

# Contents

<b>Acknowledgments</b> . . . . .	<b>xvii</b>
<b>Reader Services</b> . . . . .	<b>xviii</b>

## **PART I: THE FUNDAMENTALS OF BIG DATA**

### **CHAPTER 1: Understanding Big Data . . . . . 3**

Concepts and Terminology . . . . .	5
Datasets . . . . .	5
Data Analysis . . . . .	6
Data Analytics . . . . .	6
<i>Descriptive Analytics</i> . . . . .	8
<i>Diagnostic Analytics</i> . . . . .	9
<i>Predictive Analytics</i> . . . . .	10
<i>Prescriptive Analytics</i> . . . . .	11
Business Intelligence (BI) . . . . .	12
Key Performance Indicators (KPI) . . . . .	12
Big Data Characteristics . . . . .	13
Volume . . . . .	14
Velocity . . . . .	14
Variety . . . . .	15
Veracity . . . . .	16
Value . . . . .	16
Different Types of Data . . . . .	17
Structured Data . . . . .	18
Unstructured Data . . . . .	19
Semi-structured Data . . . . .	19
Metadata . . . . .	20
Case Study Background . . . . .	20
History . . . . .	20
Technical Infrastructure and Automation Environment . . . . .	21
Business Goals and Obstacles . . . . .	22

Case Study Example . . . . .24

    Identifying Data Characteristics . . . . .26

*Volume* . . . . .26

*Velocity* . . . . .26

*Variety* . . . . .26

*Veracity* . . . . .26

*Value* . . . . .27

    Identifying Types of Data . . . . .27

**CHAPTER 2: Business Motivations and Drivers  
for Big Data Adoption. . . . .29**

Marketplace Dynamics . . . . .30

Business Architecture . . . . .33

Business Process Management . . . . .36

Information and Communications Technology . . . . .37

    Data Analytics and Data Science . . . . .37

    Digitization . . . . .38

    Affordable Technology and Commodity Hardware . . . . .38

    Social Media . . . . .39

    Hyper-Connected Communities and Devices . . . . .40

    Cloud Computing . . . . .40

Internet of Everything (IoE) . . . . .42

Case Study Example . . . . .43

**CHAPTER 3: Big Data Adoption and Planning  
Considerations . . . . .47**

Organization Prerequisites . . . . .49

Data Procurement . . . . .49

Privacy . . . . .49

Security . . . . .50

Provenance . . . . .51

- Limited Realtime Support. . . . . 52
- Distinct Performance Challenges. . . . . 53
- Distinct Governance Requirements . . . . . 53
- Distinct Methodology . . . . . 53
- Clouds . . . . . 54
- Big Data Analytics Lifecycle . . . . . 55
  - Business Case Evaluation . . . . . 56
  - Data Identification . . . . . 57
  - Data Acquisition and Filtering . . . . . 58
  - Data Extraction . . . . . 60
  - Data Validation and Cleansing . . . . . 62
  - Data Aggregation and Representation. . . . . 64
  - Data Analysis . . . . . 66
  - Data Visualization . . . . . 68
  - Utilization of Analysis Results. . . . . 69
- Case Study Example . . . . . 71
  - Big Data Analytics Lifecycle. . . . . 73
  - Business Case Evaluation . . . . . 73
  - Data Identification . . . . . 74
  - Data Acquisition and Filtering . . . . . 74
  - Data Extraction . . . . . 74
  - Data Validation and Cleansing. . . . . 75
  - Data Aggregation and Representation. . . . . 75
  - Data Analysis . . . . . 75
  - Data Visualization . . . . . 76
  - Utilization of Analysis Results. . . . . 76

**CHAPTER 4: Enterprise Technologies and Big Data Business Intelligence . . . . . 77**

- Online Transaction Processing (OLTP) . . . . . 78
- Online Analytical Processing (OLAP) . . . . . 79
- Extract Transform Load (ETL) . . . . . 79

- Data Warehouses . . . . .80
- Data Marts . . . . .81
- Traditional BI . . . . .82
  - Ad-hoc Reports . . . . .82
  - Dashboards . . . . .82
- Big Data BI. . . . .84
  - Traditional Data Visualization . . . . .84
  - Data Visualization for Big Data . . . . .85
- Case Study Example . . . . .86
  - Enterprise Technology . . . . .86
  - Big Data Business Intelligence. . . . .87

**PART II: STORING AND ANALYZING BIG DATA**

**CHAPTER 5: Big Data Storage Concepts. . . . .91**

- Clusters . . . . .93
- File Systems and Distributed File Systems . . . . .93
- NoSQL . . . . .94
- Sharding. . . . .95
- Replication . . . . .97
  - Master-Slave. . . . .98
  - Peer-to-Peer . . . . .100
- Sharding and Replication. . . . .103
  - Combining Sharding and Master-Slave Replication. . . . .104
  - Combining Sharding and Peer-to-Peer Replication. . . . .105
- CAP Theorem. . . . .106
- ACID . . . . .108
- BASE . . . . .113
- Case Study Example . . . . .117

**CHAPTER 6: Big Data Processing Concepts . . . . . 119**

- Parallel Data Processing . . . . . 120
- Distributed Data Processing . . . . . 121
- Hadoop . . . . . 122
- Processing Workloads . . . . . 122
  - Batch . . . . . 123
  - Transactional . . . . . 123
- Cluster . . . . . 124
- Processing in Batch Mode . . . . . 125
  - Batch Processing with MapReduce . . . . . 125
  - Map and Reduce Tasks . . . . . 126
    - Map* . . . . . 127
    - Combine* . . . . . 127
    - Partition* . . . . . 129
    - Shuffle and Sort* . . . . . 130
    - Reduce* . . . . . 131
  - A Simple MapReduce Example . . . . . 133
  - Understanding MapReduce Algorithms . . . . . 134
- Processing in Realtime Mode . . . . . 137
  - Speed Consistency Volume (SCV) . . . . . 137
  - Event Stream Processing . . . . . 140
  - Complex Event Processing . . . . . 141
  - Realtime Big Data Processing and SCV . . . . . 141
  - Realtime Big Data Processing and MapReduce . . . . . 142
- Case Study Example . . . . . 143
  - Processing Workloads . . . . . 143
  - Processing in Batch Mode . . . . . 143
  - Processing in Realtime . . . . . 144

**CHAPTER 7: Big Data Storage Technology . . . . . 145**

- On-Disk Storage Devices . . . . . 147
  - Distributed File Systems . . . . . 147
  - RDBMS Databases . . . . . 149

NoSQL Databases . . . . .	152
<i>Characteristics</i> . . . . .	152
<i>Rationale</i> . . . . .	153
<i>Types</i> . . . . .	154
<i>Key-Value</i> . . . . .	156
<i>Document</i> . . . . .	157
<i>Column-Family</i> . . . . .	159
<i>Graph</i> . . . . .	160
NewSQL Databases . . . . .	163
In-Memory Storage Devices . . . . .	163
In-Memory Data Grids . . . . .	166
<i>Read-through</i> . . . . .	170
<i>Write-through</i> . . . . .	170
<i>Write-behind</i> . . . . .	172
<i>Refresh-ahead</i> . . . . .	172
In-Memory Databases . . . . .	175
Case Study Example . . . . .	179

## **CHAPTER 8: Big Data Analysis Techniques . . . . . 181**

Quantitative Analysis . . . . .	183
Qualitative Analysis . . . . .	184
Data Mining . . . . .	184
Statistical Analysis . . . . .	184
A/B Testing . . . . .	185
Correlation . . . . .	186
Regression . . . . .	188
Machine Learning . . . . .	190
Classification (Supervised Machine Learning) . . . . .	190
Clustering (Unsupervised Machine Learning) . . . . .	191
Outlier Detection . . . . .	192
Filtering . . . . .	193

- Semantic Analysis . . . . . 195
  - Natural Language Processing . . . . . 195
  - Text Analytics . . . . . 196
  - Sentiment Analysis . . . . . 197
- Visual Analysis . . . . . 198
  - Heat Maps . . . . . 198
  - Time Series Plots . . . . . 200
  - Network Graphs . . . . . 201
  - Spatial Data Mapping . . . . . 202
- Case Study Example . . . . . 204
  - Correlation . . . . . 204
  - Regression . . . . . 204
  - Time Series Plot . . . . . 205
  - Clustering . . . . . 205
  - Classification . . . . . 205

**APPENDIX A: Case Study Conclusion . . . . . 207**

**About the Authors . . . . . 211**

- Thomas Erl . . . . . 211
- Wajid Khattak . . . . . 211
- Paul Buhler . . . . . 212

**Index . . . . . 213**