Data Warehouse Performance

W.H. Inmon

Ken Rudin

Christopher K. Buss

Ryan Sousa

WILEY COMPUTER PUBLISHING

John Wiley & Sons, Inc.

New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

Contents

Acknowledgments xi

Preface xiii

Chapter One **Introduction to Data Warehouse Performance** 1

Measuring Performance 2
Productivity and Performance 3
Data Volumes and Performance 4
Why the Growth of Data? 5
Data Gets in the Way 6
Finding Hidden Data 6
User Expectations and Performance 10
Education and Performance 12
Achieving Performance 13
When Should Performance Be Considered? 13
Summary 15

PART ONE Usage, Data, and Performance

Chapter Two **User Community and Performance** 19

Know Your End Users: Farmers and Explorers 19
The World of the Explorer 22
Profile of Execution 26
World of the Farmer 27
Typing the Farmer and the Explorer 30
Clerical versus Management 31

Casual versus Power 32
Predefined versus Ad Hoc 32
Summary versus Detailed 33
Simple versus Complex 34
Summary 34

Chapter Three **Farmers and Explorers** **37**
Optimizing for the Explorer 39
Subsetting and Removing Data 40
Creating a Living Sample 42
Exploration Warehouse 44
Optimizing for The Farmer 54
Strategic Approaches to Performance 55
Tactical Approaches to Performance 57
Operational Approaches to Performance 64
Summary 72

Chapter Four **Data Marts** **75**
What is a Data Mart? 75
Data Mart Community 77
Data Mart Appeal 77
Data Mart Source 79
Building the Data Mart First 80
Different Kinds of Data Marts 81
Loading the Data Mart 82
Metadata in the Data Mart 84
Data Modeling for the Data Mart 85
Purging the Data Mart 86
Data Mart Contents 87
Structure within the Data Mart 87
Performance in the Data Mart 90
Monitoring the Data Mart Environment 91
Summary 92

Chapter Five **Dormant Data** **95**
Understanding Dormant Data 96
Summary Tables and Dormant Data 98
Misjudgment of History and Dormant Data 98
Reality of Requirements and Dormant Data 98
Insistence of Detail and Dormant Data 99

Calculating Dormant Data 99
Finding Dormant Data 101
Removing Dormant Data 102
Selecting Data to be Removed 103
Determining the Probability of Access 105
Summary 107

Chapter Six **Data Cleansing** **109**
How Dirty Data Gets In 109
Cleansing Dirty Data 111
Cleansing the Legacy Environment 111
Cleansing at the Point of Integration 113
Cleansing after Loading 114
Different Kinds of Audits 116
Managing Required Resources 116
Cleansing Data Over Time 119
Bounded Referential Integrity 120
Summary 122

Chapter Seven **Monitors** **125**
Activity Monitors 125
Finding Dormant Data 127
Understanding Dormant Data 129
Removing Dormant Data 130
Capturing Activity Information 131
Reviewing the Output 136
Resource Governors Versus Query Blocking 138
Resource Governors—Nothing New 138
Why Are Resource Governors Inadequate? 138
Query Blocking 142
Monitoring Data Content 143
Data Warehouse Alarm Clock 147
Summary 150

PART TWO Platform and Performance

Chapter Eight **Components of the High-Performance Platform** **155**
Performance Chain 156
Scalability Requirement 156

Parallelism and Its Relationship to Performance 159
What Is Parallelism? 159
Types of Parallelism 162
High-Performance Hardware 166
Symmetric Multiprocessors (SMPs) 167
Clusters 170
Massively Parallel Processors (MPPs) 174
Nonuniform Memory Access (NUMA) 178
High-Performance Databases 182
Parallel Queries 182
Shared-Disk and Shared-Nothing Database Architectures 186
Other Parts of the Performance Chain 190
The Extract/Transform/Load Component 190
End-User Access Tools 193
Scalable Application Frameworks 194
Summary 196

Chapter Nine Building a High-Performance Platform 197
System Architecture 198
Three-Tiered Architectures 199
Two-Tiered Architectures 204
The Solution: Scalable Data Marts 206
Building a Balanced Hardware System 210
Estimating the Business Requirements 211
Determining the Technical Configuration 213
Iterate 216
Designing the Physical Database for Performance 218
Denormalizing the Database 219
Indexing Your Data 226
Designing Your Disk Layout 232
Taking Advantage of I/O Parallelism 234
Striping Techniques 234
Table Partitioning Techniques 242
Optimizing the Queries 248
Execution of an SQL Query 249
Effects of Parallelism 251
CPU Utilization 253
Query-Optimization Questions 254
Summary 256

Chapter Ten **Advanced Platform Topics** **259**

Building a Performance Assurance Environment 259
Defining the Performance Assurance Metrics 260
Building Performance Assurance Tests 263
Very Large Database (VLDB) Issues with Data Warehousing 270
Custom Code Requirements 271
One-In-A-Million Odds Occur Frequently 272
Statistical Effects 273
Algorithm Changes 274
Exceeding Batch Windows 277
Data Warehouses and the Web 278
Web Access Means More Users 280
Web Access Means More Data 282
Data Warehouses and Data Mining 283
Data Mining Requires Scalability 284
The Basics of Scalable Data Mining 286
Data Warehouses and Object-Relational Databases 287
Scalable Performance for Complex Data Types 288
Scalable Functionality 290
Scalability Issues Regarding Object-Relational Technology 290
Summary 294

PART THREE Service Management and Performance

Chapter Eleven **Service Management and the Service Management Contract** **299**

Service Management Defined 300
Business Need for Service Management 302
The Service Management Contract 304
Creating the Service Management Contract 305
Step 1: Establish Relevant Data Warehouse Resources Inventory or Services Catalog 307
Step 2: Characterize Usage of Data Warehouse Resources 308
Step 3: Determine Current or Projected Service Levels 308
Step 4: Understand the Customer's Requirements 309
Step 5: Determine Cost and Feasibility of Customer's Requirements 310

Step 6: Create the Service Management Contract 311
Step 7: Track Compliance to the Service Management Contract 319
Summary 320

Chapter Twelve Putting the Service Management Contract in Motion 323

Putting the SMC in Context 324
Organization Layer 325
Service Dimensions Layer 325
Service Reporting Layer 327
Data Warehouse Administration (DWA) Organization Layer 327
Service Administrator 329
Measuring the Right Metrics 329
DWA May Not Be In IT 330
Challenges Existing DW Support Structure 330
Service Reporting Layer 333
Service Dimensions Layer 335
Query Response Time Dimension 336
User Concurrency Dimension 343
Data Storage Dimension 348
System Availability Dimension 351
Data Currency Dimension 356
Data Quality Dimension 364
Summary 369

PART FOUR Piecing Together the Elements

Chapter Thirteen Delivering a High-Performance Data Warehouse Environment 373

Case Study Overview 375
Company Background 376
Changing Business Landscape 376
Marketing Challenges 377
Justifying the Data Warehouse Environment 378
Building the Team 378
Scoping the Three- to Six-Month Deliverable 380
Delivering the Data Warehouse Environment 383
Assess the Capabilities 384

Aligning the Users, Workload, and Capabilities 388
Design the Databases 390
Data Mart and Exploration Warehouse Designs 393
Data Mart Design 394
Exploration Warehouse Design 400
Data Warehouse Design 404
Configure the Hardware 415
Servicing the Data Warehouse Environment 418
Summary 420

Recommended Reading **421**
Articles 421
Books 429
Book Reviews 430

Index **431**