Big Data Science & Analytics

A Hands-On Approach

Arshdeep Bahga · Vijay Madisetti

Contents

I BIG DATA ANALYTICS CONCEPTS 19

1 Introduction to Big Data 21

1.1 What is Analytics? 22
1.1.1 Descriptive Analytics 22
1.1.2 Diagnostic Analytics 24
1.1.3 Predictive Analytics 24
1.1.4 Prescriptive Analytics 24

1.2 What is Big Data? 25

1.3 Characteristics of Big Data 26
1.3.1 Volume 26
1.3.2 Velocity 26
1.3.3 Variety 26
1.3.4 Veracity 27
1.3.5 Value 27

1.4 Domain Specific Examples of Big Data 27
1.4.1 Web 27
1.4.2 Financial 29
1.4.3 Healthcare 29
1.4.4 Internet of Things 30
1.4.5 Environment 31
1.4.6 Logistics & Transportation 32
1.4.7 Industry 34
1.4.8 Retail 35

1.5 Analytics Flow for Big Data 35
1.5.1 Data Collection . . . 36
1.5.2 Data Preparation . . . 36
1.5.3 Analysis Types . . . 36
1.5.4 Analysis Modes . . . 36
1.5.5 Visualizations . . . 38
1.6 Big Data Stack 38
1.6.1 Raw Data Sources . . . 39
1.6.2 Data Access Connectors . . . 39
1.6.3 Data Storage . . . 41
1.6.4 Batch Analytics . . . 41
1.6.5 Real-time Analytics . . . 42
1.6.6 Interactive Querying . . . 42
1.6.7 Serving Databases, Web & Visualization Frameworks . . . 42
1.7 Mapping Analytics Flow to Big Data Stack 43
1.8 Case Study: Genome Data Analysis 46
1.9 Case Study: Weather Data Analysis 52
1.10 Analytics Patterns 55

2 Setting up Big Data Stack . . . 63
2.1 Hortonworks Data Platform (HDP) 64
2.2 Cloudera CDH Stack 76
2.3 Amazon Elastic MapReduce (EMR) 83
2.4 Azure HDInsight 87

3 Big Data Patterns . . . 89
3.1 Analytics Architecture Components & Design Styles 90
3.1.1 Load Leveling with Queues . . . 90
3.1.2 Load Balancing with Multiple Consumers . . . 90
3.1.3 Leader Election . . . 91
3.1.4 Sharding . . . 92
3.1.5 Consistency, Availability & Partition Tolerance (CAP) . . . 93
3.1.6 Bloom Filter . . . 93
3.1.7 Materialized Views . . . 94
3.1.8 Lambda Architecture . . . 95
3.1.9 Scheduler-Agent-Supervisor . . . 96
3.1.10 Pipes & Filters . . . 97
3.1.11 Web Service . . . 98
3.1.12 Consensus in Distributed Systems . . . 99

3.2 MapReduce Patterns 101
3.2.1 Numerical Summarization . . . 102
3.2.2 Top-N . . . 110
3.2.3 Filter . . . 113
3.2.4 Distinct . . . 115
3.2.5 Binning . . . 117
3.2.6 Inverted Index . . . 119
3.2.7 Sorting . . . 121
3.2.8 Joins . . . 123

4 NoSQL . . . 129

4.1 Key-Value Databases 130
4.1.1 Amazon DynamoDB . . . 131

4.2 Document Databases 135
4.2.1 MongoDB . . . 135

4.3 Column Family Databases 139
4.3.1 HBase . . . 139

4.4 Graph Databases 147
4.4.1 Neo4j . . . 147

II BIG DATA ANALYTICS IMPLEMENTATIONS 155

5 Data Acquisition . . . 157

5.1 Data Acquisition Considerations 158
5.1.1 Source Type . . . 158
5.1.2 Velocity . . . 158
5.1.3 Ingestion Mechanism . . . 158

5.2 Publish - Subscribe Messaging Frameworks 159
5.2.1 Apache Kafka . . . 160
5.2.2 Amazon Kinesis . . . 165

5.3 Big Data Collection Systems 167
5.3.1 Apache Flume . . . 167
5.3.2 Apache Sqoop . . . 180
5.3.3 Importing Data with Sqoop . . . 181
5.3.4 Selecting Data to Import . . . 182
5.3.5 Custom Connectors . . . 182
5.3.6 Importing Data to Hive . . . 182
5.3.7 Importing Data to HBase . . . 183
5.3.8 Incremental Imports . . . 183
5.3.9 Importing All Tables . . . 183
5.3.10 Exporting Data with Sqoop . . . 183

5.4 Messaging Queues 184
5.4.1 RabbitMQ . . . 184
5.4.2 ZeroMQ . . . 186
5.4.3 RestMQ . . . 187
5.4.4 Amazon SQS . . . 189
5.5 Custom Connectors 191
5.5.1 REST-based Connectors . . . 191
5.5.2 WebSocket-based Connectors . . . 194
5.5.3 MQTT-based Connectors . . . 195
5.5.4 Amazon IoT . . . 197
5.5.5 Azure IoT Hub . . . 205

6 Big Data Storage . . . 213
6.1 HDFS 214
6.1.1 HDFS Architecture . . . 214
6.1.2 HDFS Usage Examples . . . 218

7 Batch Analysis . . . 221
7.1 Hadoop and MapReduce 222
7.1.1 MapReduce Programming Model . . . 222
7.1.2 Hadoop YARN . . . 222
7.1.3 Hadoop Schedulers . . . 226
7.2 Hadoop - MapReduce Examples 228
7.2.1 Batch Analysis of Sensor Data . . . 228
7.2.2 Batch Analysis of N-Gram Dataset . . . 231
7.2.3 Find top-N words with MapReduce . . . 232
7.3 Pig 233
7.3.1 Loading Data . . . 234
7.3.2 Data Types in Pig . . . 234
7.3.3 Data Filtering & Analysis . . . 235
7.3.4 Storing Results . . . 236
7.3.5 Debugging Operators . . . 236
7.3.6 Pig Examples . . . 238
7.4 Case Study: Batch Analysis of News Articles 238
7.5 Apache Oozie 244
7.5.1 Oozie Workflows for Data Analysis . . . 244
7.6 Apache Spark 252
7.6.1 Spark Operations . . . 253
7.7 Search 257
7.7.1 Apache Solr . . . 257

8 Real-time Analysis 269
8.1 Stream Processing 270
8.1.1 Apache Storm 270
8.2 Storm Case Studies 274
8.2.1 Real-time Twitter Sentiment Analysis 274
8.2.2 Real-time Weather Data Analysis 286
8.3 In-Memory Processing 293
8.3.1 Apache Spark 293
8.4 Spark Case Studies 297
8.4.1 Real-time Sensor Data Analysis 298
8.4.2 Real-Time Parking Sensor Data Analysis for Smart Parking System 299
8.4.3 Real-time Twitter Sentiment Analysis 305
8.4.4 Windowed Analysis of Tweets 311

9 Interactive Querying 313
9.1 Spark SQL 314
9.1.1 Case Study: Interactive Querying of Weather Data 319
9.2 Hive 322
9.3 Amazon Redshift 326
9.4 Google BigQuery 335

10 Serving Databases & Web Frameworks 345
10.1 Relational (SQL) Databases 346
10.1.1 MySQL 347
10.2 Non-Relational (NoSQL) Databases 350
10.2.1 Amazon DynamoDB 351
10.2.2 Cassandra 357
10.2.3 MongoDB 360
10.3 Python Web Application Framework - Django 362
10.3.1 Django Architecture 362
10.3.2 Starting Development with Django 363
10.4 Case Study: Django application for viewing weather data 379

III ADVANCED TOPICS 387

11 Analytics Algorithms 389
11.1 Frameworks 390
11.1.1 Spark MLlib 390

11.1.2 H2O 391
11.2 Clustering **393**
11.2.1 K-Means 393
11.3 Case Study: Song Recommendation System **400**
11.4 Classification & Regression **406**
11.4.1 Performance Evaluation Metrics 407
11.4.2 Naive Bayes 408
11.4.3 Generalized Linear Model 420
11.4.4 Decision Trees 435
11.4.5 Random Forest 438
11.4.6 Gradient Boosting Machine 447
11.4.7 Support Vector Machine 458
11.4.8 Deep Learning 460
11.5 Case Study: Classifying Handwritten Digits **471**
11.5.1 Digit Classification with H2O 471
11.5.2 Digit Classification with Spark 473
11.6 Case Study: Genome Data Analysis (Implementation) **475**
11.7 Recommendation Systems **479**
11.7.1 Alternating Least Squares (ALS) 480
11.7.2 Singular Value Decomposition (SVD) 484
11.7.3 Case Study: Movie Recommendation System 484

12 Data Visualization 497
12.1 Frameworks & Libraries **498**
12.1.1 Lightning 498
12.1.2 Pygal 498
12.1.3 Seaborn 498
12.2 Visualization Examples **499**
12.2.1 Line Chart 499
12.2.2 Scatter Plot 501
12.2.3 Bar Chart 504
12.2.4 Box Plot 506
12.2.5 Pie Chart 508
12.2.6 Dot Chart 509
12.2.7 Map Chart 510
12.2.8 Gauge Chart 512
12.2.9 Radar Chart 513
12.2.10 Matrix Chart 514
12.2.11 Force-directed Graph 516
12.2.12 Spatial Graph 518
12.2.13 Distribution Plot 519
12.2.14 Kernel Density Estimate (KDE) Plot 520

12.2.15 Regression Plot 521
12.2.16 Residual Plot 522
12.2.17 Interaction Plot 523
12.2.18 Violin Plot 524
12.2.19 Strip Plot 525
12.2.20 Point Plot 526
12.2.21 Count Plot 527
12.2.22 Heatmap 528
12.2.23 Clustered Heatmap 529
12.2.24 Joint Plot 530
12.2.25 Pair Grid 532
12.2.26 Facet Grid 533

Bibliography 538

Index 539