

# Contents

|   |      |
|---|------|
| Contents .....                                    | v    |
| Preface.....                                      | xi   |
| Acknowledgments .....                             | xiii |
| 1 Introduction.....                               | 1    |
| 2 State of Computing.....                         | 10   |
| 2.1 Microprocessors and Moore’s Law .....         | 10   |
| 2.1.1 Uniprocessor Era.....                       | 10   |
| 2.1.2 The Power Wall.....                         | 12   |
| 2.1.3 Multicore Era .....                         | 14   |
| 2.1.4 Heterogeneous Core Era .....                | 17   |
| 2.2 High-Performance Computing.....               | 19   |
| 2.2.1 Traditional Supercomputing .....            | 20   |
| 2.2.2 Multiprocessing and Parallel Computing..... | 21   |
| 2.2.3 Distributed or Network Computing .....      | 22   |
| 2.2.4 Volunteer Computing .....                   | 23   |
| 2.2.5 Cluster Computing .....                     | 24   |
| 2.2.6 Grid Computing .....                        | 25   |
| 2.2.7 In-Memory Computing .....                   | 26   |
| 2.2.8 Virtual Computing .....                     | 27   |
| 2.3 New Computing Delivery Models.....            | 28   |
| 2.3.1 Utility Computing .....                     | 28   |
| 2.3.2 Cloud Computing.....                        | 28   |
| 2.4 Latency-Bandwidth Imbalance .....             | 31   |
| 2.4.1 Storage Technology.....                     | 31   |
| 2.4.2 Memory and Network .....                    | 34   |
| 2.5 Computing Trends .....                        | 36   |
| 2.5.1 End of the Free Lunch for Software .....    | 36   |
| 2.5.2 Key Future Trends in Computing .....        | 38   |
| 3 Exponential Data Growth and Big Data .....      | 39   |
| 3.1 Background.....                               | 39   |
| 3.2 What Defines a Big Data Scenario? .....       | 43   |
| 3.2.1 What Constitutes Big Data?.....             | 43   |

|         |   |    |  |   |     |
|---------|---|----|--|---|-----|
| 3.2.2   | Somewhere, Something Incredible is<br>Waiting to be Known ..... | 49 | 4.2.5                                  | Parallel Relational Model vs. Hadoop .....                      | 97  |
| 3.2.3   | The “Pale Blue Dot” Effect .....                                | 54 | 4.2.6                                  | Beyond Hadoop and MapReduce .....                               | 99  |
| 3.3     | Analytics Divide and Big Data .....                             | 55 | 4.3                                    | Massively Parallel Databases .....                              | 102 |
| 3.4     | Data Analysis Gaps .....  | 58 | 4.3.1                                  | Relational Data Warehouses.....                                 | 103 |
| 3.5     | 6 Vs of Data.....   | 61 | 4.3.2                                  | NoSQL Data Stores.....  | 104 |
| 3.5.1   | Value .....   | 61 | 4.4                                    | Real-Time Event Stream Processing .....                         | 108 |
| 3.5.2   | Volume .....  | 62 | 4.4.1                                  | Requirements of Real-Time Stream Processing.....                | 108 |
| 3.5.3   | Variety .....   | 63 | 4.4.2                                  | StreamSQL .....   | 111 |
| 3.5.4   | Velocity .....  | 65 | 4.5                                    | In-Memory Distributed Analytics (IMDA) .....                    | 113 |
| 3.5.5   | Validity (or Veracity) .....                                    | 65 | 4.6                                    | Big Data Analytics Appliances.....                              | 116 |
| 3.5.6   | Volatility.....   | 67 | 4.7                                    | Big Data Analytics in the Cloud.....                            | 118 |
| 3.6     | The Impact of the 6 Vs.....                                     | 68 | 4.8                                    | Open-Source Analytics Software .....                            | 120 |
| 3.7     | Big Data Management and Governance.....                         | 72 | 4.9                                    | Big Data Technologies’ Adoption Life Cycle .....                | 122 |
| 3.8     | Economic Value, Privacy, and Security.....                      | 74 | 5                                      | High-Performance Data Mining .....                              | 125 |
| 3.8.1   | Economic Impact .....   | 74 | 5.1                                    | Background.....   | 125 |
| 3.8.2   | Privacy Issues.....   | 75 | 5.2                                    | Data Mining Practice .....                                      | 128 |
| 3.8.3   | Security Issues .....   | 76 | 5.2.1                                  | Analytics Sophistication .....                                  | 128 |
| 4       | Big Data Analytics Revolution .....                             | 78 | 5.2.2                                  | Data Mining Methodology.....                                    | 130 |
| 4.1     | Major Developments in Big Data Analytics .....                  | 78 | 5.2.2.1                                | Business Problem  |     |
| 4.2     | Hadoop, MapReduce, and YARN .....                               | 82 | Understanding (Business-Focused) ..... | 132   |     |
| 4.2.1   | Apache Hadoop .....   | 82 | 5.2.2.2                                | Data Understanding and  |     |
| 4.2.1.1 | HDFS in a Nutshell .....  | 84 | Preparation (Data-Focused) .....       | 133   |     |
| 4.2.1.2 | MapReduce in a Nutshell .....                                   | 85 | 5.2.2.3                                | Model Development and   |     |
| 4.2.1.3 | More on Hadoop .....  | 86 | Assessment (Analysis-Focused).....     | 136   |     |
| 4.2.2   | Hadoop 2.0 and YARN.....  | 88 | 5.2.2.4                                | Deployment and Monitoring                                       |     |
| 4.2.3   | Hadoop Ecosystem .....  | 89 | (Operation-Focused) .....              | 142   |     |
| 4.2.3.1 | Databases .....   | 90 | 5.2.3                                  | Data Governance and Model Management .....                      | 143 |
| 4.2.3.2 | High-Level Interfaces .....                                     | 91 | 5.3                                    | Achieving High Performance and Scalability in Data Mining...146 |     |
| 4.2.3.3 | Metadata .....  | 92 | 5.3.1                                  | Chunking or Data Partitioning .....                             | 147 |
| 4.2.3.4 | Business Intelligence and Visualization .....                   | 92 | 5.3.2                                  | Statistical Query Model .....                                   | 148 |
| 4.2.3.5 | Machine Learning.....   | 93 | 5.4                                    | Evolution of Computing Architectures in Data Mining .....       | 151 |
| 4.2.3.6 | Unstructured Data.....  | 93 | 5.4.1                                  | Serial Computing Environment (SC).....                          | 151 |
| 4.2.3.7 | Data Collection .....   | 94 | 5.4.2                                  | Multiprocessor Computing Architecture (SMP) .....               | 152 |
| 4.2.3.8 | Management Components.....                                      | 95 | 5.4.3                                  | Cluster Computing with Shared Storage (CCSS).....               | 155 |
| 4.2.4   | Challenges .....  | 96 | 5.4.4                                  | Shared-Nothing Distributed Computing                            |     |
|         |   |    | Architecture (SN) .....                | 156   |     |

|         |  |     |         |   |     |
|---------|--|-----|---------|---|-----|
| 5.4.5   | Shared-Nothing In-Memory Distributed Computing (SNIM) .....    | 157 | 7.1.1   | Unstructured Data .....                             | 207 |
| 5.4.6   | Accelerated Processing Units.....                              | 159 | 7.2     | Example Applications .....                          | 210 |
| 5.4.6.1 | Graphical Processing Unit (GPU) .....                          | 159 | 7.2.1   | Online Product Recommendation .....                 | 211 |
| 5.4.6.2 | Field-Programmable Gate Array (FPGA).....                      | 160 | 7.2.2   | High-Performance Optimization for Marketing Use ... | 213 |
| 5.4.6.3 | Digital Signal Processor (DSP).....                            | 161 | 7.2.3   | Predicting Component Failure Using Sensor Data .... | 213 |
| 5.5     | High-Performance Data Mining vs. Big Data Analytics .....      | 162 | 7.2.4   | Clickstream Analysis .....                          | 214 |
| 5.6     | HPDM in Classic Organizations .....                            | 165 | 7.2.5   | GPS Data Analysis .....                             | 214 |
| 5.7     | Major Recent Trends in HPDM .....                              | 169 | 7.2.6   | Presidential Election .....                         | 215 |
| 5.7.1   | The Emergence of In-Database Analytics Processing ...          | 171 | 7.2.7   | Telematics in Insurance .....                       | 216 |
| 5.7.1.1 | In-Database Data Preparation .....                             | 172 | 7.2.8   | Portfolio Credit Risk .....                         | 217 |
| 5.7.1.2 | In-database Basic Analysis.....                                | 174 | 7.3     | Big Data Mining Considerations .....                | 219 |
| 5.7.1.3 | In-database Modeling and Advanced Analysis... ..               | 174 | 7.3.1   | Data Preparation.....                               | 219 |
| 5.7.1.4 | In-database Scoring .....                                      | 175 | 7.3.1.1 | Definition of “All the Data” .....                  | 220 |
| 5.7.1.5 | In-database Considerations.....                                | 177 | 7.3.1.2 | A Dream Come True .....                             | 221 |
| 5.7.2   | Cluster or GRID Analytics Processing.....                      | 178 | 7.3.2   | Generalization and Learner Complexity .....         | 224 |
| 5.7.3   | Big Data Mining .....  | 178 | 7.3.2.1 | Structured Risk Minimization .....                  | 224 |
| 6       | Big Data Analytics Applications .....                          | 179 | 7.3.2.2 | Important Observations .....                        | 227 |
| 6.1     | Background.....  | 179 | 7.3.2.3 | Linear Learners Revisited .....                     | 230 |
| 6.2     | General Applications .....                                     | 183 | 7.3.3   | Learning Algorithms .....                           | 230 |
| 6.3     | Data Storage and Archiving .....                               | 185 | 7.3.3.1 | Learner Time Complexity .....                       | 232 |
| 6.3.1   | Big Data Value Assessment.....                                 | 185 | 7.3.4   | Curse of Dimensionality.....                        | 235 |
| 6.3.1.1 | Value per Unit of Storage .....                                | 187 | 7.3.5   | The More Data, The Better .....                     | 237 |
| 6.3.1.2 | Less Expensive Storage.....                                    | 188 | 7.3.6   | Long Tail.....                                      | 240 |
| 6.3.2   | Cold Data Storage, Active Archiving, and Access .....          | 189 | 7.3.7   | Sampling Revisited, and Some Practical Advice.....  | 243 |
| 6.4     | ETL and Data Preparation .....                                 | 192 | 7.3.8   | General Principles.....                             | 250 |
| 6.5     | Extracting Simple Events, Anomalies, or Patterns .....         | 196 | 7.3.9   | Human Aspects.....                                  | 252 |
| 6.6     | Active Data Analysis, Big Data Mining, and Visualization ..... | 197 | 8       | Evolution of Analytics Environments .....           | 253 |
| 6.6.1   | comScore .....   | 198 | 8.1     | Background.....                                     | 253 |
| 6.6.2   | Walmart.....   | 199 | 8.1.1   | Adoption of Big Data Technologies.....              | 254 |
| 6.6.3   | Chevron .....  | 200 | 8.1.2   | Adoption of Hadoop .....                            | 255 |
| 6.6.4   | Sears .....  | 200 | 8.2     | Classic Analytics Environments.....                 | 257 |
| 6.6.5   | eBay/PayPal .....  | 201 | 8.3     | Modern Analytics Environments .....                 | 259 |
| 6.6.6   | JP Morgan Chase .....  | 201 | 8.4     | Data Science .....                                  | 266 |
| 6.6.7   | Other Applications.....  | 202 | 8.4.1   | Definition.....                                     | 270 |
| 7       | HPDM Applications and Considerations .....                     | 204 | 8.5     | Big Data Trends.....                                | 272 |
| 7.1     | Background.....  | 204 | 9       | Bibliography.....                                   | 273 |