The Data Webhouse Toolkit

Building the Web-Enabled Data Warehouse

Ralph Kimball

Richard Merz

WILEY COMPUTER PUBLISHING

John Wiley & Sons, Inc.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

Contents

Introduction **1**

Building the Infrastructure for the Revolution 1
The Data Webhouse 3
What This Book Is About 4
Who the Book Is For 7
What You Need to Know 8
How to Use This Book 9
The Purpose of Each Chapter 10
Goals of a Data Webhouse 18
Goals of the Book 19

PART ONE BRINGING THE WEB TO THE WAREHOUSE **21**

Chapter 1 Why Bring the Web to the Warehouse? **23**

Why the Chckstream Is Not Just Another Data Source 24
Analyzing Behavior 26
Ensuring Privacy 29
The Webhouse Architecture 30
The User and the ISP 33
The Public Web Server and Business Transactions 33
The Hot Response Cache 34
The Data Webhouse System 37
Summary 38

Chapter 2 Tracking Website User Actions 41

A Brief Catalog of User Actions 44
Steps in Product Purchase 46
Recognition of Need 46
Trying to Find What's Needed 46
Searching for Information about Alternatives 47
Selection 47
Cross-Selling and Up-Selling 47
Checkout 50
Post-Order Processing 51
Steps in Software or Content Purchase 51
Trials and Demos 52
Elements of Tracking 52
User Origin 53
Session Identification 54
User Identification 56
Behavioral Analysis 59
Entry Point 60
Dwell 60
Querying 61
Intra-Site Navigation 61
Exit Point 63
Associating Diverse Actions 63
The Requirements of Personalization 64
Recognition of Re-visits 64
User Interface and Content Personalization 65
Collateral and Impulse Sales 65
Active Collaborative Filtering 66
Calendar and Lifestyle Events 66
Localization 66
Summary 68

Chapter 3 Using the Clickstream to Make Decisions 69

Decisions About Identifying and Recognizing Customers 71
Customizing Marketing Activities by Identifying Your Customers 71
Targeting Marketing Activities by Clustering Your Customers 73
Deciding Whether to Encourage or Support a Referring Cross-Link 75
Deciding Whether a Customer Is About to Leave Us 76

Decisions About Communicating 77
Deciding Whether a Particular Web Ad Is Working *11*
Deciding If Custom Greetings Are Working 78
Deciding If a Promotion Is Profitable 80
Responding to a Customer's Life Change 81
Improving the Effectiveness of Your Website 82
Fostering a Sense of Community 83
Fundamental Decisions About Your Web Business 83
Deciding which Products and Services We Provide over the Web 84
Providing Real Time Status Tracking of Our Operations 85
Determining If Our Web Business Is Profitable 87
Summary 89

Chapter 4 Understanding the Ciickstream as a Data Source 91

Web Client/Server Interactions—A Brief Tutorial 92
Basic Client I Server Interaction 92
Advertisements 94
The Referrer 94
The Profiler 94
Composite Sites 95
Proxy Servers and Browser Caches 95
Browser Caches 97
Web Server Logs 97
Host 99
Ident 101
Authuser 101
Time 101
Request 101
Status 102
Bytes 102
Referrer 102
User-Agent 102
Filename 104
Time-to-Serve 104
IP Address 104
Server Port 104
Process ID 105
URL 105
Cookies 105
Cookie Contents 107
Cookie Tutorial—Examining Your Own Cookie File 108

Universal System Identifiers 110
Query Strings 110
Templates 111
Summary 112

Chapter 5 Designing the Website to Support Warehousing 113

Monolithic vs. Distributed Web Servers 114
Synchronize Your Servers 115
Time Synchronization Tools and Techniques 116
Content Labels for Pages 119
Content Indexes for Static HTML 120
Content Indexes for Dynamic HTML 121
A Simple Content Index Application 121
Consistent Cookies 122
Null Logging Server 123
Personal Data Repository 126
Building Trust 127
Special Issues in Collecting Information from Children 127
Summary 128

Chapter 6 Building Ciickstream Data Marts 129

A Lightning Tour of Dimensional Modeling 129
Stringing Stars Together 134
Ciickstream Dimensions 139
Calendar Date Dimension 139
Time of Day Dimension 142
Customer Dimension 143
Page Dimension 148
Event Dimension 149
Session Dimension 150
Referral Dimension 151
Product (or Service) Dimension 152
Causal Dimension 154
Business Entity Dimension 155
Ciickstream Tracking Keys 157
The Ciickstream Data Mart 158
A Ciickstream Fact Table to Analyze Complete Sessions 159

A Ciickstream Fact Table to Analyze Individual Page Use 163
Aggregate Ciickstream Fact Tables 167
Summary 168

Chapter 7 Assembling Ciickstream Value Chains 169

The Sales Transaction Data Mart 170
The Customer Communication Data Mart 171
The Web Profitability Data Mart 172
A Supply Chain for a Web Retailer 176
A Policies and Claims Chain for Insurance 178
A Sales Pipeline Chain 180
A Health Care Value Circle 182
Summary 185

Chapter 8 Implementing the Ciickstream Post-Processor 187

Post-Processor Architecture 189
The Page Event Extractor 191
The Content Resolver 192
The Session Identifier 192
Computing Dwell Time 193
Host and Referrer Resolver 195
Summary 197

PART TWO BRINGING THE WAREHOUSE TO THE WEB 199

Chapter 9 Why Bring the Warehouse to the Web? 201

The Web Pulls the Data Warehouse 202
The Web Pushes the Data Warehouse 204
Tightening the User Interface Feedback Loop 205
Mixing Query and Update 206
Speed Is Nonnegotiable 206
The Sun Never Sets on the Data Webhouse 207
Multimedia Merges into Communication 208
The Web Is Mass Customization 209
The Webhouse Is Profoundly Distributed 210
We Must Face Security and Its Cousin, Privacy 211
Summary 212

Chapter 10 Designing the User Experience 215

How the Second Revolution Differs from the First 215
Second Generation User Interface Guidelines 217
Ensure Near-Instantaneous Performance 217
Meet User Expectations 226
Make Each Page a Pleasant Experience 234
Streamline Processes 237
Reassure Users 239
Provide a Means for Resolving Problems 241
Build Trust 243
Provide Communication Hooks 246
Support International Transparency 247
Summary 248

Chapter 11 Driving Data Mining from the Webhouse 251

The Roots of Data Mining 252
The Activities of Data Mining 253
Preparing for Data Mining 255
Data Transformations for Webhouses in General 255
Data Transformations for all Forms of Data Mining 256
Special Data Transformations Depending on the Data Mining Tool 259
Handing the Data to the Data Miner 261
OLAP, Data Mining, and the Webhouse 265
Summary 266

Chapter 12 Creating an International Data Webhouse 269

The Evolving International Web 270
UNICODE 271
Parallel Hypertext and Machine Translation 273
Multilingual Search 275
Time Zone Converter Services 276
Holiday Lookup Services 277
International Webhouse Techniques 278
Synchronize Multiple Time Zones and Time Formats 278
Support Multiple National Calendars and Date Formats 280
Collect Revenue in Multiple Currencies 281
Handle International Names and Addresses 284
Support Variable Number Formats 290

Support International Telephone Numbers 290
Handle Multinational Queries, Reports, and Collating Sequences 290
Apply Localization in the Data Webhouse 292
Summary 292

Chapter 13 Data Webhouse Security 295

Recommended Security Techniques 297
Provide Two-Factor Authentication 297
Secure the Connection 300
Connect the Authenticated User to a Role 302
Access All Webhouse Objects Through the Roles 304
Manage a Security Process, Not a Solution 305
Summary 306

Chapter 14 Scaling the Webhouse 307

The Webhouse Is Not the Web Server 308
Explosive Changes in Ciickstream Activity 309
Web-Enabled Population Growth 310
Increasing Click Rates 311
User-Level Auto-Search 312
Deeper Economic Penetration 312
Sudden Fame 312
IP as a Universal Transport Protocol 313
XML—Universal Transfer 313
Explosive Changes in Demand for Data Warehouse Services 314
Critical Bottlenecks in Hardware and Software 314
Avoiding the Single Bottleneck 315
Avoiding Process Duplication 317
Physical Considerations: Co-Location 317
Operating Systems 318
Programming Languages 319
Databases 319
Query and Reporting Software 320
Balance the Use of E-Mail and Links 321
Hardware Characteristics 321
The Granularity Tradeoff 322
Summary 323

Chapter 15 Managing the Webhouse Project **325**

Define the Project 326
Identify the Roles 328
Front Office: Sponsors and Drivers 328
Coaches: Project Managers and Leads 330
Regular Lineup: Core Project Team 331
Gather Business Requirements and Audit Data 337
Plan and Manage the Implementation 339
Launch the System 340
Loop Back and Do It Again 341
Summary 341

Chapter 16 The Future of Webhousing **343**

CRM Will Continue to Drive Data Webhousing 344
Describing Behavior Better 345
We Will Finally Need Data Mining 346
ISPs Own a Gold Mine 348
Wanted: Better Search Engines 349
Is Data Winning the War over Storage and Speed? 350
Full Inversion of Databases 351
Website Application Logs 351
Everything Is a Module 352
Summary 353

Glossary of Abbreviations and Terms **355**

Bibliography **387**

Index **391**