Contents

Foreword xiii
Preface xv
About the Author xix

CHAPTER 1 Introduction to Data Virtualization 1

1.1 Introduction 1
1.2 The World of Business Intelligence Is Changing 1
1.3 Introduction to Virtualization 3
1.4 What Is Data Virtualization? 4
1.5 Data Virtualization and Related Concepts 5
1.5.1 Data Virtualization versus Encapsulation and Information Hiding 5
1.5.2 Data Virtualization versus Abstraction 6
1.5.3 Data Virtualization versus Data Federation 7
1.5.4 Data Virtualization versus Data Integration 8
1.5.5 Data Virtualization versus Enterprise Information Integration 9
1.6 Definition of Data Virtualization 9
1.7 Technical Advantages of Data Virtualization 10
1.8 Different Implementations of Data Virtualization 14
1.9 Overview of Data Virtualization Servers 14
1.10 Open versus Closed Data Virtualization Servers 15
1.11 Other Forms of Data Integration 16
1.12 The Modules of a Data Virtualization Server 18
1.13 The History of Data Virtualization 19
1.14 The Sample Database: World Class Movies 22
1.15 Structure of This Book 25

CHAPTER 2 Business Intelligence and Data Warehousing 27

2.1 Introduction 27
2.2 What Is Business Intelligence? 27
2.3 Management Levels and Decision Making 28
2.4 Business Intelligence Systems 29
2.5 The Data Stores of a Business Intelligence System 30
2.5.1 The Data Warehouse 30
2.5.2 The Data Marts 34
2.5.3 The Data Staging Area 35
2.5.4 The Operational Data Store 37
2.5.5 The Personal Data Stores 38
2.5.6 A Comparison of the Different Types of Data Stores 38

2.6 Normalized Schemas, Star Schemas, and Snowflake Schemas 39
2.6.1 Normalized Schemas 40
2.6.2 Denormalized Schemas 40
2.6.3 Star Schemas 41
2.6.4 Snowflake Schemas 43
2.7 Data Transformation with Extract Transform Load, Extract Load Transform, and Replication 44
2.7.1 Extract Transform Load 44
2.7.2 Extract Load Transform 45
2.7.3 Replication 46
2.8 Overview of Business Intelligence Architectures 47
2.9 New Forms of Reporting and Analytics 48
2.9.1 Operational Reporting and Analytics 48
2.9.2 Deep and Big Data Analytics 49
2.9.3 Self-Service Reporting and Analytics 49
2.9.4 Unrestricted Ad-Hoc Analysis 50
2.9.5 360-Degree Reporting 51
2.9.6 Exploratory Analysis 51
2.9.7 Text-Based Analysis 52
2.10 Disadvantages of Classic Business Intelligence Systems 53
2.11 Summary 56

CHAPTER 3 Data Virtualization Server: The Building Blocks 59

3.1 Introduction 59
3.2 The High-Level Architecture of a Data Virtualization Server 59
3.3 Importing Source Tables and Defining Wrappers 60
3.4 Defining Virtual Tables and Mappings 62
3.5 Examples of Virtual Tables and Mappings 66
3.6 Virtual Tables and Data Modeling 76
3.7 Nesting Virtual Tables and Shared Specifications 77
3.8 Importing Nonrelational Data 79
3.8.1 XML and JSON Documents 79
3.8.2 Web Services 84
3.8.3 Spreadsheets 86
3.8.4 NoSQL Databases 86
3.8.5 Multidimensional Cubes and MDX 89
3.8.6 Semistructured Data 92
3.8.7 Unstructured Data 95
3.9 Publishing Virtual Tables 96
3.10 The Internal Data Model 101
3.11 Updatable Virtual Tables and Transaction Management 106

CHAPTER 4 Data Virtualization Server: Management and Security 109
4.1 Introduction 109
4.2 Impact and Lineage Analysis 109
4.3 Synchronization of Source Tables, Wrapper Tables, and Virtual Tables 110
4.4 Security of Data: Authentication and Authorization 112
4.5 Monitoring, Management, and Administration 114

CHAPTER 5 Data Virtualization Server: Caching of Virtual Tables 119
5.1 Introduction 119
5.2 The Cache of a Virtual Table 119
5.3 When to Use Caching 120
5.4 Caches versus Data Marts 122
5.5 Where Is the Cache Kept? 122
5.6 Refreshing Caches 123
5.7 Full Refreshing, Incremental Refreshing, and Live Refreshing 124
5.8 Online Refreshing and Offline Refreshing 125
5.9 Cache Replication 126

CHAPTER 6 Data Virtualization Server: Query Optimization Techniques 127
6.1 Introduction 127
6.2 A Refresher Course on Query Optimization 128
6.3 The Ten Stages of Query Processing by a Data Virtualization Server 132
6.4 The Intelligence Level of the Data Stores 134
6.5 Optimization through Query Substitution 134
6.6 Optimization through Pushdown 137
6.7 Optimization through Query Expansion (Query Injection) 139
6.8 Optimization through Ship Joins 140
6.9 Optimization through Sort-Merge Joins 141
6.10 Optimization by Caching 142
6.11 Optimization and Statistical Data 142
6.12 Optimization through Hints 143
6.13 Optimization through SQL Override 143
6.14 Explaining the Processing Strategy 145

CHAPTER 7 Deploying Data Virtualization in Business Intelligence Systems 147
7.1 Introduction 147
7.2 A Business Intelligence System Based on Data Virtualization 147
7.3 Advantages of Deploying Data Virtualization 148
7.4 Disadvantages of Deploying Data Virtualization 151
7.5 Strategies for Adopting Data Virtualization 151
7.5.1 Strategy 1: Introducing Data Virtualization in an Existing Business Intelligence System 152

7.5.2 Strategy 2: Developing a New Business Intelligence System with Data Virtualization 157
7.5.3 Strategy 3: Developing a New Business Intelligence System Combining Source and Transformed Data 161
7.6 Application Areas of Data Virtualization 163
7.6.1 Unified Data Access 163
7.6.2 Virtual Data Mart 163
7.6.3 Virtual Data Warehouse—Based on Data Marts 165
7.6.4 Virtual Data Warehouse—Based on Production Databases 165
7.6.5 Extended Data Warehouse 167
7.6.6 Operational Reporting and Analytics 167
7.6.7 Operational Data Warehouse 168
7.6.8 Virtual Corporate Data Warehouse 169
7.6.9 Self-Service Reporting and Analytics 170
7.6.10 Virtual Sandbox 171
7.6.11 Prototyping 171
7.6.12 Analyzing Semistructured and Unstructured Data 172
7.6.13 Disposable Reports 173
7.6.14 Extending Business Intelligence Systems with External Users 173
7.7 Myths on Data Virtualization 174

CHAPTER 8 Design Guidelines for Data Virtualization 177
8.1 Introduction 177
8.2 Incorrect Data and Data Quality 177
8.2.1 Different Forms of Incorrect Data 178
8.2.2 Integrity Rules and Incorrect Data 179
8.2.3 Filtering, Flagging, and Restoring Incorrect Data 179
8.2.4 Examples of Filtering Incorrect Data 180
8.2.5 Examples of Flagging Incorrect Data 184
8.2.6 Examples of Restoring Misspelled Data 186
8.3 Complex and Irregular Data Structures 188
8.3.1 Codes without Names 188
8.3.2 Inconsistent Key Values 190
8.3.3 Repeating Groups 192
8.3.4 Recursive Data Structures 192
8.4 Implementing Transformations in Wrappers or Mappings 197
8.5 Analyzing Incorrect Data 197
8.6 Different Users and Different Definitions 198
8.7 Time Inconsistency of Data 199
8.8 Data Stores and Data Transmission 200
8.9 Retrieving Data from Production Systems 202

8.10 Joining Historical and Operational Data 203
8.11 Dealing with Organizational Changes 204
8.12 Archiving Data 205

CHAPTER 9 Data Virtualization and Service-Oriented Architecture 207
9.1 Introduction 207
9.2 Service-Oriented Architectures in a Nutshell 207
9.3 Basic Services, Composite Services, Business Process Services, and Data Services 209
9.4 Developing Data Services with a Data Virtualization Server 211
9.5 Developing Composite Services with a Data Virtualization Server 213
9.6 Services and the Internal Data Model 215

CHAPTER 10 Data Virtualization and Master Data Management 217
10.1 Introduction 217
10.2 Data Is a Critical Asset for Every Organization 217
10.3 The Need for a 360-Degree View of Business Objects 219
10.4 What Is Master Data? 219
10.5 What Is Master Data Management? 221
10.6 A Master Data Management System 222
10.7 Master Data Management for Integrating Data 224
10.8 Integrating Master Data Management and Data Virtualization 224

CHAPTER 11 Data Virtualization, Information Management, and Data Governance 231
11.1 Introduction 231
11.2 Impact of Data Virtualization on Information Modeling and Database Design 231
11.3 Impact of Data Virtualization on Data Profiling 234
11.4 Impact of Data Virtualization on Data Cleansing 239
11.5 Impact of Data Virtualization on Data Governance 239

CHAPTER 12 The Data Delivery Platform—A New Architecture for Business Intelligence Systems 243
12.1 Introduction 243
12.2 The Data Delivery Platform in a Nutshell 243
12.3 The Definition of the Data Delivery Platform 244
12.4 The Data Delivery Platform and Other Business Intelligence Architectures 245
12.5 The Requirements of the Data Delivery Platform 247
12.6 The Data Delivery Platform versus Data Virtualization 249
12.7 Explanation of the Name 250
12.8 A Personal Note 251

CHAPTER 13 The Future of Data Virtualization 253
13.1 Introduction 253
13.2 The Future of Data Virtualization According to Rick F. van der Lans 254
13.2.1 New and Enhanced Query Optimization Techniques 254
13.2.2 Exploiting New Hardware Technology 255
13.2.3 Extending the Design Module 256
13.2.4 Data Quality Features 258
13.2.5 Support for the Push-Model for Data Access 258
13.2.6 Blending of Data Virtualization, Extract Transform Load, Extract Load Transform, and Replication 259
13.3 The Future of Data Virtualization According to David Besemer, CTO of Composite Software 260
13.3.1 The Empowered Consumer Gains Ubiquitous Data Access 261
13.3.2 IT's Back Office Becomes the Cloud 261
13.3.3 Data Virtualization of the Future Is a Global Data Fabric 261
13.3.4 Conclusion 262
13.4 The Future of Data Virtualization According to Alberto Pan, CTO of Denodo Technologies 262
13.5 The Future of Data Virtualization According to James Markarian, CTO of Informatica Corporation 264
13.5.1 How to Maximize Return on Data with Data Virtualization 265
13.5.2 Beyond Looking Under the Hood 266

Bibliography 267

Index 269