

Contents

CHAPTER 1.1	An Introduction to Data Architecture	1
	Subdividing Data	1
	Repetitive/Nonrepetitive Unstructured Data	2
	The Great Divide of Data	3
	Textual/Nontextual Data	4
	The Different Forms of Data	5
	Business Value	5
CHAPTER 1.2	The Data Infrastructure	7
	Two Types of Repetitive Data	7
	Repetitive Structured Data	7
	Repetitive Big Data	9
	The Two Infrastructures	10
	What's Being Optimized?	10
	Comparing the Two Infrastructures	12
CHAPTER 1.3	The "Great Divide"	13
	Classifying Corporate Data	13
	The "Great Divide"	14
	Repetitive Unstructured Data	14
	Nonrepetitive Unstructured Data	16
	Different Worlds	18
CHAPTER 1.4	Demographics of Corporate Data	21
CHAPTER 1.5	Corporate Data Analysis	27
CHAPTER 1.6	The Life Cycle of Data: Understanding Data Over Time	33
CHAPTER 1.7	A Brief History of Data	39
	Paper Tape and Punch Cards	39
	Magnetic Tapes	40
	Disk Storage	40
	Data Base Management System (DBMS)	41
	Coupled Processors	41

	Online Transaction Processing	42
	Data Warehouse	43
	Parallel Data Management	43
	Data Vault	44
	Big Data	44
	The Great Divide	45
CHAPTER 2.1	The End-State Architecture—The “World Map”	47
	Architectural Components	47
	Different Kinds of Data in the End State Architecture	48
	Shaping the Data Through Models	50
	Where Is the Data Warehouse?	50
	Where Different Types of Questions Are Answered	
	Across the End State Architecture	52
	Data in the Data Lake	53
	Metadata in the End State Architecture	54
	Networked Metadata	55
	An Evolutionary Experience	55
	The Data Lake Architecture	57
CHAPTER 3.1	Transformations in the End-State Architecture	59
	Redundant Data	59
	Transformations	60
	Customizing Data	61
	Transforming Text	61
	Transforming Application Data	62
	Transforming Data Into a Customized State	63
	Transforming Data Into Bulk Storage	63
	Transforming Data Generated Automatically	64
	Transforming Bulk Data	65
	Transformation and Redundancy	66
CHAPTER 4.1	A Brief History of Big Data	67
	An Analogy—Taking the High Ground	67
	Taking the High Ground	67
	Standardization With the 360	69
	Online Transaction Processing	69
	Enter Teradata and MPP Processing	70
	Then Came Hadoop and Big Data	70
	IBM and Hadoop	71
	Holding the High Ground	71

	Why Do We Need Many to Many Link Structures?.....	147
	Primary Key Options for Data Vault 2.0.....	148
	Sequence Numbers	148
	Hash Keys	150
	Business Keys.....	152
	Source System Sequence Business Keys.....	153
	Multipart Source Business Keys.....	155
CHAPTER 6.3	Introduction to Data Vault Architecture.....	157
	What Is a Data Vault 2.0 Architecture?.....	157
	How Does NoSQL Fit in to the Architecture?.....	158
	What Are the Objectives of the Data Vault 2.0 Architecture?	159
	What Is the Objective of the Data Vault 2.0 Model?	159
	What Are Hard and Soft Business Rules?.....	160
	How Does Managed Self Service BI Fit in the Architecture?	161
CHAPTER 6.4	Introduction to Data Vault Methodology	163
	Data Vault 2.0 Methodology Overview.....	163
	How Does CMMI Contribute to the Methodology?	163
	If CMMI Is So Great, Why Should We Care About	
	Agility Then?.....	166
	Why Include PMP, SDLC If CMMI and Agile Should Be	
	All That's Needed?.....	167
	So Then, What Does Six Sigma Contribute to the	
	Data Vault 2 Methodology?.....	168
	Where Does TQM (Total Quality Management) Fit in to	
	All of This?	169
CHAPTER 6.5	Introduction to Data Vault Implementation	171
	Implementation Overview.....	171
	What's So Important About Patterns?	171
	Why Does Reengineering Happen Because of Big Data?	172
	Why Do We Need to Virtualize Our Data Marts?	174
	What Is Managed Self-Service BI?.....	175
CHAPTER 7.1	The Operational Environment: A Short History	177
	Commercial Uses of the Computer	177
	The First Applications	178
	Ed Yourdon and the Structured Revolution	178
	The SDLC	179
	Disk Technology.....	179
	Enter the DBMS	180
	Response Time and Availability	181
	Corporate Computing Today	183

- Taxonomies in Multiple Languages 115
- Commercial or Private Taxonomies?..... 115
- Dynamics of Taxonomies and Textual Disambiguation..... 116
- Taxonomies and Textual Disambiguation—Separate Technologies 116
- Different Types of Taxonomies 117
- Taxonomies—Maintenance Over Time..... 118
- CHAPTER 5.1** The Siloed Application Environment 121
 - The Challenge of Siloed Applications 121
 - Building Siloed Applications..... 124
 - What Does a Siloed Application Look Like? 126
 - Current Valued Data 126
 - Minimal Historical Data..... 127
 - High Availability 128
 - Overlap Between Siloed Applications 129
 - Frozen Business Requirements..... 130
 - Dismantling Siloed Applications 131
- CHAPTER 6.1** Introduction to Data Vault 2.0 133
 - Data Vault Origins and Background..... 133
 - The “Old” Data Vault 1.0 134
 - The New and Updated Data Vault 2.0 134
 - What Is Data Vault 2.0 Modeling? 136
 - A Business View 136
 - A Technical View..... 136
 - How Is Data Vault 2.0 Methodology Defined?..... 137
 - A Business View 137
 - A Technical View..... 137
 - Why Do We Need a Data Vault 2.0 Architecture?..... 138
 - Where Does Data Vault 2.0 Implementation Fit?..... 138
 - What Are the Business Benefits of Data Vault 2.0?..... 138
 - What Is Data Vault 1.0? 139
- CHAPTER 6.2** Introduction to Data Vault Modeling 141
 - What Is a Data Vault Model Concept? 141
 - Data Vault Model Defined..... 141
 - Components of a Data Vault Model 142
 - What Makes Business Keys So Interesting? 143
 - What Does This Have to Do With Data Vault and Data Warehousing?..... 144
 - How Does This Translate to Data Vault Modeling?..... 144
 - Why Restructure the Data From the Staging Area? 146
 - What Are the Basic Rules of the Data Vault Model?..... 146

	Why Do We Need Many to Many Link Structures?.....	147
	Primary Key Options for Data Vault 2.0.....	148
	Sequence Numbers.....	148
	Hash Keys.....	150
	Business Keys.....	152
	Source System Sequence Business Keys.....	153
	Multipart Source Business Keys.....	155
CHAPTER 6.3	Introduction to Data Vault Architecture.....	157
	What Is a Data Vault 2.0 Architecture?.....	157
	How Does NoSQL Fit in to the Architecture?.....	158
	What Are the Objectives of the Data Vault 2.0 Architecture?.....	159
	What Is the Objective of the Data Vault 2.0 Model?.....	159
	What Are Hard and Soft Business Rules?.....	160
	How Does Managed Self Service BI Fit in the Architecture?.....	161
CHAPTER 6.4	Introduction to Data Vault Methodology.....	163
	Data Vault 2.0 Methodology Overview.....	163
	How Does CMMI Contribute to the Methodology?.....	163
	If CMMI Is So Great, Why Should We Care About	
	Agility Then?.....	166
	Why Include PMP, SDLC If CMMI and Agile Should Be	
	All That's Needed?.....	167
	So Then, What Does Six Sigma Contribute to the	
	Data Vault 2 Methodology?.....	168
	Where Does TQM (Total Quality Management) Fit in to	
	All of This?.....	169
CHAPTER 6.5	Introduction to Data Vault Implementation.....	171
	Implementation Overview.....	171
	What's So Important About Patterns?.....	171
	Why Does Reengineering Happen Because of Big Data?.....	172
	Why Do We Need to Virtualize Our Data Marts?.....	174
	What Is Managed Self-Service BI?.....	175
CHAPTER 7.1	The Operational Environment: A Short History.....	177
	Commercial Uses of the Computer.....	177
	The First Applications.....	178
	Ed Yourdon and the Structured Revolution.....	178
	The SDLC.....	179
	Disk Technology.....	179
	Enter the DBMS.....	180
	Response Time and Availability.....	181
	Corporate Computing Today.....	183

CHAPTER 7.2	The Standard Work Unit	185
	Elements of Response Time	185
	An Hourglass Analogy	185
	The Racetrack Analogy	187
	Your Vehicle Runs as Fast as the Vehicle in Front of It	188
	The Standard Work Unit	188
	The SLA	188
CHAPTER 7.3	Data Modeling for the Structured Environment	191
	The Purpose of the Roadmap	191
	Granular Data Only	191
	The ERD	192
	The Dis	193
	Physical Data Base Design	194
	Relating the Different Levels of the Data Model	195
	An Example of the Linkage	196
	Generic Data Models	197
	Operational Data Models/Data Warehouse Data Models	198
CHAPTER 8.1	A Brief History of Data Architecture	199
CHAPTER 8.2	Big Data/Existing System Interface	211
	The Big Data/Existing Systems Interface	211
	The Repetitive Raw Big Data/Existing Systems Interface	211
	Exception Based Data	213
	The Nonrepetitive Raw Big Data/Existing Systems Interface	214
	Into the Existing Systems Environment	215
	The "Context Enriched" Big Data Environment	215
	Analyzing Structured Data/Unstructured Data Together	217
CHAPTER 8.3	The Data Warehouse/Operational Environment Interface	219
	The Operational/Data Warehouse Interface	219
	The Classical ETL Interface	220
	The ODS and the ETL Interface	220
	The Staging Area	221
	Changed Data Capture	222
	Inline Transformation	223
	ELT Processing	223
CHAPTER 8.4	Data Architecture: A High-Level Perspective	225
	A High Level Perspective	225
	Redundancy	225

	The System of Record	226
	Different Types of Questions.....	228
	Different Communities	229
CHAPTER 9.1	Repetitive Analytics: Some Basics	231
	Different Kinds of Analysis.....	231
	Looking for Patterns.....	232
	Heuristic Processing	234
	Freezing Data.....	236
	The Sandbox	237
	The “Normal” Profile.....	238
	Distillation, Filtering.....	238
	Subsetting Data	240
	Bias of the Sample	241
	Filtering Data	242
	Repetitive Data and Context.....	243
	Linking Repetitive Records	244
	Log Tape Records.....	245
	Analyzing Points of Data	246
	Outliers	247
	Data Over Time	247
CHAPTER 9.2	Analyzing Repetitive Data.....	251
	Log Data.....	253
	Active/Passive Indexing of Data	255
	Summary/Detailed Data	256
	Metadata in Big Data	257
	Linking Data.....	259
CHAPTER 9.3	Repetitive Analysis.....	261
	Internal, External Data	261
	Universal Identifiers	262
	Security	263
	Filtering, Distillation.....	265
	Archiving Results.....	266
	Metrics	267
CHAPTER 10.1	Nonrepetitive Data	269
	Inline Contextualization.....	272
	Taxonomy/Ontology Processing.....	273
	Custom Variables	274
	Homographic Resolution.....	275
	Acronym Resolution	277

	Negation Analysis	277
	Numeric Tagging	278
	Date Tagging	279
	Date Standardization	280
	List Processing	280
	Associative Word Processing	281
	Stop Word Processing	282
	Word Stemming	283
	Document Metadata	284
	Document Classification	284
	Proximity Analysis	285
	Functional Sequencing Within Textual ETL	286
	Internal Referential Integrity	287
	Preprocessing, Postprocessing	287
CHAPTER 10.2	Mapping	291
CHAPTER 10.3	Analytics From Nonrepetitive Data	295
	Call Center Information	295
	Medical Records	304
CHAPTER 11.1	Operational Analytics: Response Time	309
	Transaction Response Time	311
CHAPTER 12.1	Operational Analytics	319
	Different Perspectives of Data	324
	Data Marts	325
	The Operational Data Store—ODS	326
CHAPTER 13.1	Personal Analytics	331
CHAPTER 14.1	Data Models Across the End-State Architecture	337
	The Different Data Models	337
	Functional Decomposition and Data Flow Diagrams	337
	The Corporate Data Model	340
	The Star Join/Dimensional Data Model	343
	Taxonomies/Ontologies	345
	The Selective Subdivision of Data	347
	Proactive/Reactive Data Models	349
CHAPTER 15.1	The System of Record	353
	The End User Cycle of Awareness	353
	The System of Record	354
	The System of Record in the End State Architecture	355
	The Role of Age in the System of Record	356
	A Simple Example	356
	The Flow of Data in the System of Record	357
	Other Data Than the System of Record	358

Is Data Updated in the System of Record?	358
Detailed and Summary Data in the System of Record	359
Auditing Data and the System of Record	360
Text and the System of Record	360
CHAPTER 16.1 Business Value and the End-State Architecture	363
The Evolution of the End State Architecture	363
What is Meant by “Business Value”	363
Tactical Business Value/Strategic Business Value	364
Volume of Data Versus Business Value	365
The “Million in One” Syndrome	366
Where Business Value Occurs	366
Data Relevancy Over Time	367
Where Tactical Decisions Are Made	368
CHAPTER 17.1 Managing Text	371
The Challenge of Text	371
The Challenge of Context	374
The Processing Components of Textual ETL	376
Secondary Analysis	377
Visualization	378
Merging Text Based Data and Structured Data	378
CHAPTER 18.1 An Introduction to Data Visualizations	381
Introduction to Data Visualizations—Overview	381
Purpose and Context	382
Visualization—A Science and an Art	382
Visualization Framework	383
Step 1: Define	383
Step 2: Data	385
Types of Data	385
Data Sources	386
Data Organization	387
Data Quality	387
Step 3: Design	388
Forms of Visualizations	388
Step 4: Distribute	393
Purpose: To Inform or Educate	393
Purpose: To Interact or Explore	394
Data Visualization Tools and Software	394
Summary	395
GLOSSARY	397
INDEX	409