

Inhalt

Geleitwort	XIII
Über dieses Buch	XV
Die Autoren	XVII
1 Einleitung	1
1.1 Ziele dieses Buches	2
1.2 Struktur dieses Buches	3
1.3 Hinweis zur Anwendung dieses Buches	4
2 Architektur	5
2.1 Data Warehouse-Architektur	5
2.1.1 Aufbau eines Data Warehouse	6
2.1.2 Transformationsschritte	9
2.1.3 Architekturgrundsätze	10
2.2 Architektur BI-Anwendungen	13
2.2.1 Die BI-Plattform zur Integration von Datenquellen	15
2.2.2 Die BI-Plattform zur Vereinheitlichung der Frontends	17
2.3 Datenhaltung	18
2.3.1 Grenzen gängiger DWH/BI-Technologien	19
2.3.2 Datenhaltung im Hadoop-Ecosystem	20
2.3.3 In-Memory-Datenbanken	23
3 Datenmodellierung	27
3.1 Vorgehensweise	27
3.1.1 Anforderungsgetriebene Modellierung	27
3.1.2 Quellsystemgetriebene Modellierung	29
3.1.3 Kombination der Ansätze	30
3.2 Relationale Modellierung	30
3.2.1 Darstellung von relationalen Datenmodellen	31
3.2.2 Normalisierung	31

3.2.3	Stammdaten und Bewegungsdaten	32
3.2.4	Historisierung	32
3.2.5	Relationales Core	34
3.2.6	Corporate Information Factory	35
3.2.7	Data Vault Modeling	35
3.3	Dimensionale Modellierung	37
3.3.1	Implementierung von dimensionalen Modellen	38
3.3.1.1	Relationale Implementierung	39
3.3.1.2	Multidimensionale Implementierung	40
3.3.2	Dimensionen	41
3.3.2.1	Fachliche Attribute	41
3.3.2.2	Technische Attribute	41
3.3.2.3	Hierarchien	42
3.3.2.4	Conformed Dimensions	43
3.3.2.5	Slowly Changing Dimensions	44
3.3.2.6	Zeitdimension	47
3.3.2.7	Bridge Tables	48
3.3.2.8	Spezielle Dimensionen	50
3.3.3	Fakten	51
3.3.3.1	Kennzahlen	51
3.3.3.2	Typen von Fakten	51
3.3.4	Modellierung spezieller Problemstellungen	53
3.3.4.1	Fakten unterschiedlicher Granularität und Rollen	53
3.3.4.2	Gemeinsame Hierarchiestufen in verschiedenen Dimensionen	54
3.3.4.3	Modellierungsgrundsätze für Dimensionen und Fakten	55
3.3.5	Darstellung von dimensionalen Modellen	56
3.3.5.1	ADAPT-Notation	56
3.3.5.2	Entity-Relationship-Diagramme	58
3.3.5.3	Data-Warehouse-Bus-Matrix	58
3.3.6	Dimensionales Core	59
3.4	Tools zur Datenmodellierung	60
3.4.1	Tools für relationale Datenmodellierung	60
3.4.2	Tools für dimensionale Datenmodellierung	61
4	Datenintegration	63
4.1	Data Profiling	64
4.1.1	Probleme mangelnder Datenqualität	64
4.1.2	Einsatz von Data Profiling	65
4.2	ETL	66
4.2.1	Aufgaben der ETL-Prozesse	67
4.2.1.1	Extraktion aus Quellsystemen	67
4.2.1.2	Transformationen	67
4.2.1.3	Laden in die Zieltabellen	68

4.2.2	ETL-Tools	68
4.2.2.1	Funktionalität von ETL-Tools	70
4.2.2.2	ETL oder ELT?	70
4.2.2.3	Positionierung von ETL-Tools	72
4.2.3	Performance-Aspekte	72
4.2.3.1	Mengenbasierte statt datensatzbasierte Verarbeitung	72
4.2.3.2	ELT-Tool statt ETL-Tool	73
4.2.3.3	Reduktion der Komplexität	74
4.2.3.4	Frühzeitige Mengeneinschränkung	75
4.2.3.5	Parallelisierung	76
4.2.4	Steuerung der ETL-Prozesse	78
4.2.4.1	Protokollierung des ETL-Ablaufs	78
4.2.4.2	Restartfähigkeit und Wiederaufsetzpunkte	79
4.3	Extraktion und Delta-Ermittlung	80
4.3.1	Delta-Extraktion im Quellsystem	81
4.3.1.1	Änderungsmarker und Journaltabellen	81
4.3.1.2	Delta-Ermittlung und Pending Commits	82
4.3.1.3	Change Data Capture	83
4.3.2	Voll-Extraktion und Delta-Abgleich im Data Warehouse	84
4.3.2.1	Zwei Versionen des Vollabzugs in der Staging Area	85
4.3.2.2	Vorteil einer Voll-Extraktion für die Delta-Ermittlung	87
4.3.3	Wann verwende ich was?	87
4.4	Fehlerbehandlung	88
4.4.1	Fehlende Attribute	89
4.4.1.1	Filtern von fehlerhaften Datensätzen	89
4.4.1.2	Fehlerhafte Datensätze in Fehlertabelle schreiben	89
4.4.1.3	Singletons auf Attributebene	90
4.4.2	Unbekannte Codewerte	90
4.4.2.1	Filtern von fehlerhaften Datensätzen	91
4.4.2.2	Singletons auf Datensatzebene	91
4.4.2.3	Generierung von Embryo-Einträgen	91
4.4.3	Fehlende Dimensionseinträge	92
4.4.3.1	Filtern von unvollständigen Fakten	93
4.4.3.2	Referenz auf Singleton-Einträge	94
4.4.3.3	Generieren von Embryo-Einträgen	95
4.4.4	Doppelte Datensätze	96
4.4.4.1	Verwendung von DISTINCT	97
4.4.4.2	Nur ersten Datensatz übernehmen	97
4.5	Qualitätschecks	97
4.5.1	Qualitätschecks vor und während des Ladens	98
4.5.2	Qualitätschecks nach dem Laden	99
4.5.3	Qualitätschecks mithilfe von Test-Tools	99
4.6	Real-Time BI	100
4.6.1	Begriffsbestimmung	101

4.6.2	Garantierte Verfügbarkeit von Informationen zu gegebenem Zeitpunkt	101
4.6.3	Verfügbarkeit von Informationen simultan zur Entstehung	102
4.6.4	Verfügbarkeit von Informationen kurz nach ihrer Entstehung	104
4.6.4.1	Events und Batchverarbeitung	105
4.6.4.2	Real-Time-Partitionen	106
4.6.5	Zusammenfassung	107
5	Design der DWH-Schichten	109
5.1	Staging Area	110
5.1.1	Gründe für eine Staging Area	111
5.1.2	Struktur der Stage-Tabellen	112
5.1.3	ETL-Logik für Stage-Tabellen	113
5.1.3.1	Einschränkungen bei der Extraktion	114
5.1.3.2	Transformation	114
5.1.3.3	Sonstige Informationen	115
5.2	Cleansing Area	115
5.2.1	Gründe für eine Cleansing Area	115
5.2.2	Struktur der Cleanse-Tabellen	116
5.2.3	Beziehungen in der Cleansing Area	118
5.2.4	ETL-Logik für Cleanse-Tabellen	120
5.2.4.1	Einschränkungen bei der Extraktion	121
5.2.4.2	Transformation	121
5.2.4.3	Sonstige Informationen	122
5.3	Core-Datenmodell allgemein	122
5.3.1	Aufgaben und Anforderungen an das Core	123
5.3.2	Stammdaten im Core	124
5.3.3	Bewegungsdaten im Core	124
5.3.4	Beziehungen im Core	124
5.3.5	Datenmodellierungsmethoden für das Core	125
5.4	Core-Datenmodell relational mit Kopf- und Versionstabellen	126
5.4.1	Historisierung von Stammdaten mit Kopf- und Versionstabellen	127
5.4.2	Struktur der Stammdatentabellen	128
5.4.2.1	Tabellenspalten und Schlüssel	129
5.4.2.2	Beziehungen (1:n) zwischen Stammdaten	132
5.4.2.3	Beziehungen (m:n) zwischen Stammdaten	133
5.4.3	ETL-Logik für Stammdatentabellen	135
5.4.3.1	Lookups (Schritt 1)	136
5.4.3.2	Outer Join (Schritt 2)	137
5.4.3.3	Neue Datensätze (Schritt 3)	141
5.4.3.4	Schließen einer Version/Fall 1 (Schritt 4)	142
5.4.3.5	Aktualisieren/Fall 2 (Schritt 5)	142
5.4.3.6	Versionieren/Fall 3 und 4 (Schritt 6)	142
5.4.3.7	Singletons	142

5.4.4	Typen von Bewegungsdaten	143
5.4.4.1	Transaction Tables	144
5.4.4.2	Snapshot Tables	144
5.4.4.3	Snapshot Tables versioniert	145
5.4.5	Struktur der Bewegungstabellen	146
5.4.5.1	Tabellenspalten und Schlüssel	147
5.4.5.2	Beziehungen zu Stammdaten	150
5.4.6	ETL-Logik für Bewegungstabellen	153
5.4.6.1	Lookups	154
5.4.6.2	Sonstige Informationen	155
5.4.7	Views für externen Core-Zugriff	155
5.4.7.1	Views für Stammdaten	156
5.4.7.2	Views für Bewegungsdaten	160
5.5	Core-Datenmodell relational mit Data Vault	161
5.5.1	Stammdaten	161
5.5.2	Beziehungen	162
5.5.3	Bewegungsdaten	162
5.5.4	Historisierung	163
5.5.5	Struktur der Tabellen	163
5.5.5.1	Hubtabellen – Tabellenspalten und Schlüssel	163
5.5.5.2	Satellitentabellen – Tabellenspalten und Schlüssel	164
5.5.5.3	Linktabellen – Tabellenspalten und Schlüssel	165
5.5.6	ETL-Logik	166
5.5.7	Views für externen Core-Zugriff auf das Data-Vault-Datenmodell	167
5.5.7.1	Views für Stammdaten (ein Satellite pro Hub bzw. Link)	167
5.5.7.2	Views für Stammdaten (mehrere Satellites pro Hub bzw. Link)	170
5.6	Core-Datenmodell dimensional	173
5.6.1	Star- oder Snowflake-Schema	174
5.6.1.1	Star-Schema	174
5.6.1.2	Snowflake-Schema	175
5.6.2	Historisierung von Stammdaten mit SCD	177
5.6.3	Struktur der Dimensionstabellen (Snowflake)	180
5.6.3.1	Tabellenspalten und Schlüssel	181
5.6.3.2	Beziehungen zwischen Hierarchiestufen	184
5.6.4	ETL-Logik für Dimensionstabellen (Snowflake)	185
5.6.4.1	Lookup	185
5.6.4.2	Weitere Schritte	186
5.6.5	Struktur der Faktentabellen (Snowflake)	186
5.6.6	ETL-Logik für Faktentabellen (Snowflake)	188
5.6.7	n:m-Beziehungen im dimensional Core	188
5.7	Marts	190
5.7.1	ROLAP oder MOLAP?	191
5.7.2	Historisierung von Data Marts	192
5.7.3	Star- oder Snowflake-Schema (ROLAP)	193

5.7.4	Struktur der Dimensionstabellen (Star)	194
5.7.4.1	Tabellenspalten und Schlüssel	194
5.7.4.2	Beispiel für Conformed Rollup	197
5.7.4.3	Beispiel für Dimension mit mehreren Hierarchien	198
5.7.5	ETL-Logik für Dimensionstabellen (Star)	199
5.7.5.1	Extraktion aus dem relationalen Core	200
5.7.5.2	Extraktion aus dem dimensionalen Core	207
5.7.6	Struktur der Faktentabellen (Star-Schema)	209
5.7.7	ETL-Logik für Faktentabellen (Star)	210
5.7.8	Multidimensionale Data Marts	210
5.7.8.1	Dimensionen (Cube)	211
5.7.8.2	Fakten (Cube)	212
6	Physisches Datenbankdesign	215
6.1	Indexierung	216
6.1.1	Staging Area	217
6.1.2	Cleansing Area	217
6.1.3	Core	217
6.1.4	Data Marts	218
6.2	Constraints	219
6.2.1	Primary Key Constraints	219
6.2.2	Foreign Key Constraints	220
6.2.3	Unique Constraints	221
6.2.4	Check Constraints	221
6.2.5	NOT NULL Constraints	222
6.3	Partitionierung	222
6.3.1	Grundprinzip von Partitionierung	223
6.3.2	Gründe für Partitionierung	223
6.3.3	Partitionierung in Staging und Cleansing Area	224
6.3.4	Partitionierung im Core	225
6.3.5	Partitionierung in den Data Marts	225
6.4	Datenkomprimierung	226
6.4.1	Redundanz	227
6.4.2	Wörterbuchmethode/Tokenbasierte Reduktion	227
6.4.3	Entropiekodierung	227
6.4.4	Deduplikation	228
6.4.5	Komprimierung bei spaltenorientierter Datenhaltung	228
6.5	Aggregationen	229
6.5.1	Vorberechnete Aggregationen	230
6.5.2	Query Rewrite	230
6.5.3	Einsatz im Data Warehouse	231

7	BI-Anwendungen	233
7.1	Überblick	233
7.2	Standardberichte	236
7.3	Ad-hoc-Analyse	238
7.4	BI-Portale	239
8	Betrieb	241
8.1	Release-Management	241
8.1.1	Kategorisierung der Anforderungen	242
8.1.2	Schnittstellen zu Quellsystemen	243
8.1.3	Umgang mit historischen Daten	245
8.1.4	Datenbankumgebungen	246
8.2	Deployment	248
8.2.1	Manuelles Deployment	248
8.2.2	Filebasiertes Deployment	249
8.2.3	Repository-basiertes Deployment	250
8.2.4	Kombiniertes Deployment	250
8.3	Monitoring	252
8.3.1	Betriebsmonitoring	252
8.3.2	System und DB-Monitoring	252
8.3.3	ETL-Monitoring	252
8.3.4	Performance-Monitoring	253
8.4	Migration	255
8.4.1	Datenbank	256
8.4.2	ETL-Tool	257
8.4.3	BI-Tools	258
	Literatur	259
	Index	261