



Building the Data Warehouse, Fourth Edition

W. H. Inmon

ivj/

Wile\ i¹

Preface	xix	
Acknowledgments	xxvii	
Chapter 1	Evolution of Decision Support Systems	1
	The Evolution	2
	The Advent of DASD	4
	PC/4GL Technology	4
	Enter the Extract Program	5
	The Spider Web	6
	Problems with the Naturally Evolving Architecture	7
	Lack of Data Credibility	7
	Problems with Productivity	9
	From Data to Information	12
	A Change in Approach	14
	The Architected Environment	16
	Data Integration in the Architected Environment	18
	Who Is the User?	20
	The Development Life Cycle	20
	Patterns of Hardware Utilization	22
	Setting the Stage for Re-engineering	23
	Monitoring the Data Warehouse Environment	25
	Summary	28
Chapter 2	The Data Warehouse Environment	29
	The Structure of the Data Warehouse	33
	Subject Orientation	34
	Day 1 to Day <i>n</i> Phenomenon	39

	Granularity	41
	The Benefits of Granularity	42
	An Example of Granularity	43
	Dual Levels of Granularity	46
	Exploration and Data Mining	50
	Living Sample Database	50
	Partitioning as a Design Approach	53
	Partitioning of Data	53
	Structuring Data in the Data Warehouse	56
	Auditing and the Data Warehouse	61
	Data Homogeneity and Heterogeneity	61
	Purging Warehouse Data	64
	Reporting and the Architected Environment	64
	The Operational Window of Opportunity	65
	Incorrect Data in the Data Warehouse	67
	Summary	69
Chapter 3	The Data Warehouse and Design	71
	Beginning with Operational Data	71
	Process and Data Models and the Architected Environment	78
	The Data Warehouse and Data Models	79
	The Data Warehouse Data Model	81
	The Midlevel Data Model	84
	The Physical Data Model	88
	The Data Model and Iterative Development	91
	Normalization and Denormalization	94
	Snapshots in the Data Warehouse	100
	Metadata	102
	Managing Reference Tables in a Data Warehouse	103
	Cyclicity of Data — The Wrinkle of Time	105
	Complexity of Transformation and Integration	108
	Triggering the Data Warehouse Record	112
	Events	112
	Components of the Snapshot	113
	Some Examples	113
	Profile Records	114
	Managing Volume	115
	Creating Multiple Profile Records	117
	Going from the Data Warehouse to the	
	Operational Environment	117
	Direct Operational Access of Data Warehouse Data	118
	Indirect Access of Data Warehouse Data	119
	An Airline Commission Calculation System	119
	A Retail Personalization System	121
	Credit Scoring	123
	Indirect Use of Data Warehouse Data	125

	Star Joins	126
	Supporting the ODS	133
	Requirements and the Zachman Framework	134
	Summary	136
Chapter 4	Granularity in the Data Warehouse	139
	Raw Estimates	140
	Input to the Planning Process	141
	Data in Overflow	142
	Overflow Storage	144
	What the Levels of Granularity Will Be	147
	Some Feedback Loop Techniques	148
	Levels of Granularity — Banking Environment	150
	Feeding the Data Marts	157
	Summary	157
Chapter 5	The Data Warehouse and Technology	159
	Managing Large Amounts of Data	159
	Managing Multiple Media	161
	Indexing and Monitoring Data	162
	Interfaces to Many Technologies	162
	Programmer or Designer Control of Data Placement	163
	Parallel Storage and Management of Data	/ 164
	Metadata Management	165
	Language Interface	166
	Efficient Loading of Data	166
	Efficient Index Utilization	168
	Compaction of Data	169
	Compound Keys	169
	Variable-Length Data	169
	Lock Management	171
	Index-Only Processing	171
	Fast Restore	171
	Other Technological Features	172
	DBMS Types and the Data Warehouse	172
	Changing DBMS Technology	174
	Multidimensional DBMS and the Data Warehouse	175
	Data Warehousing across Multiple Storage Media	182
	The Role of Metadata in the JPata Warehouse Environment	182
	Context and Content	185
	Three Types of Contextual Information	186
	Capturing and Managing Contextual Information	187
	Looking at the Past	187
	Refreshing the Data Warehouse	188
	Testing	190
	Summary	191

Chapter 6	The Distributed Data Warehouse	193
	Types of Distributed Data Warehouses	193
	Local and Global Data Warehouses	194
	The Local Data Warehouse	197
	The Global Data Warehouse	198
	Intersection of Global and Local Data	201
	Redundancy	206
	Access of Local and Global Data	207
	The Technologically Distributed Data Warehouse	211
	The Independently Evolving Distributed Data Warehouse	213
	The Nature of the Development Efforts	213
	Completely Unrelated Warehouses	215
	Distributed Data Warehouse Development	217
	Coordinating Development across Distributed Locations	218
	The Corporate Data Model — Distributed	219
	Metadata in the Distributed Warehouse	223
	Building the Warehouse on Multiple Levels	223
	Multiple Groups Building the Current Level of Detail	226
	Different Requirements at Different Levels	228
	Other Types of Detailed Data	232
	Metadata	234
	Multiple Platforms for Common Detail Data	235
	Summary	236
Chapter 7	Executive Information Systems and the Data Warehouse	239
	EIS— The Promise	240
	A Simple Example	240
	Drill-Down Analysis	243
	Supporting the Drill-Down Process	245
	The Data Warehouse as a Basis for EIS	247
	Where to Turn	248
	Event Mapping	251
	Detailed Data and EIS	253
	Keeping Only Summary Data in the EIS	254
	Summary	255
Chapter 8	External Data and the Data Warehouse	257
	External Data in the Data Warehouse	260
	Metadata and External Data	261
	Storing External Data	263
	Different Components of External Data	264
	Modeling and External Data	265
	Secondary Reports	266
	Archiving External Data	267
	Comparing Internal Data to External Data	267
	Summary	268

Chapter 9	Migration to the Architected Environment	269
	A Migration Plan	270
	The Feedback Loop	278
	Strategic Considerations	280
	Methodology and Migration	283
	A Data-Driven Development Methodology	283
	Data-Driven Methodology	286
	System Development Life Cycles	286
	A Philosophical Observation	286
	Summary	287
Chapter 10	The Data Warehouse and the Web	289
	Supporting the eBusiness Environment	299
	Moving Data from the Web to the Data Warehouse	300
	Moving Data from the Data Warehouse to the Web	301
	Web Support	302
	Summary	302
Chapter 11	Unstructured Data and the Data Warehouse	305
	Integrating the Two Worlds	307
	Text — The Common Link	308
	A Fundamental Mismatch	310
	Matching Text across the Environments	310
	A Probabilistic Match	311
	Matching All the Information	312
	A Themed Match	313
	Industrially Recognized Themes	313
	Naturally Occurring Themes	* 316
	Linkage through Themes and Themed Words	317
	Linkage through Abstraction and Metadata	318
	A Two-Tiered Data Warehouse	320
	Dividing the Unstructured Data Warehouse	321
	Documents in the Unstructured Data Warehouse	322
	Visualizing Unstructured Data	323
	A Self-Organizing Map (SOM)	324
	The Unstructured Data Warehouse	325
	Volumes of Data and the Unstructured Data Warehouse	326
	Fitting the Two Environments Together	327
	Summary	330
Chapter 12	The Really Large Data Warehouse	331
	Why the Rapid Growth?	332
	The Impact of Large Volumes of Data	333
	Basic Data-Management Activities	334
	The Cost of Storage	335
	The Real Costs of Storage	336
	The Usage Pattern of Data in the Face of Large Volumes	f 336

; ij
 v
 a;
 !,!
 ;
 ;

A Simple Calculation	337
Two Classes of Data	338
Implications of Separating Data into Two Classes	339
Disk Storage in the Face of Data Separation	340
Near-Line Storage	341
Access Speed and Disk Storage	342
Archival Storage	343
Implications of Transparency	345
Moving Data from One Environment to Another	346
The CMSM Approach	347
A Data Warehouse Usage Monitor	348
The Extension of the Data Warehouse across Different Storage Media	349
Inverting the Data Warehouse	350
Total Cost	351
Maximum Capacity	352
Summary	354
Chapter 13 The Relational and the Multidimensional Models as a Basis for Database Design	357
The Relational Model	357
The Multidimensional Model	360
Snowflake Structures	361
Differences between the Models	362
The Roots of the Differences	363
Reshaping Relational Data	364
Indirect Access and Direct Access of Data	365
Servicing Future Unknown Needs	366
Servicing the Need to Change Gracefully	367
Independent Data Marts	370
Building Independent Data Marts	371
Summary	375
Chapter 14 Data Warehouse Advanced Topics	377
End-User Requirements and the Data Warehouse	377
The Data Warehouse and the Data Model	378
The Relational Foundation	378
The Data Warehouse and Statistical Processing	379
Resource Contention in the Data Warehouse	380
The Exploration Warehouse	380
The Data Mining Warehouse	382
Freezing the Exploration Warehouse	383
External Data and the Exploration Warehouse	384
Data Marts and Data Warehouses in the Same Processor	384
The-Life Cycle of Data	386
Mapping the Life Cycle to the Data Warehouse Environment	387
Testing and the Data Warehouse	388

Tracing the Flow of Data through the Data Warehouse	390
Data Velocity in the Data Warehouse	391
"Pushing" and "Pulling" Data	393
Data Warehouse and the Web-Based eBusiness Environment	393
The Interface between the Two Environments	394
The Granularity Manager	394
Profile Records	396
The ODS, Profile Records, and Performance	397
The Financial Data Warehouse	397
The System of Record	399
A Brief History of Architecture — Evolving to the Corporate Information Factory	402
Evolving from the CIF	404
Obstacles	406
CIF — Into the Future	406
Analytics	406
ERP/SAP	407
Unstructured Data	408
Volumes of Data	409
Summary	410
Chapter 15 Cost-Justification and Return on Investment for a Data Warehouse	413
Copying the Competition	413
The Macro Level of Cost-Justification	414
A Micro Level Cost-Justification	415
Information from the Legacy Environment	418
The Cost of New Information	419
Gathering Information with a Data Warehouse	419
Comparing the Costs	420
Building the Data Warehouse	420
A Complete Picture	421
Information Frustration	422
The Time Value of Data	422
The Speed of Information	423
Integrated Information	424
The Value of Historical Data	425
Historical Data and CRM	426
Summary	426
Chapter 16 The Data Warehouse and the ODS	429
Complementary Structures	430
Updates in the ODS	430
Historical Data and the ODS	431
Profile Records	432
Different Classes of ODS	434
Database Design — A Hybrid Approach	435