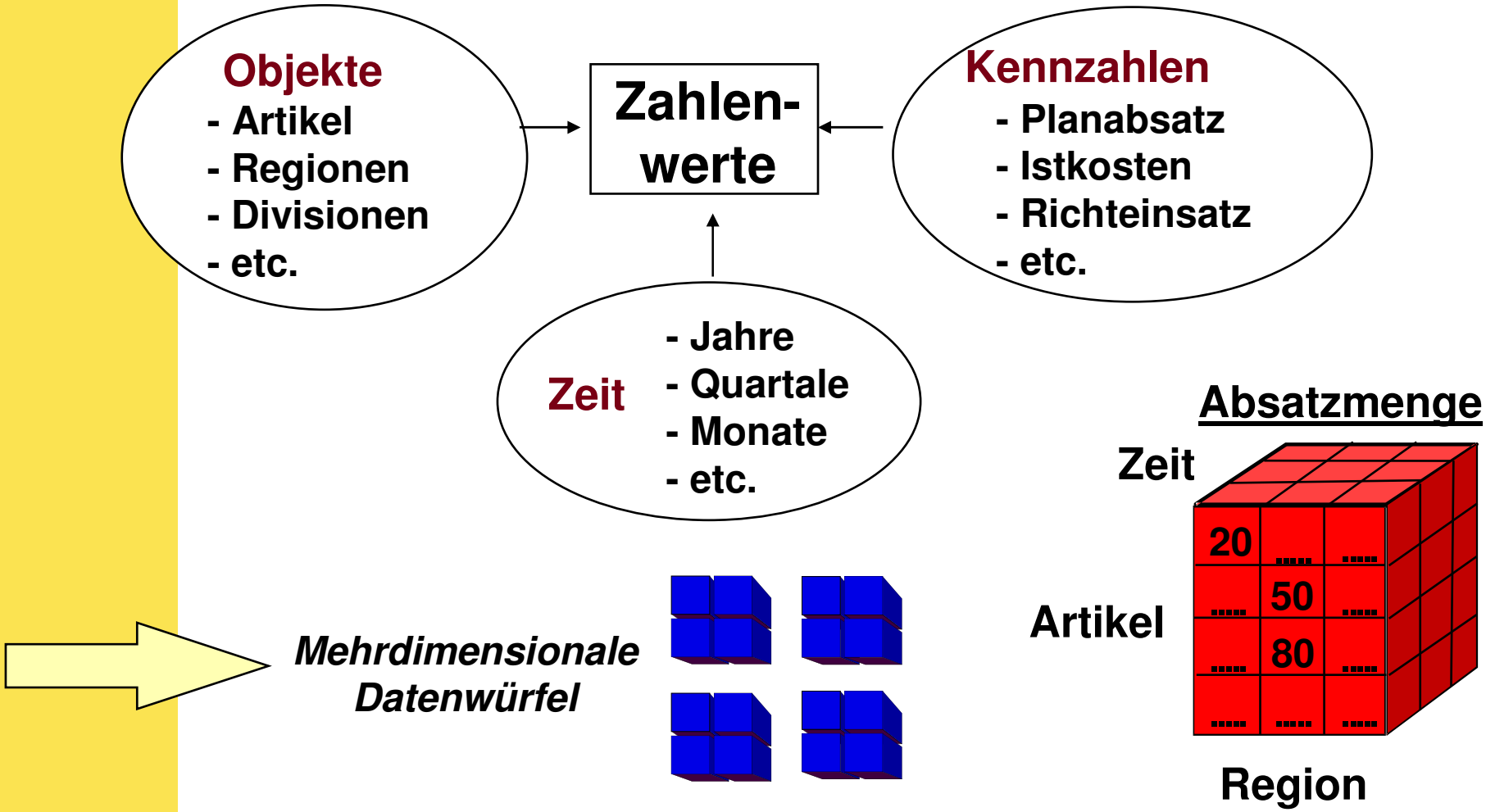




OLAP - Würfel


Währung
Währung Dollar
Währung Euro
Zeit
2002
2003
Nord
12116
West
11814
Mitte
10414
Ost
10850
Süd
10630
Vertrieb

*OLAP Multidimensionalität der Informationsversorgung



OLAP - Analysemöglichkeiten

Drill-Down / Roll-up

Detaillieren und Verdichten in einer Dimension

Slice

Fokussierung auf eine bestimmte Kategorie für die Analyse

Dice

Austauschen von Dimensionen in der Darstellung

Visualisierung

Auswahl der Darstellungsart (Matrizen, Balken, Zahlen, ...)

Drill-through

Detaillierung auf Einzelsatzebene

OLAP - Leistungsübersicht Tools

Abfragemöglichkeiten	Erstellung von Ad-hoc Abfragen ohne Programmierkenntnisse (Query by Example und Drill-Down)
Simulation	Analysen zur Trendberechnung und Planung (Was-wäre-wenn-Analysen)
Betriebswirtschaftliche Funktionen	Auf den Benutzer speziell abgestimmte betriebswirtschaftliche Analysen (z.B. Portfolioanalyse)
Automatisierung	Systeme, die ohne menschliche Hilfe Entscheidungsunterstützung bieten (elementare Diagnosen und Problemlösungen)
Präsentation	Darstellung der Information mit Hilfe von Berichten, Texten, Tabellen, Kreuztabellen und Grafiken

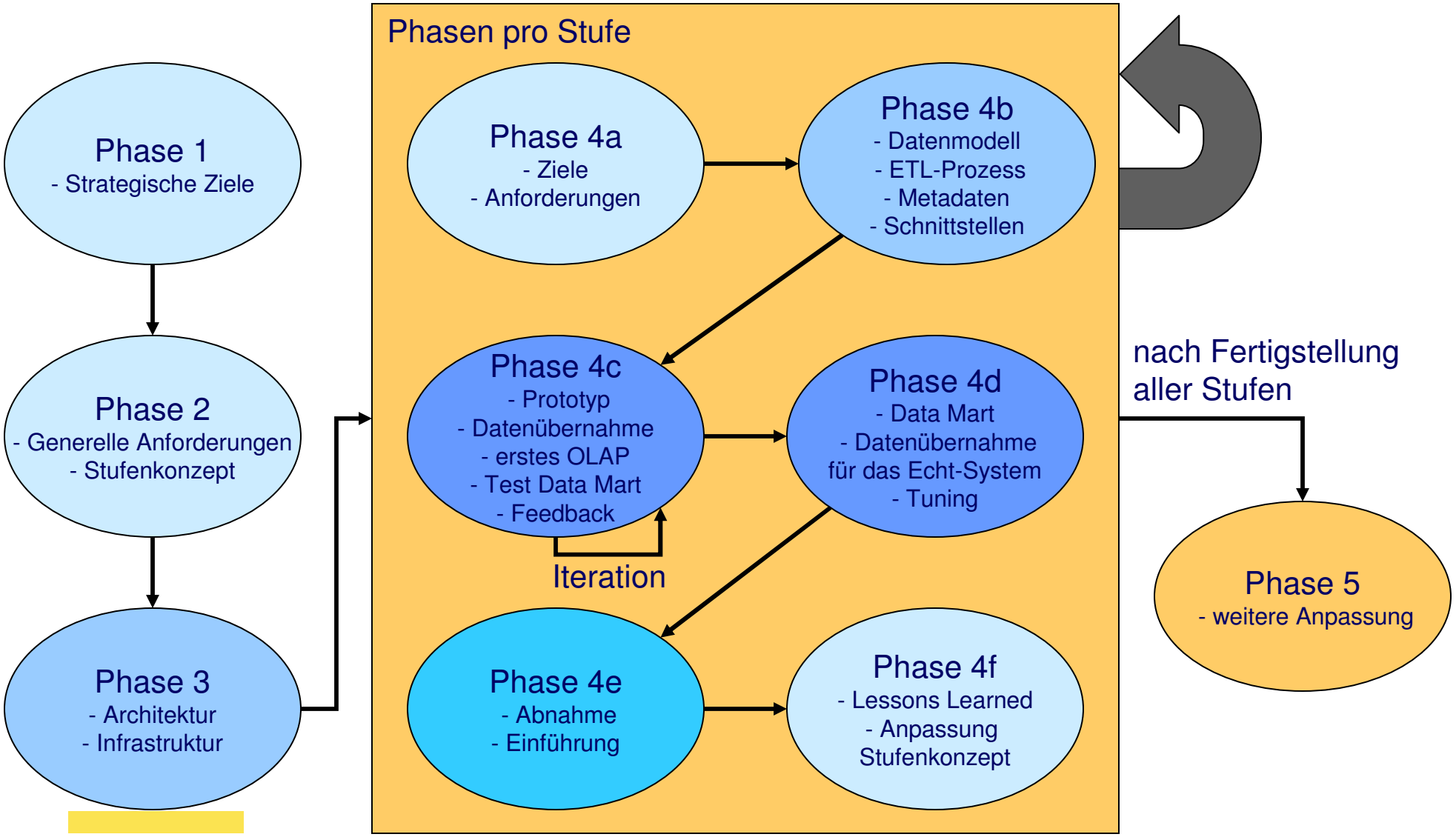
*Das Projektteam

- ProjektleiterIn mit den notwendigen Kompetenzen
- VertreterIn der jeweiligen Fachabteilungen mit genügend Engagement und Zeitkapazitäten
- VertreterIn aus dem IT/EDV Bereich
  - Mitarbeiter, die sich sehr gut mit den operativen Systemen auskennen
  - Mitarbeiter, deren hauptsächliche Aufgabe der Aufbau und die EDV-Betreuung des Neuen ist
- Pate / FürsprecherIn in der Geschäftsleitung

*Der Projektplan

- Realistisch planen
- Beanspruchung durch das Tagesgeschäft berücksichtigen
- Meilensteine in überschaubaren Abständen setzen
- Schrittweise vorgehen
- Stufenkonzept erstellen
- nach Möglichkeit die Projektstufen innerhalb einer kurzen Zeit (0,5-1 Jahr) durchziehen
- iterativ – Feedback von Nutzern einplanen

Projektphasen



Projektphase 1

- Erarbeiten der strategischen Ziele des DWH
- Ein realistisches Ziel/Aufgabenstellung als Challenge herausarbeiten; z.B. Reduzierung der Bereitstellungszeiten von Analysen um 50%.

Projektphase 2

- Klärung des Informationsbedarf und der wesentlichen Anforderungen aller späteren Anwender zur Bewältigung ihres Tagesgeschäftes und zur Umsetzung der strategischen Ziele durch umfangreiche Interviews mit den späteren Nutzern
- Analyse der bestehenden und zukünftigen Geschäftsprozesse des Unternehmens und der Datenbestände des operativen EDV-Systems und Analyse möglicher externer Daten
- Aus den Anforderungen unter Berücksichtigung der priorisierten Ziele wird ein detailliertes Stufenkonzept für den Aufbau des DWH entwickelt.

Projektphase 3

- Generelle Architektur
- Generelle Infrastruktur

Projektphase 4a

- Detailziele die aktuelle Stufe festlegen
- Anforderungsklärung für die aktuelle Stufe durchführen

Projektphase 4b

- Nur für die erste Ausbaustufe: Definition der Datenqualität und der Bewirtschaftungsprozesse für das DWH und die Data Marts; z.B. Update Zyklen, Dimensionen, Kennziffern, ...
- Nur für die erste Ausbaustufe: Corporate Data Model festlegen. Dieses Datenmodell muß ausbaubar, flexibel und den Geschäftsprozessen angepaßt sein
- Detail Architekturkonzept festlegen
- Toolauswahl
- Infrastruktur
- Erarbeiten und Implementierung des konkreten Datenmodells und Definition der Metadaten dieser Ausbaustufe I.
- Definition der Schnittstellen

Projektphase 4c

- Entwicklung eines Prototyp-DWH für diese Ausbaustufe
- Aufbereitung und Übernahme der Daten aus den operativen Systemen ins das Prototyp-DWH
- Testen des Prototyp-DWH durch die Fachabteilungen und gegebenenfalls Überarbeitung und Weiterentwicklung
- Test-Anbindung von Auswertungs- Tools (OLAP) und Test-Aufbau der ersten Data Marts
- Prototyping eines OLAP basierten Reportings
- Iterative Verbesserung des Prototyps zur ersten Aufbaustufe des DWH

Projektphase 4d

- Aufbereitung und Übernahme der Daten aus den operativen Systemen ins DWH für das Echt-System (Ausbaustufe I)
- Tuning des DWH für die jetzt bekannten hauptsächlichen Abfragen/Analysen
- Anbindung von Auswertungs- Tools (OLAP) und Aufbau der ersten Data Marts
- Aufbau eines OLAP basierten Reportings

Projektphase 4e

- Abnahme der Ausbaustufe
- Einführung

Projektphase 4f

- Lessons learned und Review
- ggf. Anpassung der nachfolgenden Ausbaustufen oder Veränderung des Stufenkonzepts
- Übergang zu 4a für die nächste Ausbaustufe

Projektphase 5

- Nach Inbetriebnahme der letzten Ausbaustufe muß konstant und konsequent Ausbau und Anpassung des DWH und der Data Marts an die sich entwickelnden Geschäftsprozesse und Aufgaben/Analysen erfolgen.

Nur ein System, das die aktuell anstehenden Entscheidungen unterstützt, wird akzeptiert und genutzt

Risiken auf dem Weg zum erfolgreichen DWH

- Ein Data Warehouse wird nur deshalb mit Informationen gefüllt, weil die Daten schon vorhanden und verfügbar sind, ohne dass sie für das Erreichen der strategischen Ziele relevant sind.
- Das Datenbankdesign eines Data Warehouse gleicht dem einer transaktionsorientierten Datenbank
- Dem Projektleiter/-team liegt eine technische Lösung näher als die Interessen und die Akzeptanz der Anwender
- Es werden nur die Analysen durchgeführt, die man schon immer gemacht hat
- Prozesse der Dateneingabe und Verarbeitung weichen von den analysierten Geschäftsprozessen ab oder Daten können außerhalb des eigentlichen Geschäftsprozesses sogar von Hand nachträglich manipuliert werden
- Zu starke Orientierung am Tool

Erfolgsfaktoren eines DWH-Projektes - I

- Detaillierte Analyse aller zuliefernden EDV-Systeme, Geschäftsprozesse und möglicher externer Daten
- Genaue Definition der Dimensionen und Kennzahlen und der Standards für die Datenqualität
- Klare Verteilung der Verantwortung für die Datenqualität sowohl innerhalb als auch bei den zuliefernden Systemen
- Enge Zusammenarbeit zwischen Projektteam, Geschäftsführung und Anwender
- Flexibles und erweiterbares Design des Data Warehouse,
- Unterstützung durch die Geschäftsführung, um nötige Ressourcen, Datensicherheit und -qualität durchzusetzen

Erfolgsfaktoren eines DWH-Projektes - II

- Ausführliche Auseinandersetzung mit den strategischen Zielen
  - Erst alle für das Unternehmen im Moment denkbaren Einsatzgebiete sammeln
  - Die gesammelten Einsatzgebiete bezüglich ROI betrachten und bewerten
  - Sich das Einsatzgebiet mit dem besten ROI, bzw. der größten Dringlichkeit für die Ausbaustufe I herausgreifen
- Neue Geschäftsprozesse/-felder werden zeitnah im DWH integriert
- Nur realistische Erwartungshaltungen an die Leistungsfähigkeit des DWH wecken, z.B. nicht jede gewünschte Information gibt es sofort und auf Knopfdruck

Der Erfolg eines DWH mißt sich vor allem an der Zufriedenheit seiner Benutzer

Kernfragen vor der Einführung des DWH

WARUM soll ein DWH eingeführt werden?

Für WEN soll ein DWH eingeführt werden?

WIE soll das DWH eingeführt werden?