

# Building a Scalable Data Warehouse with Data Vault 2.0

**Daniel Linstedt**

**Michael Olschimke**



AMSTERDAM • BOSTON • HEIDELBERG • LONDON  
NEW YORK • OXFORD • PARIS • SAN DIEGO  
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Morgan Kaufmann is an Imprint of Elsevier



# Contents

Authors Biography.....	xiii
Foreword.....	xv
Preface.....	xvii
Acknowledgments.....	xix
 <b>CHAPTER 1 Introduction to Data Warehousing.....</b>	 <b>1</b>
<b>1.1 History of Data Warehousing .....</b>	<b>2</b>
1.1.1 Decision Support Systems .....	3
1.1.2 Data Warehouse Systems .....	4
<b>1.2 The Enterprise Data Warehouse Environment .....</b>	<b>5</b>
1.2.1 Access.....	5
1.2.2 Multiple Subject Areas .....	5
1.2.3 Single Version of Truth .....	5
1.2.4 Single Version of Facts .....	6
1.2.5 Mission Criticality .....	6
1.2.6 Scalability .....	6
1.2.7 Big Data.....	7
1.2.8 Performance Issues .....	7
1.2.9 Complexity .....	8
1.2.10 Auditing and Compliance.....	9
1.2.11 Costs .....	10
1.2.12 Other Business Requirements.....	11
<b>1.3 Introduction to Data Vault 2.0 .....</b>	<b>11</b>
<b>1.4 Data Warehouse Architecture.....</b>	<b>12</b>
1.4.1 Typical Two-Layer Architecture.....	12
1.4.2 Typical Three-Layer Architecture .....	13
References.....	14
 <b>CHAPTER 2 Scalable Data Warehouse Architecture .....</b>	 <b>17</b>
<b>2.1 Dimensions of Scalable Data Warehouse Architectures .....</b>	<b>17</b>
2.1.1 Workload .....	18
2.1.2 Data Complexity.....	18
2.1.3 Analytical Complexity.....	19
2.1.4 Query Complexity .....	19
2.1.5 Availability .....	20
2.1.6 Security .....	20

2.2	Data Vault 2.0 Architecture .....	21
2.2.1	Business Rules Definition .....	22
2.2.2	Business Rules Application .....	23
2.2.3	Staging Area Layer .....	25
2.2.4	Data Warehouse Layer .....	26
2.2.5	Information Mart Layer .....	27
2.2.6	Metrics Vault .....	27
2.2.7	Business Vault .....	28
2.2.8	Operational Vault .....	29
2.2.9	Managed Self-Service BI .....	30
2.2.10	Other Features .....	31
	References .....	31
<b>CHAPTER 3</b>	<b>The Data Vault 2.0 Methodology .....</b>	<b>33</b>
3.1	Project Planning .....	33
3.1.1	Capability Maturity Model Integration .....	39
3.1.2	Managing the Project .....	42
3.1.3	Defining the Project .....	50
3.1.4	Estimation of the Project .....	54
3.2	Project Execution .....	62
3.2.1	Traditional Software Development Life-Cycle .....	63
3.2.2	Applying Software Development Life-Cycle to the Data Vault 2.0 Methodology .....	67
3.2.3	Parallel Teams .....	69
3.2.4	Technical Numbering .....	71
3.3	Review and Improvement .....	73
3.3.1	Six Sigma .....	74
3.3.2	Total Quality Management .....	81
	References .....	86
<b>CHAPTER 4</b>	<b>Data Vault 2.0 Modeling .....</b>	<b>89</b>
4.1	Introduction to Data Vault Modeling .....	89
4.2	Data Vault Modeling Vocabulary .....	90
4.2.1	Hub Entities .....	91
4.2.2	Link Entities .....	91
4.2.3	Satellite Entities .....	91
4.3	Hub Definition .....	93
4.3.1	Definition of a Business Key .....	95
4.3.2	Hub Entity Structure .....	98
4.3.3	Hub Examples .....	100
4.4	Link Definition .....	101
4.4.1	Reasons for Many-to-Many Relationships .....	103

4.4.2 Flexibility of Links .....	105
4.4.3 Granularity of Links .....	106
4.4.4 Link Unit-of-Work .....	109
4.4.5 Link Entity Structure .....	110
4.4.6 Link Examples .....	111
<b>4.5 Satellite Definition .....</b>	<b>112</b>
4.5.1 Importance of Keeping History .....	114
4.5.2 Splitting Satellites .....	114
4.5.3 Satellite Entity Structure .....	116
4.5.4 Satellite Examples .....	118
4.5.5 Link Driving Key .....	119
References .....	121
<b>CHAPTER 5 Intermediate Data Vault Modeling .....</b>	<b>123</b>
<b>5.1 Hub Applications .....</b>	<b>123</b>
5.1.1 Business Key Consolidation .....	124
<b>5.2 Link Applications .....</b>	<b>127</b>
5.2.1 Link-on-Link .....	127
5.2.2 Same-as Links .....	129
5.2.3 Hierarchical Links .....	129
5.2.4 Nonhistorized Links .....	132
5.2.5 Nondescriptive Links .....	136
5.2.6 Computed Aggregate Links .....	137
5.2.7 Exploration Links .....	139
<b>5.3 Satellite Applications .....</b>	<b>139</b>
5.3.1 Overloaded Satellites .....	139
5.3.2 Multi-Active Satellites .....	141
5.3.3 Status Tracking Satellites .....	143
5.3.4 Effectivity Satellites .....	145
5.3.5 Record Tracking Satellites .....	146
5.3.6 Computed Satellites .....	149
References .....	150
<b>CHAPTER 6 Advanced Data Vault Modeling .....</b>	<b>151</b>
<b>6.1 Point-in-Time Tables .....</b>	<b>151</b>
6.1.1 Point-in-Time Table Structure .....	153
6.1.2 Managed PIT Window .....	156
<b>6.2 Bridge Tables .....</b>	<b>158</b>
6.2.1 Bridge Table Structure .....	159
6.2.2 Comparing PIT Tables with Bridge Tables .....	160
<b>6.3 Reference Tables .....</b>	<b>160</b>
6.3.1 No-History Reference Tables .....	161

6.3.2 History-Based Reference Tables .....	163
6.3.3 Code and Descriptions .....	164
Reference .....	169
<b>CHAPTER 7 Dimensional Modeling .....</b>	<b>171</b>
7.1 Introduction.....	171
7.2 Star Schemas.....	172
7.2.1 Fact Tables.....	174
7.2.2 Dimension Tables .....	176
7.2.3 Querying Star Schemas .....	177
7.3 Multiple Stars.....	179
7.3.1 Conformed Dimensions.....	179
7.4 Dimension Design .....	180
7.4.1 Slowly Changing Dimensions .....	181
7.4.2 Hierarchies.....	183
7.4.3 Snowflake Design.....	189
References.....	193
<b>CHAPTER 8 Physical Data Warehouse Design.....</b>	<b>195</b>
8.1 Database Workloads .....	195
8.1.1 Workload Characteristics .....	196
8.2 Separate Environments for Development, Testing, and Production.....	197
8.2.1 Blue-Green Deployment.....	198
8.3 Microsoft Azure Cloud Computing Platform .....	200
8.4 Physical Data Warehouse Architecture on Premise .....	203
8.4.1 Hardware Architectures and Databases.....	203
8.4.2 Processor Options .....	206
8.4.3 Memory Options.....	207
8.4.4 Storage Options .....	207
8.4.5 Network Options .....	209
8.5 Database Options .....	210
8.5.1 Tempdb Options .....	210
8.5.2 Partitioning .....	211
8.5.3 Filegroups .....	212
8.5.4 Data Compression .....	212
8.6 Setting up the Data Warehouse.....	213
8.6.1 Setting up the Stage Area .....	213
8.6.2 Setting up the Data Vault .....	217
8.6.3 Setting up Information Marts .....	222
8.6.4 Setting up the Meta, Metrics, and Error Marts.....	226
References.....	228

<b>CHAPTER 9 Master Data Management.....</b>	<b>229</b>
9.1 Definitions.....	229
9.1.1 Master Data .....	229
9.1.2 Data Management.....	230
9.1.3 Master Data Management.....	230
9.2 Master Data Management Goals .....	231
9.3 Drivers for Managing Master Data.....	232
9.4 Operational vs. Analytical Master Data Management.....	235
9.5 Master Data Management as an Enabler for Managed Self-Service BI.....	238
9.6 Master Data Management as an Enabler for Total Quality Management .....	239
9.6.1 MDS Object Model .....	241
9.6.2 Master Data Manager .....	249
9.6.3 Explorer .....	250
9.6.4 Version Management.....	252
9.6.5 Integration Management.....	253
9.6.6 System Administration .....	254
9.6.7 User and Group Permissions .....	255
9.7 Creating a Model .....	256
9.7.1 Creating Entities .....	258
9.7.2 Creating Business Rules .....	261
9.8 Importing a Model .....	263
9.9 Integrating MDS with the Data Vault and Operational Systems.....	265
9.9.1 Stage Tables.....	267
9.9.2 Subscription Views.....	278
References.....	282
 <b>CHAPTER 10 Metadata Management.....</b>	 <b>283</b>
10.1 What is Metadata? .....	283
10.1.1 Business Metadata .....	284
10.1.2 Technical Metadata.....	286
10.1.3 Process Execution Metadata .....	287
10.2 Implementing the Meta Mart.....	287
10.2.1 SQL Server BI Metadata Toolkit.....	288
10.2.2 Naming Conventions .....	292
10.2.3 Capturing Source System Definitions.....	296
10.2.4 Capturing Hard Rules .....	298
10.2.5 Capturing Metadata for the Staging Area .....	300
10.2.6 Capturing Requirements to Source Tables .....	301
10.2.7 Capturing Source Tables to Data Vault Tables.....	302
10.2.8 Capturing Soft Rules.....	311
10.2.9 Capturing Data Vault Tables to Information Mart Table Mappings .....	315

10.2.10 Capturing Requirements to Information Mart Tables.....	317
10.2.11 Capturing Access Control Lists and Other Security Measures.....	318
<b>10.3 Implementing the Metrics Vault.....</b>	<b>320</b>
10.3.1 Capturing Performance Data in SQL Server Integration Services .....	323
<b>10.4 Implementing the Metrics Mart.....</b>	<b>333</b>
<b>10.5 Implementing the Error Mart.....</b>	<b>335</b>
10.5.1 Capturing Erroneous Data in SQL Server Integration Services .....	336
References.....	342

## **CHAPTER 11 Data Extraction..... 343**

<b>11.1 Purpose of Staging Area .....</b>	<b>343</b>
<b>11.2 Hashing in the Data Warehouse .....</b>	<b>347</b>
11.2.1 Hash Functions Revisited .....	350
11.2.2 Applying Hash Functions to Data.....	351
11.2.3 Risks of Using Hash Functions.....	355
11.2.4 Hashing Business Keys.....	360
11.2.5 Hashing for Change Detection.....	364
<b>11.3 Purpose of the Load Date .....</b>	<b>370</b>
<b>11.4 Purpose of the Record Source.....</b>	<b>372</b>
<b>11.5 Types of Data Sources .....</b>	<b>373</b>
<b>11.6 Sourcing Flat Files.....</b>	<b>375</b>
11.6.1 Control Flow .....	375
11.6.2 Flat File Connection Manager .....	380
11.6.3 Data Flow.....	383
<b>11.7 Sourcing Historical Data.....</b>	<b>399</b>
11.7.1 SSIS Example for Sourcing Historical Data.....	401
<b>11.8 Sourcing the Sample Airline Data .....</b>	<b>403</b>
11.8.1 Authenticating with Google Drive.....	404
11.8.2 Control Flow .....	406
11.8.3 GoogleSheets Connection Manager.....	411
11.8.4 Data Flow.....	414
<b>11.9 Sourcing Denormalized Data Sources .....</b>	<b>422</b>
<b>11.10 Sourcing Master Data from MDS.....</b>	<b>425</b>
References.....	427

## **CHAPTER 12 Loading the Data Vault..... 429**

<b>12.1 Loading Raw Data Vault Entities .....</b>	<b>432</b>
12.1.1 Hubs .....	434
12.1.2 Links .....	446
12.1.3 No-History Links .....	457
12.1.4 Satellites.....	465
12.1.5 End-Dating Satellites .....	486

12.1.6	Separate New from Changed Rows .....	491
12.1.7	No-History Satellites.....	496
12.1.8	Soft-Deleting Data in Hubs and Links.....	499
12.1.9	Dealing with Missing Data .....	501
<b>12.2</b>	<b>Loading Reference Tables .....</b>	<b>505</b>
12.2.1	No-History Reference Tables .....	506
12.2.2	History-Based Reference Tables.....	509
12.2.3	Code and Descriptions .....	511
12.2.4	Code and Descriptions with History .....	514
<b>12.3</b>	<b>Truncating the Staging Area.....</b>	<b>517</b>
	References.....	518
<b>CHAPTER 13</b>	<b>Implementing Data Quality .....</b>	<b>519</b>
<b>13.1</b>	<b>Business Expectations Regarding Data Quality .....</b>	<b>519</b>
<b>13.2</b>	<b>The Costs of Low Data Quality .....</b>	<b>520</b>
<b>13.3</b>	<b>The Value of Bad Data .....</b>	<b>521</b>
<b>13.4</b>	<b>Data Quality in the Architecture.....</b>	<b>523</b>
<b>13.5</b>	<b>Correcting Errors in the Data Warehouse.....</b>	<b>524</b>
<b>13.6</b>	<b>Transform, Enhance and Calculate Derived Data .....</b>	<b>525</b>
13.6.1	T-SQL Example.....	526
<b>13.7</b>	<b>Standardization of Data .....</b>	<b>528</b>
13.7.1	T-SQL Example.....	528
<b>13.8</b>	<b>Correct and Complete Data .....</b>	<b>530</b>
13.8.1	T-SQL Example.....	531
13.8.2	DQS Example .....	532
13.8.3	SSIS Example .....	537
<b>13.9</b>	<b>Match and Consolidate Data.....</b>	<b>548</b>
13.9.1	SSIS Example .....	550
<b>13.10</b>	<b>Creating Dimensions from Same-as Links .....</b>	<b>560</b>
	References.....	566
<b>CHAPTER 14</b>	<b>Loading the Dimensional Information Mart .....</b>	<b>567</b>
<b>14.1</b>	<b>Using the Business Vault as an Intermediate to the Information Mart.....</b>	<b>567</b>
14.1.1	Computed Satellite.....	567
14.1.2	Building an Exploration Link .....	569
<b>14.2</b>	<b>Materializing the Information Mart.....</b>	<b>579</b>
14.2.1	Loading Type 1 Dimensions.....	580
14.2.2	Loading Type 2 Dimensions.....	582
14.2.3	Loading Fact Tables.....	585
14.2.4	Loading Aggregated Fact Tables .....	590
<b>14.3</b>	<b>Leveraging PIT and Bridge Tables for Virtualization .....</b>	<b>592</b>
14.3.1	Factors that Affect Performance of Virtualized Facts .....	594



14.3.2 Advantages of Virtualization .....	595
14.3.3 Loading PIT Tables .....	596
14.3.4 Creating Virtualized Dimensions.....	601
14.3.5 Loading Bridge Tables.....	604
14.3.6 Creating Virtualized Facts .....	608
<b>14.4</b> Implementing Temporal Dimensions .....	614
<b>14.5</b> Implementing Data Quality Using PIT Tables .....	616
<b>14.6</b> Dealing with Reference Data.....	618
<b>14.7</b> About Hash Keys in the Information Mart .....	620
14.7.1 Advantages of Using Hash Keys in the Information Mart .....	620
14.7.2 Reduce the Number of Dimensions in Cube.....	620
14.7.3 Use Fixed Binary Data Type for Hash Values.....	620
14.7.4 Reduce the Size of the Hash Key.....	621
14.7.5 Introduce Additional Sequence Numbers .....	621
References.....	621
<b>CHAPTER 15 Multidimensional Database .....</b>	<b>623</b>
<b>15.1</b> Accessing the Information Mart .....	624
15.1.1 Creating a Data Source .....	624
15.1.2 Creating Data Source View .....	626
<b>15.2</b> Creating Dimensions .....	631
15.2.1 Date Dimension .....	633
<b>15.3</b> Creating Cubes.....	639
15.3.1 Processing the Cube.....	645
<b>15.4</b> Accessing the Cube.....	646
References.....	647
 Subject Index .....	 649