

# ***Data Warehouse Grundlagen***

***Seminarunterlage***

***Version: 2.18***

Dieses Dokument wird durch die ORDIX AG veröffentlicht.

Copyright ORDIX AG. Alle Rechte vorbehalten.

Alle Produkt- und Dienstleistungs-Bezeichnungen sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Firmen und beziehen sich auf Eintragungen in den USA oder USA-Warenzeichen.

Weitere Logos und Produkt- oder Handelsnamen sind eingetragene Warenzeichen oder Warenzeichen der jeweiligen Unternehmen.

Kein Teil dieser Dokumentation darf ohne vorherige schriftliche Genehmigung der ORDIX AG weitergegeben oder benutzt werden.

### **Adressen der ORDIX AG**

Die ORDIX AG besitzt folgende Geschäftsstellen

ORDIX AG  
Karl-Schurz-Straße 19a  
D-33100 Paderborn  
Tel.: (+49) 0 52 51 / 10 63 - 0  
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG  
An der alten Ziegelei 5  
D-48157 Münster  
Tel.: (+49) 02 51 / 9 24 35 – 00  
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG  
Welser Straße 9  
D-86368 Gersthofen  
Tel.: (+49) 08 21 / 507 492 – 0  
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG  
Kreuzberger Ring 13  
D-65205 Wiesbaden  
Tel.: (+49) 06 11 / 7 78 40 – 00  
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG  
Wikingerstraße 18-20  
D-51107 Köln  
Tel.: (+49) 02 21 / 8 70 61 – 0  
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG  
Südwestpark 67/2  
D-90449 Nürnberg  
Tel.: (+49) 0 52 51 / 10 63 - 0  
Fax.: (+49) 01 80 / 1 67 34 90

Internet: <https://www.ordix.de>

Email: [seminare@ordix.de](mailto:seminare@ordix.de)

## Inhaltsverzeichnis

<b>1</b>	<b>Einführung .....</b>	<b>8</b>
1.1	Geschichtliches .....	9
1.2	Motivation .....	10
1.2.1	Hintergrund .....	11
1.3	Ist-Situation .....	12
1.4	Modernes Informationssystem .....	13
1.5	Was ist ein Data Warehouse? .....	14
1.5.1	Multiple Datenquellen .....	15
1.5.2	Unternehmensspezifisch skalierbar .....	16
1.5.3	Universelle Abfragen möglich .....	17
1.5.4	Hochleistungsplattform für Reporting .....	18
1.5.5	Analysen (Vergangenheit/Zukunft) .....	19
1.6	Ziele .....	20
1.7	Gründe für ein separates DWH .....	21
1.8	Abgrenzung zu OLTP .....	23
1.8.1	Abgrenzung zu OLTP-Anfragen .....	24
1.8.2	Abgrenzung zu OLTP-Daten .....	25
1.8.3	Abgrenzung zu OLTP-Anwender .....	26
1.8.4	Abgrenzung zu OLTP - Zusammenfassung .....	27
1.9	Definition .....	28
1.9.1	Definition nach Inmon .....	29
1.9.1.1	Themenorientierung .....	30
1.9.1.2	Integrierte Datenbasis .....	31
1.9.1.3	Persistente Datenbasis .....	32
1.9.1.4	Chronologisierte Daten .....	33
1.10	Anwendungsgebiete .....	34
<b>2</b>	<b>Multidimensionales Datenmodell .....</b>	<b>36</b>
2.1	Überblick .....	37
2.2	Normalisierung .....	38
2.2.1	Überblick .....	38
2.2.2	Normalisierung von Daten .....	39
2.2.2.1	Redundanzen .....	40
2.2.2.2	0. Normalform .....	41
2.2.2.3	1. Normalform .....	42
2.2.2.4	2. Normalform .....	44
2.2.2.5	3. Normalform .....	47
2.2.2.6	Zusammenfassung Normalformen .....	49
2.3	Kennzahlen .....	50
2.3.1	Additive Kennzahlen .....	51
2.3.2	Semi-Additive Kennzahlen .....	52
2.3.3	Nicht-Additive Kennzahlen .....	53
2.4	Dimensionen .....	54
2.4.1	Eigenschaften von Dimensionen .....	55
2.4.1.1	Einfache Hierarchie .....	56
2.4.1.2	Parallele Hierarchie .....	57
2.4.2	Aufbau Dimensionstabelle .....	58
2.4.2.1	Beispiel 1 .....	59
2.4.2.2	Beispiel 2 .....	60
2.4.3	Junk Dimension .....	61
2.4.3.1	Junk Dimension – Beispiel .....	62
2.5	Fakten- und Dimensionstabellen .....	63
2.5.1	Aufbau Faktentabelle .....	64
2.5.2	Varianten von Fakten .....	65
2.5.2.1	Ereignis Fakt .....	65
2.5.2.2	Snapshot Fakt .....	66
2.5.3	Beispiel Faktentabelle .....	67
2.6	DWH-Datenmodelle .....	68

2.6.1	Star-Schema .....	69
2.6.1.1	Vor- und Nachteile .....	70
2.6.1.2	Abfragen im Starschema .....	71
2.6.2	Snowflake-Schema .....	72
2.6.2.1	Schematische Darstellung .....	73
2.6.2.2	Vorteile .....	74
2.6.2.3	Nachteile .....	75
2.6.2.4	Abfragen im Snowflake-Schema .....	76
2.6.2.5	Mischformen .....	77
2.6.3	Galaxy-Schema – Überblick .....	78
2.7	Data Vault – Grundlagen .....	79
2.7.1	Data Vault – Elemente .....	80
2.7.2	Data Vault – Beispiel .....	81
2.7.3	Data Vault – Vorteile .....	82
2.7.4	Data Vault – Nachteile .....	83
2.8	Slowly Changing Dimensions .....	84
2.8.1	Slowly Changing Dimension – Anwendungsbereiche .....	85
2.8.2	Slowly Changing Dimensions – Fachlicher Schlüssel .....	86
2.8.3	Typ 1 – keine Historisierung .....	87
2.8.4	Typ 2 – Historisierung .....	88
2.8.4.1	Typ 2 – Ablaufprinzip .....	89
2.8.4.2	Typ 2 – Beispiel .....	90
2.8.5	Typ 3 – Teilweise Historisierung .....	92
2.9	Designprozess .....	93
2.9.1	Überblick .....	93
2.9.2	Beispiel .....	94
<b>3</b>	<b>Grundlagen der Architektur .....</b>	<b>96</b>
3.1	Überblick .....	97
3.1.1	Architekturschichten .....	97
3.1.2	Schematischer Aufbau .....	98
3.2	ETL-Tools .....	99
3.3	Staging Area .....	100
3.4	Landing Area .....	101
3.5	Cleansing Area .....	102
3.6	Metadaten .....	103
3.7	Core Data Warehouse .....	104
3.8	Data Marts .....	105
3.8.1	Überblick .....	105
3.8.2	Extraktarten .....	106
3.8.3	Vorteile .....	107
3.8.4	Data Mart Arten .....	108
3.8.4.1	Abhängige Data Marts .....	108
3.8.4.2	Unabhängige Data Marts .....	109
3.8.4.3	Virtuelle Data Marts .....	110
<b>4</b>	<b>Entwurf eines Data Warehouse Systems .....</b>	<b>111</b>
4.1	Vorgehensmodell .....	112
4.2	Modellierungsschritte .....	113
4.3	Analyse des Informationssystems / Anforderungsanalyse .....	114
4.3.1	Informationsbedarfsanalyse .....	115
4.3.2	Analysemodell / Prozessmodell .....	118
4.3.3	Objektmodell .....	120
4.4	Konzeptioneller Entwurf .....	121
4.4.1	Beispiel MERM Diagramm .....	122
4.5	Logischer Entwurf .....	123
4.6	Technische Implementierung .....	124
4.7	Test .....	125
4.8	Softwareauswahl .....	126
4.8.1	Auswahlkriterien .....	127

4.8.2	Marktrecherche .....	128
4.8.3	Bewertung der Auswahl .....	129
4.8.4	Kosten der Software .....	130
<b>5</b>	<b>Speicherstrukturen .....</b>	<b>131</b>
5.1	Überblick .....	132
5.2	ROLAP .....	133
5.3	MOLAP .....	134
5.4	HOLAP .....	136
5.5	DOLAP .....	137
5.6	Multidimensional Expressions (MDX) .....	138
<b>6</b>	<b>Befüllung .....</b>	<b>139</b>
6.1	Überblick .....	140
6.2	ETL-Tool .....	141
6.3	Monitoring Quellsystem .....	142
6.3.1	Überblick .....	142
6.3.2	Triggerbasiert .....	143
6.3.3	Replikationsbasiert .....	144
6.3.4	Zeitstempelbasierte Monitoringstrategie .....	145
6.3.5	Log-basierte Monitoringstrategie .....	146
6.3.6	Snapshot-basierte Monitoringstrategie .....	147
6.4	Extraktionsstrategien .....	148
6.5	Staging Area .....	150
6.5.1	Überblick .....	150
6.5.2	Ausprägungen .....	152
6.6	Cleansing Area .....	153
6.7	Transformation .....	154
6.7.1	Ursache von fehlerhaften Daten .....	154
6.7.2	Überblick .....	155
6.7.3	Filterung .....	156
6.7.3.1	Überblick .....	156
6.7.3.2	Klassen .....	157
6.7.3.3	Beispiele .....	158
6.7.4	Harmonisierung .....	159
6.7.5	Aggregation .....	162
6.7.6	Anreicherung .....	163
6.8	Laden .....	164
6.9	Data Mart .....	165
6.10	Change Data Capture .....	166
6.10.1	Überblick .....	166
6.10.2	Vorteile .....	167
6.11	Deploymentprozess .....	168
6.11.1	Deploymentprozess – Aufbau der Testdatenbank .....	168
6.11.2	Deploymentprozess – ETL Prozesse Entwickeln .....	169
6.11.3	Deploymentprozess – Testen .....	170
6.11.4	Deploymentprozess – ETL Prozesse entwickeln .....	171
<b>7</b>	<b>Multidimensionale Operatoren .....</b>	<b>172</b>
7.1	OLAP-Operatoren .....	173
7.1.1	Standard-Operatoren .....	174
7.1.2	Bewegen im Multidimensionalen Datenmodell .....	175
7.1.3	Pivotierung/Rotation .....	176
7.1.4	Roll-Up / Drill-Down .....	177
7.1.4.1	Beispiel .....	178
7.1.5	Drill Across .....	179
7.1.6	Drill Through .....	180
7.1.7	Slice / Dice .....	181
7.1.8	Slice .....	182
7.1.9	Dice .....	183

7.1.10	Split / Merge .....	184
7.1.10.1	Beispiel .....	185
<b>8</b>	<b>Reporting.....</b>	<b>186</b>
8.1	Frontend Tools .....	187
8.2	Verteilung der Anwender.....	188
8.3	Dashboard.....	189
8.4	Statische Reports.....	190
8.5	Dynamische Berichte .....	191
8.6	Ad-Hoc Berichte .....	192
8.7	Data Mining .....	193
8.8	Auswahl Reporting Tool.....	194
<b>9</b>	<b>Datenbankoptimierung .....</b>	<b>196</b>
9.1	Überblick .....	197
9.2	Laden von Daten.....	198
9.2.1	Überblick .....	198
9.2.2	Einzelatzverarbeitung .....	199
9.2.3	Ladetool.....	200
9.2.4	Ladetool – Beispiel Oracle .....	201
9.2.5	Ladetool – Beispiel DB2.....	202
9.2.6	Ladetool – Beispiel Informix .....	203
9.3	Partitionierung .....	204
9.3.1	Überblick .....	204
9.3.2	Range Partitioning.....	205
9.3.3	List Partitioning.....	209
9.3.4	Hash Partitioning.....	212
9.4	Komprimierung.....	216
9.5	Datenbank Caches.....	219
9.6	Blockgröße – Seitengröße .....	220
9.7	Reservierter Freiplatz bei der Tabellenerstellung .....	221
9.8	Referenzielle Integrität .....	222
9.9	Materialisierte Sichten und Tabellen.....	223
9.10	Merge Anweisung .....	226
9.10.1	Beispiel.....	227
9.11	Parallelisierung.....	228
9.12	Spaltenorientierte Speicherung und In-Memory-Funktionalität.....	229
9.12.1	Überblick .....	229
9.12.2	Beispiel Oracle .....	230
9.12.3	Beispiel DB2.....	233
9.13	Hardwareoptimierungen.....	234
<b>10</b>	<b>Big Data .....</b>	<b>235</b>
10.1	Wie "big" ist Big Data? .....	236
10.2	3V-Modell .....	237
10.3	Welche technischen Probleme sollen gelöst werden? .....	238
10.4	Welche fachlichen Probleme sollen gelöst werden? .....	239
10.5	Big Data in Aktion.....	240
10.5.1	Beispiel Handel: Onlineshop .....	240
10.5.2	Mögliche Interpretationen.....	241
10.6	Verteilte Datenhaltung im Cluster .....	242
10.7	Typische Probleme in verteilten Systemen.....	243
10.8	Verfügbarkeit im Cluster.....	244
10.9	Horizontale Skalierung mit Commodity Hardware .....	245
10.10	PC Cluster .....	246
10.11	Yahoo's Hadoop Cluster (2007).....	247
10.12	Big Data ist nicht nur ein Tool .....	248
10.13	Zusammenfassung.....	249
<b>11</b>	<b>Streaming.....</b>	<b>250</b>

11.1	Anwendungsfall.....	251
11.2	Apache Kafka.....	252
11.3	Begriffe.....	253
11.4	Streaming – Analogie.....	254
11.5	Klassische Datenverarbeitung.....	255
	11.5.1 Nachteile.....	256
11.6	Von Batch zu Streaming.....	257
11.7	Vorteile von Streaming.....	258
11.8	Plattform – Analogie.....	259
11.9	Klassische Datenübertragung.....	260
	11.9.1 Nachteile.....	261
11.10	Publish-Subscribe.....	262
	11.10.1 Vorteile.....	263
11.11	Distributed – Analogie.....	264
11.12	Distribution.....	265
11.13	Gesamtbild.....	266
<b>12</b>	<b>Übungen / Lösungen.....</b>	<b>267</b>
12.1	Übungen.....	268
	12.1.1 Normalisierung von Daten.....	268
	12.1.2 Multidimensionales Datenmodell.....	271
	12.1.3 Snowflake-Schema.....	273
	12.1.4 Befüllung.....	274
12.2	Lösungen.....	276
	12.2.1 Normalisierung von Daten.....	276
	12.2.2 Multidimensionales Datenmodell.....	280
	12.2.3 Snowflake-Schema.....	282
	12.2.4 Befüllung.....	283