

Clickstream Data Warehousing

Mark Sweiger
Mark R. Madsen
Jimmy Langston
Howard Lombard



Contents

Chapter 1	A Typical E-Business	3
	Simplistic View of E-Business Architecture	4
	Internet Service Providers	5
	Multiple Internet-Connected Services	6
	Multiple Physical Web Servers	7
	Different Types of Replicated Web Servers	8
	Clustered Business Transaction Systems	9
	The Clickstream Data Warehouse	10
	The Canonical E-Business Architecture	13
	Summary	13
Chapter 2	The Web Application Environment	15
	The Stateless HTTP Transaction Model	15
	Passing Information between HTTP Transactions	17

Using Query Strings	19
Cookies, User Identity, and Web Server Log Records	19
Site Hits, Page Visits, and User Sessions	20
Calling Other Executables Using CGI	21
Using Scripting Languages to Log User Behavior	23
Web Servers, Application Servers, and Dynamically Generated Web Pages	25
Dynamically Generated Web Pages and Search Engines	26
Summary	28
Chapter 3 Clickstream Data Sources and Web Server Log Files	29
Web Server Logs	30
Standard Log File Formats	30
Extending the Web Server Logs	52
Cookies	58
Internal Data Sources	63
Web Site and Log File Analysis Tools	63
Other Web Server-Related Systems	70
Business Applications	76
Customer Contact Systems	77
External Data Sources	79
Content Caching Services	80
Partners and Affiliates	80
Online Advertising Data	80
Syndicated Consumer or Business Data	82
Summary	83
Chapter 4 Using Cookies and Other Mechanisms to Track User Identity	85
Web Programming Techniques for Maintaining Application State	85
Cookies	86
The Query String and URL Rewriting	91
Hidden Form Fields	92
Managing Sessions and Tracking Users	93
Using Cookies to Track Sessions	93

	Using URL Rewriting to Track Sessions	97
	Using Hidden Fields to Track Sessions	97
	Session Management Design Techniques and Trade-offs	98
	Tracking Users	100
	User Identification and User Profiles	107
	Classes of Online Users	108
	User Identification	111
	Processing User Identity Data	117
	Building User Profiles	122
	Web Site Personalization and User Profiles	129
	Basic Personalization Methods	130
	Types of Personalization	131
	Links Between Warehouse and Web Site Profiles	136
	Implications of Privacy Policies for the Clickstream Data Warehouse	137
	Summary	144
Part 2	Building a Clickstream Data Warehouse, Step-by-Step	145
Chapter 5	Planning, Managing, and Staffing a Clickstream Data Warehouse Project	147
	Introduction to the Clickstream Data Warehouse Project Flow	148
	Managing the Project	149
	Phase 1: Project Definition and Planning	150
	Phase 2: Business Requirements Analysis	157
	Phase 3: Data Warehouse Design	165
	Phase 4: Data Warehouse Implementation	175
	Phase 5: Deployment	185
	Project Staffing and Organization	192
	Project Roles	193
	Project Organization	197
	Project Staffing	201
	Summary	208

Chapter 6	The Clickstream Data Warehouse Meta-Schema	209
	Evolving the Meta-Schema from a Sales Analysis Base	211
	The CRM Customer Dimension	211
	The User Activity/Site Hit Fact Table	212
	The User Dimension	214
	The Fiscal and User Time Dimensions	214
	The Physical, Web, and Site Geography Dimensions	216
	The Content and Activity Dimensions	217
	The Internal and External Promotion Dimensions	218
	The User Activity/Site Hit Meta-Schema	219
	Meta-Schema Attributes	219
	Fiscal Time Attributes	220
	User Time Attributes	222
	Physical Geography Attributes	223
	Web Geography Attributes	224
	Site Geography Attributes	226
	User Dimension Attributes	228
	Content Dimension Attributes	229
	Activity Dimension Attributes	229
	Internal Promotion Attributes	231
	External Promotion Attributes	231
	User Activity/Site Hit Fact Table Attributes	232
	The Page Activity and Session Activity Aggregates	237
	The Page Dimension	239
	The Session Dimension	240
	The Session Aggregate	241
	Variation 1: B2B Applications of the Meta-Schema	243
	Adjusting the User Dimension for Business Use	244
	Variation 2: Adding Clickstream Characteristics to Existing Business-Oriented Schemas	246
	Variation 3: Supporting a Large Site with Multiple Replicated Web Servers	250
	Summary	251

Chapter 7	Implementing the Appropriate Clickstream Data Warehouse Technology Infrastructure	253
	Database Support for Clickstream Data Warehouses	254
	Bulk/Batch RDBMS Loaders	255
	Partitioning	257
	Indexing	264
	Specialized Joins	273
	Aggregate Creation, Awareness, and Management	284
	Parallelism	297
	Useful Analytical Extensions to SQL	307
	Disk Drive and Volume Management	311
	Logical Volume Management	311
	Database Objects	321
	Guidelines for Database Object Disk Layout	325
	Choosing the Appropriate Infrastructure Vendors	327
	Database Software	328
	Logical Volume Management (LVM)	
	Software and Disk Subsystems	330
	Summary	330
Chapter 8	Building the Clickstream Extract, Transformation, and Load Mechanism	331
	Extract, Transformation, and Load Architecture	331
	Clickstream ETL Architecture	335
	The More Complex Clickstream Environment	336
	The Clickstream ETL Architecture	337
	Building the ETL Subsystem	339
	Step 1: Data Analysis	339
	Step 2: Making the Web Site Clickstream-Friendly	340
	Step 3: Create the High-Level ETL Design and Architecture	346
	Step 4: Design the Clickstream-Specific Components	358
	Step 5: Design and Build the Dimension Table ETL Components	373

Step 6: Design and Build the Fact Table ETL Components	391
Step 7: Build the Data Loading Mechanism and Integrate the ETL Programs	397
Step 8: Build Support for Data Administration	399
Summary	401
Chapter 9 Analyzing the Data in the Clickstream Data Warehouse	403
OLAP Tools	403
MOLAP Overview	404
ROLAP Overview	406
HOLAP Overview	408
OLAP Today	410
Analytical Features and Techniques	410
Query Tool Related Features and Techniques	410
Data Model–Related Features and Techniques	414
Database Engine Related Features and Techniques	418
Summary	420
Appendix A	423
Index	433