Referenzmodell für die Architektur von Data-Warehouse-Systemen (Referenzarchitektur)

GSE-Working Group "Software Engineering“
10. November 2008, DKV, Köln
Thomas Zeh, E. Merck, Darmstadt

Referenzarchitektur

Gliederung

1. Ziele und Zweck der Präsentation
2. Entstehungsgeschichte
3. Definition der Begriffe Referenzmodell und Referenzarchitektur
4. Komponenten der Referenzarchitektur und ihr Zusammenspiel
5. Zu den vier Datenschichten der Architektur
6. Organisatorische Aspekte des Data Warehousing

   Diskussion

Referenzarchitektur

1. Ziele und Zweck der Präsentation

- Aufzeigen einer idealtypischen Data-Warehouse-Architektur (Referenzarchitektur)
- Überdenken des bisherigen Data-Warehouse-Begriffs
- Aufzeigen der erweiterten Einsatzmöglichkeiten

Referenzarchitektur

2. Entstehungsgeschichte

Die Referenzarchitektur ist wesentlicher Bestandteil des Buches

Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung
Herausgeber: A. Bauer und H. Günzel, Uni Erlangen

- Mitte 1999 als Idee des Arbeitskreises der GI „Konzepte des Data Warehousing“ geboren
- bis Mitte 2000 durch den Dialog/Diskurs von ca. 50 Autoren aus dem deutschsprachigen Raum entwickelt
- als Lehrbuch für Ausbildungszwecke wie auch als Nachschlagewerk für den Praktiker gedacht
- Näheres unter www.data-warehouse-systeme.de

Referenzarchitektur

3. Definitionen

3.1 Definition eines Referenzmodells (allgemein)

Definition: Ein Modell wird **Referenzmodell** (bezüglich eines Sachverhalts) genannt, wenn es die beiden folgende Eigenschaften aufweist:

1. Das Modell erlaubt Vergleiche zwischen Modellen, die den Sachverhalt beschreiben.

2. Auf Basis des Modells können spezielle Modelle (als Grundlage für die Konstruktion eines bestimmten Sachverhalts) geplant werden.

Das Referenzmodell stellt somit ein **Modellmuster** dar; es kann als ideal-typisches Modell für die Klasse der zu modellierender Sachverhalte (hier die Menge der existierenden und geplanten Data-Warehouse-Systeme) betrachtet werden. → Kandidat für einen Standard

Referenzarchitektur

3. Definitionen

3.2 Definition der Referenzarchitektur

Der Sachverhalt auf den das Referenzmodell angewandt wird, ist die Architektur von DW-Systemen. Unter Architektur eines Systems versteht der Autor (neben der Baukunst) all die die Struktur des Systems definierenden Komponenten und deren Anordnung (hier schwerpunktmäßig Daten und Funktionen – nicht jedoch Hardware, Organisation).

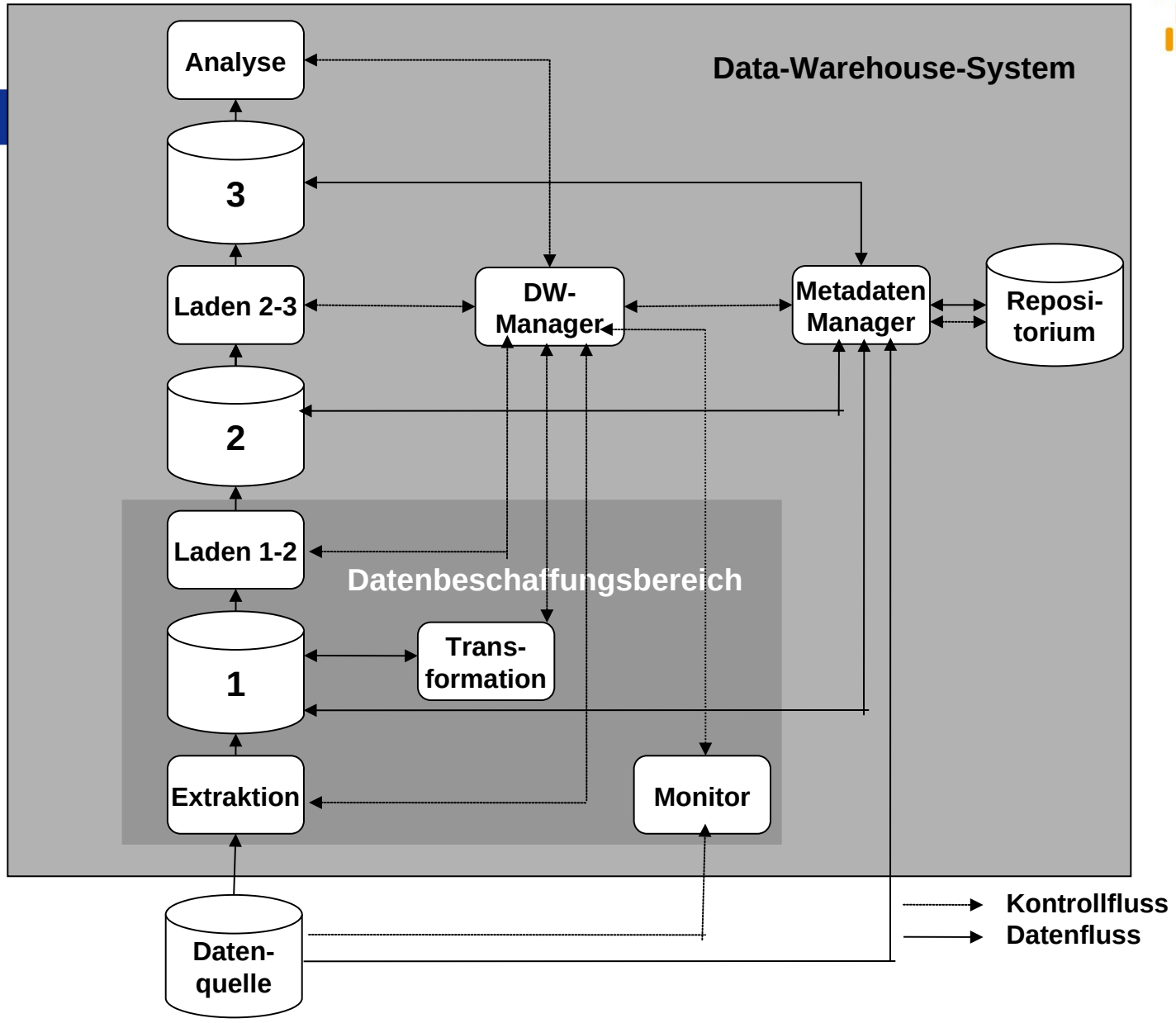
Architekturen sollten nach Vitruv auf folgenden Prinzipien beruhen:

- Das System soll die Anforderungen erfüllen (utilitas)
- Es soll robust gegenüber Änderungen sein (firmitas)
- Es soll ästhetisch / anmutig wirken (venustas).

Definition: Die **Referenzarchitektur** ist das Referenzmodell für die Architektur von Data-Warehouse-Systemen.

Referenzarchitektur

4.



4. Komponenten der Referenzarchitektur

4.1 Komponente Monitor

Aufgaben und Eigenschaften des Monitors:

- Er stellt das Bindeglied zwischen Datenquelle und DW-System dar.
- Er beobachtet die Datenquelle auf Änderungen.
- Er meldet Änderungen dem Data-Warehouse-Manager.

Praxisaspekte:

- Monitore sind oft im Umfeld des Anwendungssystems zu finden.
- Das Beobachten kann durch Auswertungen von Logfiles geschehen (z.B. sog. Belegdateien bei SAP).

4.2 Komponente Extraktion

Aufgaben und Eigenschaften der Extraktionskomponente:

- Sie selektiert die relevanten Daten einer Datenquelle.
- Sie überträgt die Daten aus einer Datenquelle in den Datentopf 1.

Praxisaspekte:
Übertragung der Daten erfolgt meist inkrementell (Delta-Technik) unter Nutzung von Standard-Datenbank-Schnittstellen wie z.B. ODBC.
Extraktion kann erfolgen

- periodisch,
- ereignisgesteuert (z.B. beim Erreichen von 1000 Änderungen),
- sofort bei Änderung in der Datenquelle,
- auf Anfrage.

4.3 Komponente Transformation

Aufgaben und Eigenschaften der Transformationskomponente:

„Homogenisierung“ der Daten (Anpassung von Daten und Schemata)

- Datenbereinigung (Plausibilitätsprüfung, Korrektur/Ergänzung,Streichung)
- Datenintegration (z.B. Schlüsselbehandlung, Anpassung / Vereinheitlichung von Datentypen (z.B. Date & Time), Code-Konvertierung, Voraggregation)
- Ladefähigkeit (Aufbereitung / Umstrukturierung für den Ladeprozeß)

Praxisaspekte:
Datenbereitstellung, -bereinigung und -integration erfordern häufig ca. 80 % des Budgets eines DWH-Projekts.

4.4 Komponente Datentopf 1

Der Datentopf 1 ist die zentrale Datenhaltungskomponente des Daten-beschaffungsbereiches.

Aufgaben und Eigenschaften des Datentopfes 1:

- Er ist Arbeitsspeicher für die Transformationskomponente.
- Er dient als (Zwischen-)Speicher der Daten auf ihrem Weg von der Datenquelle zum Langfristspeicher Datentopf 2.

Praxisaspekte:
Datentopf 1 ist ein Speicher für temporär zu haltende Daten.

4.5 Komponente Laden 1 → 2

Aufbereitete Daten müssen aus dem Datenbeschaffungsbereich in den "Nutzungsbereich" übertragen werden.

Aufgaben und Eigenschaften der Ladekomponente 1 → 2:

- Sie überträgt die Daten auf ihrem Weg vom Datentopf 1 zum Datentopf 2

Praxisaspekte:

- Nach dem erstmaligen Laden wird meist inkrementell geladen (Delta-Technik).

4.6 Komponente Datentopf 2

Aufgaben und Eigenschaften des Datentopfs 2:

- Er nimmt alle notwendigen Daten auf und stellt das (logisch) zentrale Datenlager über den ganzen relevanten Zeithorizont dar (Sammelfunktion).

- Er versorgt die Anwendungen mit den jeweiligen anwendungsspezifischen Daten (Verteilungsfunktion).

Praxisaspekte:
Da der Datentopf 2 alle Daten für jedwede Auswertung enthalten soll, ist er möglichst anwendungsneutral und redundanzfrei zu strukturieren.
der Devise "one fact - one place" zu strukturieren → 3NF.

4. Komponenten der Referenzarchitektur

4.7 Komponente Datentopf 3 und 4.8 Komponente Laden 2 → 3

Daten aus dem zentralen Datenlager werden anwendungsspezifisch aufbereitet und an einem separaten Ort bereitgestellt, dem Datentopf 3. Den Transport von Datentopf 2 nach 3 leistet die Ladekomponente 2 → 3.

Aufgaben und Eigenschaften des Datentopfes 3:

- Er ist die Datenbasis, die für die jeweilige Anwendung zugeschnitten ist, was Struktur und Inhalt (insbesondere den zeitlichen Horizont) angeht.

Praxisaspekte:

- Der Datentopf 3 tritt vielfach auf; je nach Anwendung bzw. Bereich.
- Die Strukturierung des Datentopfs 3 richtet sich nach den Auswertungs-wünschen; er ist nicht selten multidimensional strukturiert → OLAP.

4.9 Komponente Data-Warehouse-Manager

Der Data-Warehouse-Manager ist die zentrale Prozeßsteuerungs-Komponente des Data-Warehouse-Systems. Jeglicher Kontrollfluß strömt durch den Data-Warehouse-Manager. Sein “Gedächtnis” ist das Repositorium.

Aufgaben und Eigenschaften des Data-Warehouse-Managers:

- Initiierung des Datenbeschaffungsprozesses
- Steuerung und Überwachung der einzelnen Prozesse von der Extraktion bis hin zur Auswertung.

Praxisaspekte:

- Data-Warehouse-Manager hat Scheduler-/Workflow-Funktion.

4.10 Komponente Repositorium

Das Repositorium ist die Datenhaltungskomponente für alle **Metadaten**.
Es enthält jede Art von Information, die für den Aufbau und die Benutzung des Data-Warehouse-Systems und der Datenquellen benötigt werden

Aufgaben und Eigenschaften des Repositorium:

- Es dient der Integration (von Schemata und Daten).
- Es ermöglicht die Automatisierung des Ablaufs.
- Es ist die Ablage für datenschutz- und datensicherheitsrelevante Daten.

Praxisaspekte:

- Meist werden die Metadaten gestreut und teilweise redundant bei den jeweiligen Komponenten wie ETL-Tool, DBMS für die jeweiligen Datentöpfe und die Auswertungskomp. (insbes. Daten zum Zugriffsschutz) gehalten.

4.11 Komponente Metadatenmanager

Alle Änderungen an den Metadaten wie auch Anfragen an das Repositorium laufen zentral über den Metadatenmanager.

Aufgaben und Eigenschaften des Metadatenmanagers:

- Der Metadatenmanager verwaltet die Metadaten.

Praxisaspekte:

- Da die Metadaten heute meist verteilt bei den jeweiligen Komponenten gehalten werden, sind die zugehörigen Verwaltungsprogramme dann meist ebenfalls gestreut (z.B. beim ETL-Tool, bei Datentopf 2 und 3).

4.12 Komponente Analyse

Der Analysebegriff umfasst alle Operationen,
die mit den Daten im Datentopf 3 durchgeführt werden.

Aufgaben und Eigenschaften der Analysekomponente:

- Berechnen (von einfacher Arithmetik bis hin zu Data Mining)
- Suchen und Finden (Retrieval)
- Generierung von Sichten (z.B. für spezielle Datenbereitstellungen)
- Abrufen und Präsentieren / Visualisieren

Praxisaspekte:

- Analyse ist meist beschränkt auf Online Analytical Processing (OLAP); d.h. Auswerten von multidimensional strukturierten Daten.

Referenzarchitektur

5. Zu den vier Datenschichten der Architektur

5.1 Zu den Bezeichnungen der drei Datentöpfe:

Datentopf 1: → Arbeitsbereich, Staging Area (SA)

Datentopf 2: → Data Warehouse oder
Basisdatenbank (Bauer,Günzel) oder
Basis Data Warehouse (BDW) (Schwinn et.al.)

Datentopf 3: → Data Mart oder
Data Warehouse (Bauer,Günzel) oder
Funktionales Data Warehouse (FDW) (Schwinn et.al.)

Referenzarchitektur

5. Zu den vier Datenschichten der Architektur

5.2 Zur Zweckmäßigkeit der vier Datenschichten

1. Die Datenschicht Quelldaten (QD) ist gegeben.

2. Separate Staging Area (SA):
   Unabhängigkeit vom Betrieb der Quelldatensysteme
   wie auch vom BDW und der Analyse in den FDWs

3. Separates Basis Data Warehouse (BDW):
- Skalierbarkeit des BDW und Mehrfachverwendung der Daten durch anwendungsneutrale Strukturierung
- Stabile Datensituation (bis auf Zeiten während des Ladens)

4. Separates Funktionales Data Warehouse (FDW):
- Flexibilität durch anwendungsspezifische Strukturierung durch Unabhängigkeit von den Datenstrukturen des BDW (z.B. für performantes OLAP)
- Reproduktionsmöglichkeit der Auswertungen

5. Zu den vier Datenschichten der Architektur

5.3 Erweiterte Einsatzmöglichkeiten

Aus Sicht des Datenmanagements bieten Architekturen auf Basis der vorgestellten Referenzarchitektur weitergehende Möglichkeiten als das klassische Data Warehousing gemäß Definition von Inmon.

Wichtigstes Kriterium für ein (Basis) Data Warehouse:

- integrierte Sammlung von Daten („integrated collection of data“).

Fragwürdige Restriktionen des Data-Warehouse-Begriffs von Inmon sind:

- themenorientiert (subject oriented),
- schnappschussorientiert (time-variant),
- dauerhaft (non volatile),
- Managemententscheidungen unterstützend („in support of management´s decision-making process“).

Referenzarchitektur

5. Zu den vier Datenschichten der Architektur

5.4 Beispiele für erweiterte Einsatzmöglichkeiten

- **Referenzsystem für Dokumente**
  Um Anwendern eine integrierte Sicht über gestreut gehaltene Dokumente zu ermöglichen, wurde für die Suche ein Data Warehouse mit allen Werten zu den Suchattributen und den Verweisen auf die zugehörigen Dokumente beschickt. Für die Anzeige der gefundenen Dokumente wurde eine Zugriffsschicht auf die verteilten Datenhaltungssysteme geschaffen.

- **Metadirectory**
  Performante Datenbank mit standardisierter Zugriffsmethode LDAP
  z.B. für Daten von Mitarbeitern und Equipment weltweit als Data Warehouse

- **Produktdaten weltweit für Mitarbeiter aus F&E, Produktion, Lager, Vertrieb**
  Ein Automobilkonzern benutzt ein separates SAP R/3 als Data Warehouse ausschließlich für die Integration von Materialdaten (i.w. Konvertierung von Material-Identifikationen und Berücksichtigung lokal unterschiedlicher Stücklisten) aus den diversen lokalen Produktions- und Lagerstätten.

6. Organisatorische Aspekte des Data Warehousing (optional)

Data-Warehouse-Kompetenzzentrum: Zweckmäßige Institution im Vorfeld einer Organisationseinheit Data Warehousing wie auch nach der Etablierung

1. Aufgaben

   Anforderungs- und Machbarkeitsanalysen, Toolevaluation, Konzepterstellung für die Aufbauorganisation, Sicherstellung des Betriebs

2. Rollen / Zuständigkeiten

   Datenmanager: korrekter Datenfluss (inhaltlich)
   Ablaufadministrator: korrekter Datenfluss (technisch)
   Datenadministrator: konzeptionelles Schema
   Datenbankadministrator: internes Schema und User Views
   User Help Desk: Anwenderbetreuung
   und ggf. Qualitätsmanager

6. Organisatorische Aspekte des Data Warehousing (optional)

Rollen bei Merck (je Datenkomplex):

1. Data Owner

   Ein Manager (z.B. ein Bereichsleiter), der **Entscheidungen** über den Datenkomplex trifft.

2. Data Steward

   Ein „Kümmerer“, der für die im folgenden beschriebenen **Aufgaben** zum Datenkomplex **zuständig** ist.

6. Organisatorische Aspekte des Data Warehousing (optional)

Aufgaben des Data Steward

Die Verantwortung eines Data Steward umfasst nach M.Mayer und R.Winter drei Aspekte (2)

1. Dateninhalte und Verwendung

2. Entwicklung und Datenbereitstellung

3. Algorithmen und Methoden

Bei den im folgenden aufgeführten Aufgaben und Zuständigkeiten wird davon ausgegangen, dass im Regelfall die Arbeiten in einem Team innerhalb eines IT-Projektes durchgeführt werden. Hierbei kann der Data Steward Aufgaben delegieren, ebenso wie später im laufenden Betrieb.
Nichts desto trotz trägt er die Verantwortung für die aufgeführten Aufgaben.

6. Organisatorische Aspekte des Data Warehousing (optional)

Aufgaben des Data Steward

1. **Dateninhalte und Verwendung**
   Fachliche Seite der Daten im Hinblick auf Bedeutung, Qualität und Verwendungszweck.
   Fachlich korrekte Abbildung der Realität auf die Daten.
   Definition des Qualitätsanspruchs (z.B. Vollständigkeit, Datenintegrität, Aktualität)
   Hinweis: die Sicherstellung, dass der Anspruch auch erfüllt wird, obliegt anderen
   z.B. einem Prozessowner oder einem Verfahrensleiter

2. **Entwicklung und Datenbereitstellung**
   Festlegung der Zuständigkeiten für Erhebung, erstmalige Erfassung und laufende Pflege
   Definition des Schutz- und Sicherheitsanspruchs
   Archivierung und Wiederherstellung
   Auskunftsbereitschaft über Bedeutung, Zuständigkeiten,...

3. **Algorithmen und Methoden**
   Definition von Logiken und Extraktions- und Transformationsregeln zur Umwandlung von feingranularen Daten in die managementorientierten Sichten.
   Definition von abgeleiteten Daten (z.B. Aggregation).