Logische Modellierung von Data Warehouses

Vertiefungsarbeit von Karin Schäuble

Gliederung

1. Einführung

- Informationen werden zum entscheidenden Wettbewerbsfaktor
- Weltweite Vernetzung zwingt Unternehmen sich auch an der Vernetzung zu beteiligen
- Unternehmen haben durch ERP-Systeme erste Wege gefunden
- Fokus verschiebt sich Richtung Data Warehouse
- Aus brachliegenden Daten sollen wichtige Informationen herausgefiltert werden, um dann als Wissensgrundlage zu dienen
- Extraktion von verwertbarem Wissen wird immer wichtiger

1. Einführung

- DWH - Idee ist nicht ganz neu
- MIS wurde bereits Ende der 60er Jahre geprägt
- Früher mangelte es neben der technischen Infrastruktur an einer ausreichenden elektronischen Datenbasis
- Heute geht es vor allem darum das vorhandene Potential zu erschließen
- Das DWH der 90er zielt darauf ab, alle entscheidungsrelevanten Informationen verfügbar zu machen
- DWH frischt das Konzept der 70er Jahre wieder auf
- Unterschied : Daten liegen redundant vor
  es handelt sich nur um einen Teil der Informationen, der der dem jeweiligen Analysezweck dient

1. Einführung

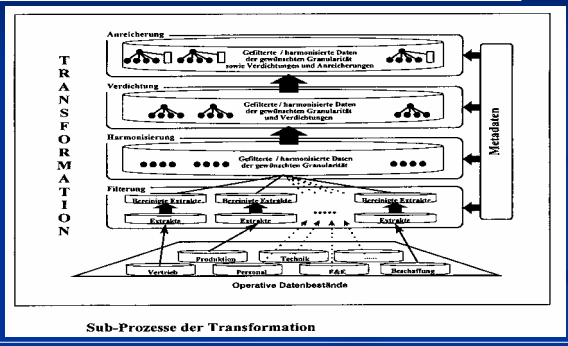
- Wichtiger Faktor bei der DWH-Modellierung ist die logische Modellierung
- DWH's setzen meist auf relationalen Datenbanken auf
- Alternativer Ansatz wird benötigt, der den Anforderungen der analytischen Systeme gerecht wird

2.1 Abgrenzung

Datenbankmodellierung vs. ETL-Prozeß:

- Datenbankmodellierung ist die Festlegung der Grundstruktur der Datenbank
- ETL: Modellierung der Daten

2.1 Abgrenzung



Sub-Prozesse der Transformation

2.1 Abgrenzung

Semantische vs. Logische vs. Physische Modellierung:

- Datenbankmodellierung ist von 2 Sichtweisen geprägt, die der Anwender und Entscheider und die der DWH-Entwickler
- 3 Modellierungsebenen
- Semantische Modellierung:
  - Dient als Schnittstelle zwischen den beiden Gruppen
  - Fachkonzept mit der Darstellung der betriebswirtschaftlichen Problemstellung
  - unabhängig von der Datenbanktechnologie
  - Begriffserklärung, Informationsbedarfsanalyse und Dokumentation

2.1 Abgrenzung

- Logische Modellierung:
  - setzt semantisches Modell in DV-Konzept um
  - noch unabhängig von der physischen Realisierung
  - ist auf die konkret eingerichtete Datenbanktechnologie ausgerichtet
- Physische Modellierung:
  - technische Implementierung
  - konkrete Umsetzung der logischen Datenmodellierung
  - abhängig von dem verwendeten Datenbanksystem

2.1 Abgrenzung

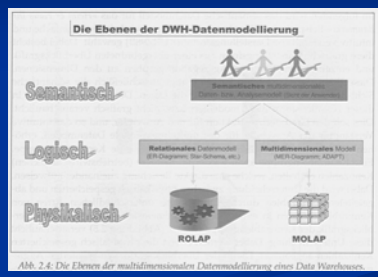


Abb. 2.4: Die Ebenen der multidimensionalen Datenmodellierung eines Data Warehouses.

2.1 Abgrenzung

OLTP vs. OLAP:

	OLTP	OLAP
Zweck	Unterstützung und Abwicklung der Geschäftsprozesse (dient der tägl. Arbeit) transaktionsorientiert	Informationssystem für Entscheidungsunterstützung (dient als Datenspeicher für Analyse) analyseorientiert
Datenbankgröße	Gigabyte-Bereich	Gigabyte/Terabyte-Bereich
Modellierung	Anwendungs-, funktions- und prozeßorientiert	subjektorientiert
Datenmenge je Transaktion	Gering	Groß
Datenaktualisierung	Permanent pro Transaktion	Periodisch
Abfragekomplexität	Niedrig	Hoch

2.1 Abgrenzung

OLTP vs. OLAP:

	OLTP	OLAP
Operationen	Lesen, schreiben, löschen	Im wesentl. nur lesen
Alter der Daten	Aktuelle (max. 90 Tage)	Mehr als 10 Jahre (historisch)
Datenquellen	Meist eine	mehrere
Niveau der Daten	Detailiert	Verdichtet, aufbereitet
Normalisierung	Wird i.a. eingehalten	Weniger wichtig
Zugriffsmuster	Vorhersehbar	Ad hoc
Behandlung der Zeit	Keinen Schlüssel	Dimension Zeit
Optimierungsziele	Höherer Durchsatz, sehr kurze Antwortzeiten, hohe Verfügbarkeit	Gute Antwortzeiten für komplexe Anfragen, hohe Flexibilität

2.2.1 Normalisierung

Gründe für die Normalisierung:

- Vermeidung von unerwünschten Anomalien
- Vermeidung von Redundanzen

Gründe für die Denormalisierung:

- Reduktion der Datenbankzugriffe

2.2.1 Normalisierung

1. NF:

Eine Relation befindet sich in der ersten Normalform, wenn alle Ihre Attribute nur einfache Attributwerte aufweisen.

2.2.1 Normalisierung

Unnormalisierte Relation

PersNr	Name	Vorname	AbtNr	Abteilung	ProjektNr	Beschreibung	Zeit
1	Lorenz	Sophia	1	Personal	2	Verkaufspromotion	83
2	Hohl	Tatjana	2	Einkauf	3	Konkurrenzanalyse	29
3	Willschrein	Theodor	1	Personal	1,2,3	Kundenumfrage, Verkaufspromotion, Konkurrenzanalyse	140, 92, 110
4	Richter	Hans-Otto	3	Verkauf	2	Verkaufspromotion	67
5	Wiesenland	Brunhilde	2	Einkauf	1	Kundenumfrage	160

2.2.1 Normalisierung

Relation Firma

PersNr	Name	Vorname	AbtNr	Abteilung	ProjektNr	Beschreibung	Zeit
1	Lorenz	Sophia	1	Personal	2	Verkaufspromotion	83
2	Hohl	Tatjana	2	Einkauf	3	Konkurrenzanalyse	29
3	Willschrein	Theodor	1	Personal	1	Kundenumfrage	140
3	Willschrein	Theodor	1	Personal	2	Verkaufspromotion	92
3	Willschrein	Theodor	1	Personal	3	Konkurrenzanalyse	110
4	Richter	Hans-Otto	3	Verkauf	2	Verkaufspromotion	67
5	Wiesenland	Brunhilde	2	Einkauf	1	Kundenumfrage	160

2.2.1 Normalisierung

2. NF:

Eine Relation befindet sich in der zweiten Normalform, wenn

1. sie in der ersten Normalform ist und
2. jedes Nicht-Schlüsselattribut vom gesamten Primärschlüssel voll funktional abhängig ist.

2.2.1 Normalisierung

Relation Firma

PersNr	Name	Vorname	AbtNr	Abteilung	ProjektNr	Beschreibung	Zeit
1	Lorenz	Sophia	1	Personal	2	Verkaufspromotion	83
2	Hohl	Tatjana	2	Einkauf	3	Konkurrenzanalyse	29
3	Willschrein	Theodor	1	Personal	1	Kundenumfrage	140
3	Willschrein	Theodor	1	Personal	2	Verkaufspromotion	92
3	Willschrein	Theodor	1	Personal	3	Konkurrenzanalyse	110
4	Richter	Hans-Otto	3	Verkauf	2	Verkaufspromotion	67
5	Wiesenland	Brunhilde	2	Einkauf	1	Kundenumfrage	160

2.2.1 Normalisierung

Relation Personal

PersNr	Name	Vorname	AbtNr	Abteilung
1	Lorenz	Sophia	1	Personal
2	Hohl	Tatjana	2	Einkauf
3	Willschrein	Theodor	1	Personal
4	Richter	Hans-Otto	3	Verkauf
5	Wiesenland	Brunhilde	2	Einkauf

Relation Firma

PersNr	ProjektNr	Zeit
1	2	83
2	3	29
3	1	140
3	2	92
3	3	110
4	2	67
5	1	160

Relation Projekt

ProjektNr	Beschreibung
2	Verkaufspromotion
3	Konkurrenzanalyse
1	Kundenumfrage

2.2.1 Normalisierung

3. NF:

Eine Relation befindet sich in der dritten Normalform, wenn

1. sie in der zweiten Normalform ist und
2. wenn keine transitiven Abhängigkeiten existieren.

2.2.1 Normalisierung

Relation Personal

PersNr	Name	Vorname	AbtNr	Abteilung
1	Lorenz	Sophia	1	Personal
2	Hohl	Tatjana	2	Einkauf
3	Willschrein	Theodor	1	Personal
4	Richter	Hans-Otto	3	Verkauf
5	Wiesenland	Brunhilde	2	Einkauf

Relation Firma

PersNr	ProjektNr	Zeit
1	2	83
2	3	29
3	1	140
3	2	92
3	3	110
4	2	67
5	1	160

Relation Projekt

ProjektNr	Beschreibung
2	Verkaufspromotion
3	Konkurrenzanalyse
1	Kundenumfrage

2.2.1 Normalisierung

Relation Personal

PersNr	Name	Vorname	AbtNr
1	Lorenz	Sophia	1
2	Hohl	Tatjana	2
3	Willschrein	Theodor	1
4	Richter	Hans-Otto	3
5	Wiesenland	Brunhilde	2

Relation Firma

PersNr	ProjektNr	Zeit
1	2	83
2	3	29
3	1	140
3	2	92
3	3	110
4	2	67
5	1	160

Relation Projekt

ProjektNr	Beschreibung
2	Verkaufspromotion
3	Konkurrenzanalyse
1	Kundenumfrage

Relation Abteilung

AbtNr	Abteilung
1	Personal
2	Einkauf
3	Verkauf

2.2.2 Aggregation

- Prozess bei dem Daten zusammengefasst werden
- Aggregationen können bei jeder Abfrage neu gebildet werden oder physisch in einer Tabelle abgespeichert werden

- **Vorteil:** Verbesserung der Antwortzeit einer Abfrage
- **Nachteil:** vergrößert den Umfang der Datenspeicherung und den Aufwand der Datenverwaltung

2.2.2 Aggregation

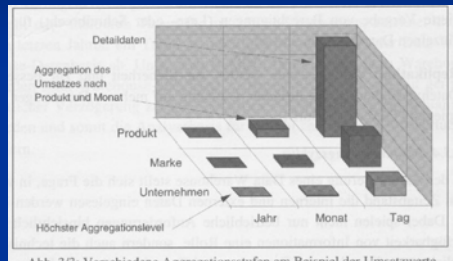


Abb. 3/3: Verschiedene Aggregationsstufen am Beispiel der Umsatzwerte

3. Anforderungen

- Auswertung und Analyse von Daten ist eine der Schlüsselkomponenten im DWH
- Greift auf OLAP-Technik zurück
- Codd stellte mehrere Regeln auf, die OLAP definieren
- Pendse und Creeth fassen die Anforderungen unter dem Akronym FASMI zusammen:

  Fast analysis of shared multidimensional information

3. Anforderungen

- Performance
- Flexibilität
- Datenvolumen
- Speicherplatz
- Verständlichkeit
- Wartung

4. Logische Modellierung

4.1 Methoden
- 4.1.1 Star Schema
- 4.1.2 Galaxy-Schema
- 4.1.3 Fact-Constellation-Schema
- 4.1.4 Snowflake-Schema

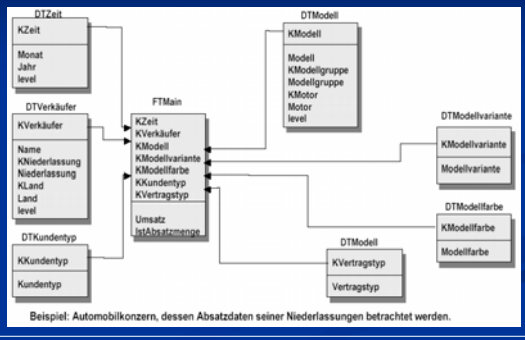
4.2 Software-Tools

4.1.1 Star-Schema

Eigenschaften:

- Daten werden in einer Fakttabelle und mehreren Dimensionstabellen organisiert, die sternförmig angeordnet sind
- Kennzahlen werden in der Fakttabelle und Dimensionselemente in den Dimensionstabellen abgelegt
- verbunden über Schlüsselelemente
- Daten liegen in den Dimensionstabellen denormalisiert vor
- Verdichtete Werte werden in den jeweiligen Tabellen abgelegt, gekennzeichnet durch ein Level-Attribut

4.1.1 Star-Schema



Beispiel: Automobilkonzern, dessen Absatzdaten seiner Niederlassungen betrachtet werden.

4.1.1 Star-Schema

KZeit	Monat	Jahr
1	Januar	1996
2	Februar	1996
3	März	1996
4	April	1996
5	Mai	1996
6	Juni	1996
7	Juli	1996
8	August	1996
9	September	1996
10	Oktober	1996
11	November	1996
12	Dezember	1996
13	Null	1996
14	Null	Null

Dimensionstabelle „Zeit"

4.1.1 Star-Schema

KVerkäufer	Name	KNieder-lassung	Nieder-lassung	KLand	Land	level
1	Meier	1	Düsseldorf	1	NRW	0
2	Schneider	1	Düsseldorf	1	NRW	0
3	Bäcker	2	Essen	1	NRW	0
4	Müller	2	Essen	1	NRW	0
5	Null	1	Düsseldorf	1	NRW	1
6	Null	2	Essen	1	NRW	1
7	Null	Null	Null	1	NRW	2
8	Null	Null	Null	Null	Null	3

Dimensionstabelle „Verkäufer“

4.1.1 Star-Schema

KModell	Modell	KModellgruppe	Modell-gruppe	KMotor	Motor	level
1	316i	1	3er	1	1,6 Liter	Modelle
2	318i	1	3er	2	1,8 Liter	Modelle
3	320i	1	3er	3	2,0 Liter	Modelle
12	Null	1	3er	Null	Null	Modellgruppen
13	Null	2	3er	Null	Null	Modellgruppen
14	Null	3	7er	Null	Null	Modellgruppen
15	Null	Null	Null	1	1,6 Liter	Motoren
16	Null	Null	Null	2	1,8 Liter	Motoren
17	Null	Null	Null	3	2,0 Liter	Motoren
23	Null	Null	Null	Null	Null	Total

Dimensionstabelle „Modelle“

4.1.1 Star-Schema

KModell-variante	Modellvariante
1	Stufenheck
2	Touring
3	Null

KModell-farbe	Modellfarbe
1	Rot
2	Schwarz
3	Blau
4	Null

KKundentyp	Kundentyp
1	Altkunde
2	Neukunde
3	Null

KModell-farbe	Modellfarbe
1	Barverkauf
2	Finazierung
3	Leasing
4	Null

Dimensionstabelle mit flacher Hierarchie

4.1.1 Star-Schema

KZeit	KVerkäufer	KModell	KModell-variante	KModell-farbe	KKundentyp	KVertragstyp	Umsatz	IstAbsatz-menge
1	1	2	1	1	2	2	60000	2
1	1	3	1	2	2	2	120000	3
1	3	3	1	1	1	2	80000	2
1	4	2	2	2	2	2	30000	1
1	4	3	2	2	2	1	40000	1
1	4	4	1	1	2	2	100000	2
2	1	4	1	3	1	1	50000	1
2	2	3	1	3	2	2	160000	4
2	2	2	2	2	2	2	90000	3
2	4	4	1	1	2	2	50000	1
....								
13	4	23	3	4	3	4	220000	5
....								
13	8	23	3	4	3	4	780000	20
....								

Faktentabelle im Sternschema

4.1.1 Star-Schema

Vorteile:

- Geringe Anzahl von Tabellen
- Wenige Joins
- Denormalisierte Dimensionstabellen

4.1.1 Star-Schema

Nachteile:

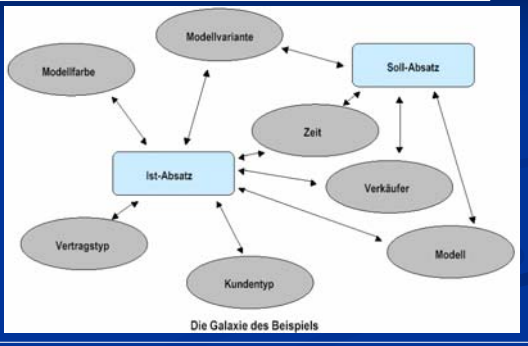
- Fakttabelle kann sehr groß werden
- Speicherung atomarer und aggregierter Werte in derselben Tabelle
- Level-Attribut muss in allen Abfragen mitgeführt werden
- Redundante Einträge
- Aktualisierung von Datensätzen

4.1.2 Galaxy-Schema

Eigenschaften:

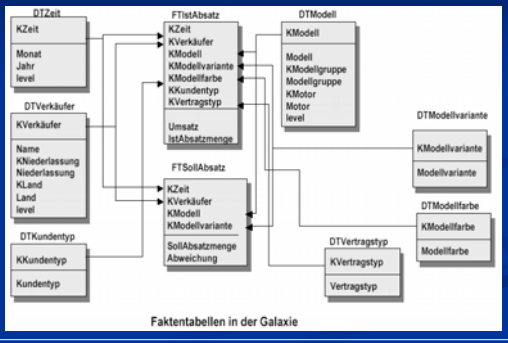
- Trennung in mehrere Fakttabellen, in welchen nur Fakten gleicher Dimensionierung gespeichert werden
- Dimensionen wie beim Star Schema

4.1.2 Galaxy-Schema



Die Galaxie des Beispiels

4.1.2 Galaxy-Schema



Faktentabellen in der Galaxie

4.1.2 Galaxy-Schema

Vorteile:

- Mehrere Fakttabellen erhöhen die Performance

Nachteile:

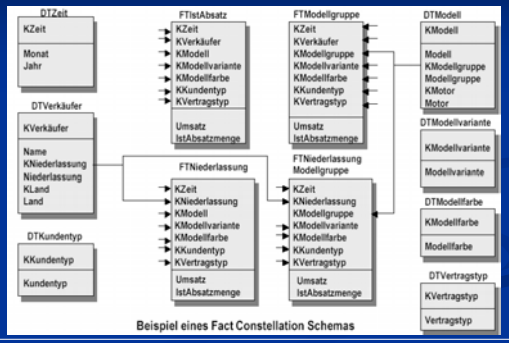
- Erschwert die Navigation im Datenbestand
- Eingeschränkte Abfragemöglichkeit
- Level-Einträge
- Speicherung atomarer und aggregierter Werte in derselben Tabelle

4.1.3 Fact-Constellation-Schema

Eigenschaften:

- Aggregierte Werte aus der Fakttabelle werden in separate Fakttabellen ausgelagert

4.1.3 Fact-Constellation-Schema



Beispiel eines Fact Constellation Schemas

4.1.3 Fact-Constellation-Schema

Vorteile:

- Kein Level-Attribut
- Kleinere Tabellen
- Schnellerer Zugriff auf aggregierte Werte

Nachteile:

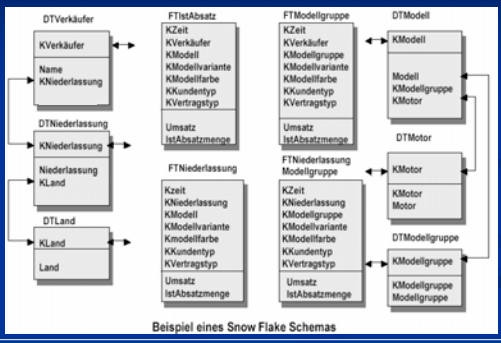
- Mehr Joins
- Mit der Anzahl der Dimensionstabellen steigt die Anzahl der Fakttabellen explosionsartig an

4.1.4 Snowflake-Schema

Eigenschaften:

- Normalisierung der Dimensionstabellen
- Fakttabellen ergeben sich wie beim Fact Constellation Schema

4.1.4 Snowflake-Schema



Beispiel eines Snow Flake Schemas

4.1.4 Snowflake-Schema

Vorteile:

- Geringe Redundanz
- Sehr flexibel

4.1.4 Snowflake-Schema

Nachteile:

- Hohe Komplexität, sehr unübersichtlich
- Viele Joins
- Hoher Wartungsaufwand

4.2 Software-Tools

Global Player:

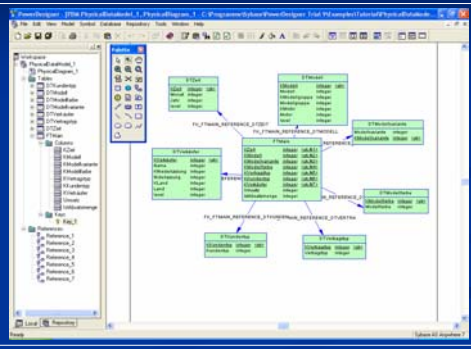
- Oracle
- Cognos
- Business Objects
- Crystal Decisions
- Sybase

4.2 Software-Tools

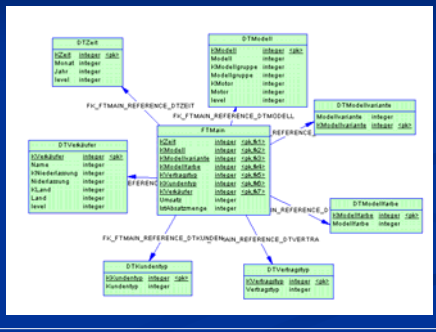
Sybase Power Designer:

- Business Process Architect: Modellierung von Unternehmensabläufen
- Data Architect: Datenbankmodellierung
- Physical Architect: Erstellung physikalischer Datenbankmodelle
- Developer: UML-Modellierung
- Object Architect: objektorientierte Modellierung
- Studio: kombiniert Prozess Modellierung mit UML-Modellierung (für Unternehmensverantwortliche)
- Viewer: bietet eine Ansicht aller Modellierungs-Informationen für alle

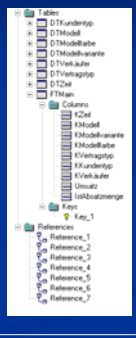
4.2 Software-Tools



4.2 Software-Tools



4.2 Software-Tools



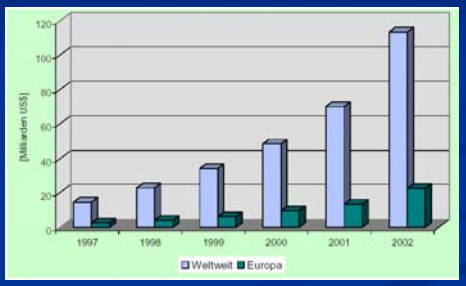
5. Zusammenfassung

- Es gibt kein Modell, dass eindeutig am besten ist
- Anforderungen des Unternehmen müssen bei der Auswahl der Modellierungswahl genau analysiert werden
  - welche Analysen sollen durchgeführt werden
  - welche Kennzahlen sollen dargestellt werden
  - wie viele und welche Dimensionen beschreiben die Kennzahlen
  - wie wichtig ist die Perfomance
  - wie groß soll das Data Warehouse werden
  - wie wichtig ist die Flexibilität

5. Ausblick

- Data Warehouse ist die ideale Basis für eine entscheidungsorientierte Auswertung der Unternehmensdaten
- OLAP und Data Mining erschließen verborgene Geschäftserfahrungen
- Data Mining ermöglicht das „Unternehmensgedächtnis" Data Warehouse systematisch, konsequent und erschöpfend zu nutzen
- in unmittelbarer Zukunft ist im Data Mining ein ähnlicher Investitionsboom zu erwarten
- Immer mehr Unternehmen versuchen Data Mining Software auf den Markt zu bringen
- Trotz aller bisherigen Entwicklungen steht das Data Mining noch am Anfang der Forschung

5. Ausblick



Vielen Dank für Ihre

Aufmerksamkeit!