

# Contents

<b>About the Author .....</b>	<b>xi</b>
<b>About the Technical Reviewers .....</b>	<b>xiii</b>
<b>■Chapter 1: The Big (Data) Problem .....</b>	<b>1</b>
<b>Identifying Big Data Symptoms .....</b>	<b>1</b>
Size Matters.....	1
Typical Business Use Cases.....	2
<b>Understanding the Big Data Project's Ecosystem .....</b>	<b>3</b>
Hadoop Distribution .....	3
Data Acquisition.....	6
Processing Language .....	7
Machine Learning .....	10
NoSQL Stores.....	10
<b>Creating the Foundation of a Long-Term Big Data Architecture.....</b>	<b>12</b>
Architecture Overview .....	12
Log Ingestion Application .....	13
Learning Application.....	13
Processing Engine .....	14
Search Engine.....	15
<b>Summary .....</b>	<b>15</b>

<b>Chapter 2: Early Big Data with NoSQL</b> .....	<b>17</b>
NoSQL Landscape .....	17
Key/Value.....	17
Column .....	18
Document .....	18
Graph .....	19
NoSQL in Our Use Case.....	20
Introducing Couchbase.....	21
Architecture .....	22
Cluster Manager and Administration Console .....	24
Managing Documents.....	28
Introducing Elasticsearch .....	30
Architecture .....	30
Monitoring Elasticsearch.....	34
Search with Elasticsearch.....	36
Using NoSQL as a Cache in a SQL-based Architecture .....	38
Caching Document .....	38
ElasticSearch Plug-in for Couchbase with Couchbase XDCR .....	40
ElasticSearch Only .....	40
Summary.....	40
<b>Chapter 3: Defining the Processing Topology</b> .....	<b>41</b>
First Approach to Data Architecture .....	41
A Little Bit of Background.....	41
Dealing with the Data Sources .....	42
Processing the Data.....	45
Splitting the Architecture.....	49
Batch Processing.....	50
Stream Processing .....	52
The Concept of a Lambda Architecture .....	53
Summary.....	55

<b>Chapter 4: Streaming Data .....</b>	<b>57</b>
Streaming Architecture .....	57
Architecture Diagram.....	57
Technologies.....	58
The Anatomy of the Ingested Data .....	60
Clickstream Data .....	60
The Raw Data .....	62
The Log Generator .....	63
Setting Up the Streaming Architecture.....	64
Shipping the Logs in Apache Kafka .....	64
Draining the Logs from Apache Kafka .....	72
Summary.....	79
<b>Chapter 5: Querying and Analyzing Patterns.....</b>	<b>81</b>
Defining an Analytics Strategy .....	81
Continuous Processing.....	81
Real-Time Querying.....	82
Process and Index Data Using Spark .....	82
Preparing the Spark Project .....	82
Understanding a Basic Spark Application.....	84
Implementing the Spark Streamer .....	86
Implementing a Spark Indexer .....	89
Implementing a Spark Data Processing .....	91
Data Analytics with Elasticsearch .....	93
Introduction to the aggregation framework.....	93
Visualize Data in Kibana .....	100
Summary.....	103

<b>Chapter 6: Learning From Your Data? .....</b>	<b>105</b>
Introduction to Machine Learning .....	105
Supervised Learning.....	105
Unsupervised Learning.....	107
Machine Learning with Spark.....	108
Adding Machine Learning to Our Architecture.....	108
<b>Adding Machine Learning to Our Architecture .....</b>	<b>112</b>
Enriching the Clickstream Data .....	112
Labelizing the Data.....	117
Training and Making Prediction.....	119
<b>Summary .....</b>	<b>121</b>
<b>Chapter 7: Governance Considerations .....</b>	<b>123</b>
<b>Dockerizing the Architecture .....</b>	<b>123</b>
Introducing Docker .....	123
Installing Docker.....	125
Creating Your Docker Images .....	125
Composing the Architecture .....	128
<b>Architecture Scalability .....</b>	<b>132</b>
Sizing and Scaling the Architecture.....	132
Monitoring the Infrastructure Using the Elastic Stack.....	135
Considering Security .....	136
<b>Summary .....</b>	<b>137</b>
<b>Index.....</b>	<b>139</b>