Open Source Data Warehousing and Business Intelligence

Lakshman Bulusu
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreword</td>
<td>xvii</td>
</tr>
<tr>
<td>Introduction</td>
<td>xix</td>
</tr>
<tr>
<td>What Does This Book Cover?</td>
<td>xxii</td>
</tr>
<tr>
<td>Who Should Read This Book?</td>
<td>xxii</td>
</tr>
<tr>
<td>Why a Separate Book?</td>
<td>xxiii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>xxvii</td>
</tr>
<tr>
<td>About the Author</td>
<td>xxix</td>
</tr>
<tr>
<td>Chapter 1</td>
<td>Introduction</td>
</tr>
<tr>
<td>1.1</td>
<td>In This Chapter</td>
</tr>
<tr>
<td>1.2</td>
<td>Data Warehousing and Business Intelligence: What, Why, How, When, When Not?</td>
</tr>
<tr>
<td>1.2.1</td>
<td>Taking IT Intelligence to Its Apex</td>
</tr>
<tr>
<td>1.3</td>
<td>Open Source DW and BI: Much Ado about Anything-to-Everything DW and BI, When Not, and Why So Much Ado?</td>
</tr>
<tr>
<td>1.3.1</td>
<td>Taking Business Intelligence to Its Apex: Intelligent Content for Insightful Intent</td>
</tr>
<tr>
<td>1.4</td>
<td>Summary</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>Data Warehousing and Business Intelligence: An Open Source Solution</td>
</tr>
<tr>
<td>2.1</td>
<td>In This Chapter</td>
</tr>
<tr>
<td>2.2</td>
<td>What Is Open Source DW and BI, and How &quot;Open&quot; Is This Open?</td>
</tr>
</tbody>
</table>
2.3 What's In, What's Not: Available and Viable Options for Development and Deployment 19
  2.3.1 Semantic Analytics 19
  2.3.2 Testing for Optimizing Quality and Automation—Accelerated! 19
  2.3.3 Business Rules, Real-World Perspective, Social Context 19
  2.3.4 Personalization Through Customizable Measures 20
  2.3.5 Leveraging the Cloud for Deployment 21

2.4 The Foundations Underneath: Architecture, Technologies, and Methodologies 21

2.5 Open Source versus Proprietary DW and BI Solutions: Key Differentiators and Integrators 27

2.6 Open Source DW and BI: Uses and Abuses 28
  2.6.1 An Intelligent Query Accelerator Using an Open Cache In, Cache Out Design 30

2.7 Summary 31

---

Chapter 3  Open Source DW & BI: Successful Players and Products 33

3.1 In This Chapter 33

3.2 Open Source Data Warehousing and Business Intelligence Technology 35
  3.2.1 Licensing Models Followed 35
  3.2.2 Community versus Commercial Open Source 37

3.3 The Primary Vendors: Inventors and Presenters 38
  3.3.1 Oracle: MySQL Vendor 38
  3.3.2 PostgreSQL Vendor 39
  3.3.3 Infobright 41
  3.3.4 Pentaho: Mondrian Vendor 41
  3.3.5 Jedox: Palo Vendor 42
  3.3.6 EnterpriseDB Vendor 42
  3.3.7 Dynamo BI and Eigenbase: LucidDB Vendor 42
  3.3.8 GreenPlum Vendor 43
  3.3.9 Hadoop Project 43
  3.3.10 HadoopDB 44
  3.3.11 Talend 44
3.4 The Primary Products and Tools Set: Inclusions and Exclusions 45
  3.4.1 Open Source Databases 45
  3.4.2 Open Source Data Integration 65
  3.4.3 Open Source Business Intelligence 70
  3.4.4 Open Source Business Analytics 81
3.5 The Primary Users: User, End-User, Customer and Intelligent Customer 89
  3.5.1 MySQL 89
  3.5.2 PostgreSQL 89
  3.5.3 Mondrian Customers 91
  3.5.4 Palo Customers 91
  3.5.5 EnterpriseDB Customers 91
  3.5.6 LucidDB Customers 91
  3.5.7 Greenplum Customers 92
  3.5.8 Talend Customers 92
3.6 Summary 92
3.7 References 93

Chapter 4 Analysis, Evaluation and Selection 99
4.1 In This Chapter 99
4.2 Essential Criteria for Requirements Analysis of an Open Source DW and BI solution 100
4.3 Key and Critical Deciding Factors in Selecting a Solution 102
  4.3.1 The Selection-Action Preview 103
  4.3.2 Raising your BIQ: Five Things Your Company Can Do Now 107
4.4 Evaluation Criteria for Choosing a Vendor-Specific Platform and Solution 110
4.5 The Final Pick: An Information-Driven, Customer-Centric Solution, and a Best-of-Breed Product/Platform and Solution Convergence Key Indicator Checklist 115
4.6 Summary 116
4.7 References 118

Chapter 5 Design and Architecture: Technologies and Methodologies by Dissection 119
5.1 In This Chapter 119
5.2 The Primary Aspects of DW and BI from a Usability Perspective: Strategic BI, Pervasive BI, Operational BI, and BI On-Demand 120
5.3 Design and Architecture Considerations for the Primary BI Perspectives 121
   5.3.1 The Case for Architecture as a Precedence Factor 122
5.4 Information-Centric, Business-Centric, and Customer-Centric Architecture: A Three-in-One Convergence, for Better or Worse 123
5.5 Open Source DW and BI Architecture 125
   5.5.1 Pragmatics and Design Patterns 126
   5.5.2 Components 127
5.6 Why and How an Open Source Architecture Delivers a Better Enterprisewide Solution 128
5.7 Open Source Data Architecture: Under the Hood 131
5.8 Open Source Data Warehouse Architecture: Under the Hood 133
5.9 Open Source BI Architecture: Under the Hood 136
5.10 The Vendor/Platform Product(s)/Tools(s) That Fit into the Open DW and BI Architecture 139
   5.10.1 Information Integration, Usability and Management (Across Data Sources, Applications and Business Domains) 141
   5.10.2 EDW: Models to Management 143
   5.10.3 BI: Models to Interaction to Management to Strategic Business Decision Support (via Analytics and Visualization) 144
5.11 Best Practices: Use and Reuse 146
5.12 Summary 147

Chapter 6 Operational BI and Open Source 149

6.1 In This Chapter 149
6.2 Why a Separate Chapter on Operational BI and Open Source? 150
6.3 Operational BI by Dissection 151
6.4 Design and Architecture Considerations for Operational BI 156
6.5 Operational BI Data Architecture: Under the Hood 157
6.6 A Reusable Information Integration Model: From Real-Time to Right Time 160
6.7 Operational BI Architecture: Under the Hood 161
6.8 Fitting Open Source Vendor/Platform Product(s)/Tools(s) into the Operational BI Architecture 164
6.8.1 Talend Data Integration 164
6.8.2 expressor 3.0 Community Edition 164
6.8.3 Advanced Analytics Engines for Operational BI 165
6.8.4 Astera's Centerprise Data Integration Platform 165
6.8.5 Actuate BIRT BI Platform 165
6.8.6 JasperSoft Enterprise 166
6.8.7 Pentaho Enterprise BI Suite 166
6.8.8 KNIME (Konstanz Information Miner) 167
6.8.9 Pervasive DataRush 167
6.8.10 Pervasive DataCloud2 167
6.9 Best Practices: Use and Reuse 167
6.10 Summary 169

Chapter 7 Development and Deployment 171

7.1 In this Chapter 171
7.2 Introduction 171
7.3 Development Options, Dissected 172
7.4 Deployment Options, Dissected 179
7.5 Integration Options, Dissected 182
7.6 Multiple Sources, Multiple Dimensions 185
7.7 DW and BI Usability and Deployment: Best Solution versus Best-Fit Solution 186
7.8 Leveraging the Best-Fit Solution: Primary Considerations 188
7.9 Better, Faster, Easier as the Hitchhiker's Rule 189
7.9.1 Dynamism and Flash—Real Output in Real Time in the Real World 190
7.9.2 Interactivity 190
7.10 Better Responsiveness, User Adoptability, and Transparency 191
7.11 Fitting the Vendor/Platform Product(s)/Tools(s): A Development and Deployment Standpoint 195
7.12 Best Practices: Use and Reuse 202
7.13 Summary 204
Chapter 8  Best Practices for Data Management  205

8.1   In This Chapter  205
8.2   Introduction  205
8.3   Best Fit of Open Source in EDW Implementation  206
8.4   Best Practices for Using Open Source as a BI-Only Methodology for Data/Information Delivery  208
  8.4.1   Mobile BI and Pervasive BI  208
8.5   Best Practices for the Data Lifecycle in a Typical EDW Lifecycle  210
  8.5.1   Data Quality, Data Profiling, and Data Loss Prevention Components  212
  8.5.2   The Data Integration Component  219
8.6   Best Practices for the Information Lifecycle as It Moves into the BI Lifecycle  230
  8.6.1   The Data Analysis Component: The Dimensions of Data Analysis in Terms of Online Analytics vs. Predictive Analytics vs. Real-Time Analytics vs. Advanced Analytics  230
  8.6.2   Data to Information Transformation and Presentation  236
8.7   Best Practices for Auditing Data Access, as It Makes Its Way via the EDW and Directly (Bypassing the EDW) to the BI Dashboard  252
8.8   Best Practices for Using XML in the Open Source EDW/BI Space  254
8.9   Best Practices for a Unified Information Integrity and Security Framework  255
8.10   Object to Relational Mapping: A Necessity or Just a Convenience?  260
  8.10.1   Synchrony Maintenance  260
  8.10.2   Dynamic Language Interoperability  261
8.11   Summary  262

Chapter 9  Best Practices for Application Management  265

9.1   In This Chapter  265
9.2   Introduction  266
9.3   Using Open Source as an End-to-End Solution Option: How Best a Practice Is It?  266
9.4 Accelerating Application Development: Choice, Design, and Suitability Aspects 267
  9.4.1 Visualization of Content: For Better or Best Fit 271
  9.4.2 Best Practices for Autogenerating Code: A Codeless Alternative to Information Presentation 272
  9.4.3 Automating Querying: Why and When 273
  9.4.4 How Fine Is Fine-Grained? Drawing the Line between Representation of Data at the Lowest Level and a Best-Fit Metadata Design and Presentation 275
9.5 Best Practices for Application Integrity 275
  9.5.1 Sharing Data between EDW and the BI Tiers: Isolation or a Tightrope Methodology 278
  9.5.2 Breakthrough BI: Self-Serviceable BI via a Self-Adaptable Solution 279
  9.5.3 Data-In, Data-Out Considerations: Data-to-Information I/O 280
9.6 Best Practices for Intra- and Interapplication Integration and Interaction 281
  9.6.1 Continuous Activity Monitoring and Event Processing 286
  9.6.2 Best Practices to Leverage Cloud-Based Methodologies 290
9.7 Best Practices for Creative BI Reporting 292
9.8 Summary 297

Chapter 10 Best Practices Beyond Reporting: Driving Business Value 299
10.1 In This Chapter 299
10.2 Introduction 299
10.3 Advanced Analytics: The Foundation for a Beyond-Reporting Approach (Dynamic KPI, Scorecards, Dynamic Dashboarding, and Adaptive Analytics) 300
10.4 Large Scale Analytics: Business-centric and Technology-centric Requirements and Solution Options

10.4.1 Business-centric Requirements 310
10.4.2 Technology-centric Requirements 313

10.5 Accelerating Business Analytics: What to Look for, Look at, and Look Beyond 320

10.6 Delivering Information on Demand and Thereby Performance on Demand

10.6.1 Design Pragmatics 326
10.6.2 Demo Pragmatics 328

10.7 Summary 329

Chapter 11 EDW/BI Development Frameworks 331

11.1 In This Chapter 331
11.2 Introduction 332
11.3 From the Big Bang to the Big Data Bang: The Past, Present, and Future 332
11.4 A Framework for BI Beyond Intelligence

11.4.1 Raising the Bar on BI Using Embeddable BI and BI in the Cloud 335
11.4.2 Raising the Bar on BI: Good to Great to Intelligent 335
11.4.3 Raising the Bar on the Social Intelligence Quotient (SIQ) 338
11.4.4 Raising the Bar on BI by Mobilizing BI: BI on the Go 341

11.5 A Pragmatic Framework for a Customer-Centric EDW/BI Solution 343
11.6 A Next-Generation BI Framework

11.6.1 Taking EDW/BI to the Next Level: An Open Source Model for EDW/BI–EPM 352
11.6.2 Open Source Model for an Open Source DW–BI/EPM Solution Delivering Business Value 353
11.6.3 Open Source Architectural Framework for a Best-Fit Open Source BI/EPM Solution 355
11.6.4 Value Proposition 356
11.6.5 The Road Ahead . . . 357
11.7 A BI Framework for a Reusable Predictive Analytics Model 357
11.9 Summary 360

Chapter 12 Best Practices for Optimization 363
12.1 In This Chapter 363
12.2 Accelerating Application Testing: Choice, Design, and Suitability 364
12.3 Best Practices for Performance Testing: Online and On Demand Scenarios 366
12.4 A Fine Tuning Framework for Optimality 369
12.5 Looking Down the Customer Experience Trail, Leaving the Customer Alone: Customer Feedback Management (CFM)—Driven and APM-Oriented Tuning 372
12.6 Codeful and Codeless Design Patterns for Business-Savvy and IT-Friendly QOS Measurements and In-Depth Impact Analysis 373
12.7 Summary 375

Chapter 13 Open Standards for Open Source: An EDW/BI Outlook 377
13.1 Introduction 377
13.2 Summary 384
13.3 References 385

Index 387