

# Big Data Science & Analytics

A Hands-On Approach



Arshdeep Bahga • Vijay Madisetti



# Contents

<b>I</b>	<b>BIG DATA ANALYTICS CONCEPTS</b>	<b>19</b>
<b>1</b>	<b>Introduction to Big Data</b>	<b>21</b>
<b>1.1</b>	<b>What is Analytics?</b>	<b>22</b>
1.1.1	Descriptive Analytics	22
1.1.2	Diagnostic Analytics	24
1.1.3	Predictive Analytics	24
1.1.4	Prescriptive Analytics	24
<b>1.2</b>	<b>What is Big Data?</b>	<b>25</b>
<b>1.3</b>	<b>Characteristics of Big Data</b>	<b>26</b>
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
<b>1.4</b>	<b>Domain Specific Examples of Big Data</b>	<b>27</b>
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

<b>1.5</b>	<b>Analytics Flow for Big Data</b>	<b>35</b>
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
<b>1.6</b>	<b>Big Data Stack</b>	<b>38</b>
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
<b>1.7</b>	<b>Mapping Analytics Flow to Big Data Stack</b>	<b>43</b>
<b>1.8</b>	<b>Case Study: Genome Data Analysis</b>	<b>46</b>
<b>1.9</b>	<b>Case Study: Weather Data Analysis</b>	<b>52</b>
<b>1.10</b>	<b>Analytics Patterns</b>	<b>55</b>
<b>2</b>	<b>Setting up Big Data Stack</b>	<b>63</b>
<b>2.1</b>	<b>Hortonworks Data Platform (HDP)</b>	<b>64</b>
<b>2.2</b>	<b>Cloudera CDH Stack</b>	<b>76</b>
<b>2.3</b>	<b>Amazon Elastic MapReduce (EMR)</b>	<b>83</b>
<b>2.4</b>	<b>Azure HDInsight</b>	<b>87</b>
<b>3</b>	<b>Big Data Patterns</b>	<b>89</b>
<b>3.1</b>	<b>Analytics Architecture Components &amp; Design Styles</b>	<b>90</b>
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

<b>3.2</b>	<b>MapReduce Patterns</b>	<b>101</b>
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
<b>4</b>	<b>NoSQL</b>	<b>129</b>
<b>4.1</b>	<b>Key-Value Databases</b>	<b>130</b>
4.1.1	Amazon DynamoDB	131
<b>4.2</b>	<b>Document Databases</b>	<b>135</b>
4.2.1	MongoDB	135
<b>4.3</b>	<b>Column Family Databases</b>	<b>139</b>
4.3.1	HBase	139
<b>4.4</b>	<b>Graph Databases</b>	<b>147</b>
4.4.1	Neo4j	147
<b>II</b>	<b>BIG DATA ANALYTICS IMPLEMENTATIONS</b>	<b>155</b>
<b>5</b>	<b>Data Acquisition</b>	<b>157</b>
<b>5.1</b>	<b>Data Acquisition Considerations</b>	<b>158</b>
5.1.1	Source Type	158
5.1.2	Velocity	158
5.1.3	Ingestion Mechanism	158
<b>5.2</b>	<b>Publish - Subscribe Messaging Frameworks</b>	<b>159</b>
5.2.1	Apache Kafka	160
5.2.2	Amazon Kinesis	165
<b>5.3</b>	<b>Big Data Collection Systems</b>	<b>167</b>
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

<b>5.4</b>	<b>Messaging Queues</b>	<b>184</b>
5.4.1	RabbitMQ .....	184
5.4.2	ZeroMQ .....	186
5.4.3	RestMQ .....	187
5.4.4	Amazon SQS .....	189
<b>5.5</b>	<b>Custom Connectors</b>	<b>191</b>
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
<b>6</b>	<b>Big Data Storage</b> .....	<b>213</b>
<b>6.1</b>	<b>HDFS</b>	<b>214</b>
6.1.1	HDFS Architecture .....	214
6.1.2	HDFS Usage Examples .....	218
<b>7</b>	<b>Batch Analysis</b> .....	<b>221</b>
<b>7.1</b>	<b>Hadoop and MapReduce</b>	<b>222</b>
7.1.1	MapReduce Programming Model .....	222
7.1.2	Hadoop YARN .....	222
7.1.3	Hadoop Schedulers .....	226
<b>7.2</b>	<b>Hadoop - MapReduce Examples</b>	<b>228</b>
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
<b>7.3</b>	<b>Pig</b>	<b>233</b>
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
<b>7.4</b>	<b>Case Study: Batch Analysis of News Articles</b>	<b>238</b>
<b>7.5</b>	<b>Apache Oozie</b>	<b>244</b>
7.5.1	Oozie Workflows for Data Analysis .....	244
<b>7.6</b>	<b>Apache Spark</b>	<b>252</b>
7.6.1	Spark Operations .....	253
<b>7.7</b>	<b>Search</b>	<b>257</b>
7.7.1	Apache Solr .....	257

<b>8</b>	<b>Real-time Analysis</b> .....	<b>269</b>
<b>8.1</b>	<b>Stream Processing</b>	<b>270</b>
8.1.1	Apache Storm .....	270
<b>8.2</b>	<b>Storm Case Studies</b>	<b>274</b>
8.2.1	Real-time Twitter Sentiment Analysis .....	274
8.2.2	Real-time Weather Data Analysis .....	286
<b>8.3</b>	<b>In-Memory Processing</b>	<b>293</b>
8.3.1	Apache Spark .....	293
<b>8.4</b>	<b>Spark Case Studies</b>	<b>297</b>
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
<b>9</b>	<b>Interactive Querying</b> .....	<b>313</b>
<b>9.1</b>	<b>Spark SQL</b>	<b>314</b>
9.1.1	Case Study: Interactive Querying of Weather Data .....	319
<b>9.2</b>	<b>Hive</b>	<b>322</b>
<b>9.3</b>	<b>Amazon Redshift</b>	<b>326</b>
<b>9.4</b>	<b>Google BigQuery</b>	<b>335</b>
<b>10</b>	<b>Serving Databases &amp; Web Frameworks</b> .....	<b>345</b>
<b>10.1</b>	<b>Relational (SQL) Databases</b>	<b>346</b>
10.1.1	MySQL .....	347
<b>10.2</b>	<b>Non-Relational (NoSQL) Databases</b>	<b>350</b>
10.2.1	Amazon DynamoDB .....	351
10.2.2	Cassandra .....	357
10.2.3	MongoDB .....	360
<b>10.3</b>	<b>Python Web Application Framework - Django</b>	<b>362</b>
10.3.1	Django Architecture .....	362
10.3.2	Starting Development with Django .....	363
<b>10.4</b>	<b>Case Study: Django application for viewing weather data</b>	<b>379</b>
<b>III</b>	<b>ADVANCED TOPICS</b>	<b>387</b>
<b>11</b>	<b>Analytics Algorithms</b> .....	<b>389</b>
<b>11.1</b>	<b>Frameworks</b>	<b>390</b>
11.1.1	Spark MLlib .....	390

11.1.2	H2O	391
<b>11.2</b>	<b>Clustering</b>	<b>393</b>
11.2.1	K-Means	393
<b>11.3</b>	<b>Case Study: Song Recommendation System</b>	<b>400</b>
<b>11.4</b>	<b>Classification &amp; Regression</b>	<b>406</b>
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
<b>11.5</b>	<b>Case Study: Classifying Handwritten Digits</b>	<b>471</b>
11.5.1	Digit Classification with H2O	471
11.5.2	Digit Classification with Spark	473
<b>11.6</b>	<b>Case Study: Genome Data Analysis (Implementation)</b>	<b>475</b>
<b>11.7</b>	<b>Recommendation Systems</b>	<b>479</b>
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
<b>12</b>	<b>Data Visualization</b>	<b>497</b>
<b>12.1</b>	<b>Frameworks &amp; Libraries</b>	<b>498</b>
12.1.1	Lightning	498
12.1.2	Pygal	498
12.1.3	Seaborn	498
<b>12.2</b>	<b>Visualization Examples</b>	<b>499</b>
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
<b>Bibliography .....</b>	<b>538</b>
<b>Index .....</b>	<b>539</b>