

BIG DATA

CONCEPTS, WAREHOUSING, AND ANALYTICS

MARIBEL YASMINA SANTOS

CARLOS COSTA



CONTENTS

List of Figures	XI
List of Tables	XVII
The Authors	XIX
Acknowledgments	XXI
Foreword	XXIII
Notation	XXV
1. Introduction	1
1.1. Objectives of this Book	4
1.2. Intended Audience	7
1.3. Book Structure	7
2. Big Data Concepts, Techniques, and Technologies	9
2.1. Big Data Relevance	10
2.2. Big Data Characteristics	12
2.3. Big Data Challenges	16
2.3.1. Big Data General Dilemmas	16
2.3.2. Challenges in the Big Data Life Cycle	17
2.3.3. Big Data in Secure, Private, and Monitored Environments	19
2.3.4. Organizational Change	20
2.4. Techniques for Big Data Solutions	21
2.4.1. Big Data Life Cycle and Requirements	23
2.4.1.1. General Steps to Process and Analyze Big Data	23
2.4.1.2. Architectural and Infrastructural Requirements	25

2.4.2.	The Lambda Architecture	27
2.4.3.	Towards Standardization: the NIST Reference Architecture	28
2.5.	Big Data Technologies	30
2.5.1.	Hadoop and Related Projects	30
2.5.2.	Landscape of Distributed SQL Engines	32
2.5.3.	Other Technologies for Big Data Analytics	35
3.	OLTP-oriented Databases for Big Data Environments	37
3.1.	NoSQL and NewSQL: an Overview	38
3.2.	NoSQL Databases	41
3.2.1.	Key-value Databases	41
3.2.1.1.	Overview	41
3.2.1.2.	Redis	42
3.2.2.	Column-oriented Databases	49
3.2.2.1.	Overview	50
3.2.2.2.	HBase	51
3.2.2.3.	From Relational Models to HBase Data Models	57
3.2.3.	Document-oriented Databases	69
3.2.3.1.	Overview	69
3.2.3.2.	MongoDB	71
3.2.4.	Graph Databases	79
3.2.4.1.	Overview	79
3.2.4.2.	Neo4j	82
3.3.	NewSQL Databases and Translytical Databases	88
4.	OLAP-oriented Databases for Big Data Environments	93
4.1.	Hive: the <i>De Facto</i> SQL-on-Hadoop Engine	94
4.1.1.	Data Storage Formats	98
4.1.1.1.	Text File	99
4.1.1.2.	Sequence File	100
4.1.1.3.	RCFile	105
4.1.1.4.	ORC File	107
4.1.1.5.	Avro File	111
4.1.1.6.	Parquet	112
4.1.2.	Partitions and Buckets	113

4.2.	From Dimensional Models to Tabular Models	119
4.2.1.	Primary Data Tables	121
4.2.2.	Derived Data Tables	125
4.3.	Optimizing OLAP workloads with Druid	131
5.	Design and Implementation of Big Data Warehouses	143
5.1.	Big Data Warehousing: an Overview	144
5.2.	Model of Logical Components and Data Flows	147
5.2.1.	Data Provider and Data Consumer	149
5.2.2.	Big Data Application Provider	149
5.2.3.	Big Data Framework Provider	151
5.2.3.1.	Messaging/Communications, Resource Management, and Infrastructures	152
5.2.3.2.	Processing	153
5.2.3.3.	Storage: Data Organization and Distribution	154
5.2.4.	System Orchestrator and Security, Privacy, and Management	157
5.3.	Model of Technological Infrastructure	158
5.4.	Method for Data Modeling	163
5.4.1.	Analytical Objects and their Related Concepts	164
5.4.2.	Joining, Uniting, and Materializing Analytical Objects	167
5.4.3.	Dimensional Big Data with Outsourced Descriptive Families	169
5.4.4.	Data Modeling Best Practices	171
5.4.4.1.	Using Null Values	171
5.4.4.2.	Date, Time, and Spatial Objects vs. Separate Temporal and Spatial Attributes	172
5.4.4.3.	Immutable vs. Mutable Records	173
5.4.5.	Data Modeling Advantages and Disadvantages	174
6.	Big Data Warehouses Modeling: From Theory to Practice	177
6.1.	Multinational Bicycle Wholesale and Manufacturing	178
6.1.1.	Fully Flat or Fully Dimensional Data Models	180
6.1.2.	Nested Attributes	181
6.1.3.	Streaming and Random Access on Mutable Analytical Objects	182
6.2.	Brokerage Firm	183
6.2.1.	Unnecessary Complementary Analytical Objects and Update Problems	183

6.2.1.1.	The Traditional Way of Handling SCD-like Scenarios	185
6.2.1.2.	A New Way of Handling SCD-like Scenarios	185
6.2.2.	Joining Complementary Analytical Objects	186
6.2.3.	Data Science Models and Insights as a Core Value	186
6.2.4.	Partition Keys for Streaming and Batch Analytical Objects	187
6.3.	Retail	188
6.3.1.	Simpler Data Models: Dynamic Partitioning Schemas	189
6.3.2.	Considerations for Spatial Objects	189
6.3.3.	Analyzing Non-Existing Events	190
6.3.4.	Wide Descriptive Families	190
6.3.5.	The Need for Joins in Data CPE Workloads	191
6.4.	Code Version Control System	192
6.5.	A Global Database of Society – The GDELT Project	193
6.6.	Air Quality	194
7.	Fueling Analytical Objects in Big Data Warehouses	197
7.1.	From Traditional Data Warehouses	198
7.2.	From OLTP NoSQL Databases	200
7.3.	From Semi-structured Data Sources	202
7.4.	From Streaming Data Sources	204
7.5.	Using Data Science Models	210
7.5.1.	Data Mining/Machine Learning Models for Structured Data	211
7.5.2.	Text Mining, Image Mining, and Video Mining Models	216
8.	Evaluating the Performance of Big Data Warehouses	219
8.1.	The SSB+ Benchmark	220
8.1.1.	Data Model and Queries	220
8.1.2.	System Architecture and Infrastructure	221
8.2.	Batch OLAP	223
8.2.1.	Comparing Flat Analytical Objects with Star Schemas	223
8.2.2.	Improving Performance with Adequate Data Partitioning	227
8.2.3.	The Impact of Dimensions' Size in Star Schemas	230
8.2.4.	The Impact of Nested Structures in Analytical Objects	232
8.2.5.	Drill Across Queries and Window and Analytics Functions	234

8.3.	Streaming OLAP	236
8.3.1.	The Impact of Data Volume in the Streaming Storage Component	236
8.3.2.	Considerations for Effective and Efficient Streaming OLAP	239
8.4.	SQL-on-Hadoop Systems under Multi-User Environments	242
9.	Big Data Warehousing in Smart Cities	245
9.1.	Logical Components, Data Flows, and Technological Infrastructure	246
9.1.1.	SusCity Architecture	247
9.1.2.	SusCity Infrastructure	250
9.2.	SusCity Data Model	251
9.2.1.	Buildings Characteristics as an Outsourced Descriptive Family	254
9.2.2.	Nested Structures in Analytical Objects	255
9.3.	The Inter-storage Pipeline	255
9.4.	The SusCity Data Visualization Platform	256
9.4.1.	City's Energy Consumption	257
9.4.2.	City's Energy Grid Simulations	258
9.4.3.	Buildings' Performance Analysis and Simulation	258
9.4.4.	Mobility Patterns Analysis	260
10.	Conclusion	263
10.1.	Synopsis of the Book	265
10.2.	Contributions to the State of the Art	270
	References	271
	Index	281