



Contents

Introduction		xv
Part I	What Is Big Data?	1
Chapter 1	Industry Needs and Solutions	3
	What's So <i>Big</i> About Big Data?	4
	A Brief History of Hadoop	5
	Google	5
	Nutch	6
	What Is Hadoop?	6
	Derivative Works and Distributions	7
	Hadoop Distributions	8
	Core Hadoop Ecosystem	9
	Important Apache Projects for Hadoop	11
	The Future for Hadoop	17
	Summary	17
Chapter 2	Microsoft's Approach to Big Data	19
	A Story of "Better Together"	19
	Competition in the Ecosystem	20
	SQL on Hadoop Today	21
	Hortonworks and Stinger	21
	Cloudera and Impala	23
	Microsoft's Contribution to SQL in Hadoop	25
	Deploying Hadoop	25
	Deployment Factors	26
	Deployment Topologies	29
	Deployment Scorecard	33
	Summary	36

Part II	Setting Up for Big Data with Microsoft	37
Chapter 3	Configuring Your First Big Data Environment	39
	Getting Started	39
	Getting the Install	40
	Running the Installation	40
	On-Premise Installation: Single-Node Installation	41
	HDInsight Service: Installing in the Cloud	51
	Windows Azure Storage Explorer Options	52
	Validating Your New Cluster	55
	Logging into HDInsight Service	55
	Verify HDP Functionality in the Logs	57
	Common Post-Setup Tasks	58
	Loading Your First Files	58
	Verifying Hive and Pig	60
	Summary	63
Part III	Storing and Managing Big Data	65
Chapter 4	HDFS, Hive, HBase, and HCatalog	67
	Exploring the Hadoop Distributed File System	68
	Explaining the HDFS Architecture	69
	Interacting with HDFS	72
	Exploring Hive: The Hadoop Data Warehouse Platform	75
	Designing, Building, and Loading Tables	76
	Querying Data	77
	Configuring the Hive ODBC Driver	77
	Exploring HCatalog: HDFS Table and Metadata Management	78
	Exploring HBase: An HDFS Column-Oriented Database	80
	Columnar Databases	81
	Defining and Populating an HBase Table	82
	Using Query Operations	83
	Summary	84
Chapter 5	Storing and Managing Data in HDFS	85
	Understanding the Fundamentals of HDFS	86
	HDFS Architecture	87
	NameNodes and DataNodes	89
	Data Replication	90
	Using Common Commands to Interact with HDFS	92
	Interfaces for Working with HDFS	92
	File Manipulation Commands	94
	Administrative Functions in HDFS	97
	Moving and Organizing Data in HDFS	100
	Moving Data in HDFS	100
	Implementing Data Structures for Easier Management	101
	Rebalancing Data	102
	Summary	103

Chapter 6	Adding Structure with Hive	105
	Understanding Hive's Purpose and Role	106
	Providing Structure for Unstructured Data	107
	Enabling Data Access and Transformation	114
	Differentiating Hive from Traditional RDBMS Systems	115
	Working with Hive	116
	Creating and Querying Basic Tables	117
	Creating Databases	117
	Creating Tables	118
	Adding and Deleting Data	121
	Querying a Table	123
	Using Advanced Data Structures with Hive	126
	Setting Up Partitioned Tables	126
	Loading Partitioned Tables	128
	Using Views	129
	Creating Indexes for Tables	130
	Summary	131
Chapter 7	Expanding Your Capability with HBase and HCatalog	133
	Using HBase	134
	Creating HBase Tables	134
	Loading Data into an HBase Table	136
	Performing a Fast Lookup	138
	Loading and Querying HBase	139
	Managing Data with HCatalog	140
	Working with HCatalog and Hive	140
	Defining Data Structures	141
	Creating Indexes	143
	Creating Partitions	143
	Integrating HCatalog with Pig and Hive	145
	Using HBase or Hive as a Data Warehouse	149
	Summary	150
Part IV	Working with Your Big Data	151
Chapter 8	Effective Big Data ETL with SSIS, Pig, and Sqoop	153
	Combining Big Data and SQL Server Tools for Better Solutions	154
	Why Move the Data?	154
	Transferring Data Between Hadoop and SQL Server	155
	Working with SSIS and Hive	156
	Connecting to Hive	157
	Configuring Your Packages	161
	Loading Data into Hadoop	165
	Getting the Best Performance from SSIS	167
	Transferring Data with Sqoop	167
	Copying Data from SQL Server	168
	Copying Data to SQL Server	170

	Using Pig for Data Movement	171
	Transforming Data with Pig	171
	Using Pig and SSIS Together	174
	Choosing the Right Tool	175
	Use Cases for SSIS	175
	Use Cases for Pig	175
	Use Cases for Sqoop	176
	Summary	176
Chapter 9	Data Research and Advanced Data Cleansing with Pig and Hive	177
	Getting to Know Pig	178
	When to Use Pig	178
	Taking Advantage of Built-in Functions	179
	Executing User-defined Functions	180
	Using UDFs	182
	Building Your Own UDFs for Pig	189
	Using Hive	192
	Data Analysis with Hive	192
	Types of Hive Functions	192
	Extending Hive with Map-reduce Scripts	195
	Creating a Custom Map-reduce Script	198
	Creating Your Own UDFs for Hive	199
	Summary	201
Part V	Big Data and SQL Server Together	203
Chapter 10	Data Warehouses and Hadoop Integration	205
	State of the Union	206
	Challenges Faced by Traditional Data Warehouse Architectures	207
	Technical Constraints	207
	Business Challenges	213
	Hadoop's Impact on the Data Warehouse Market	216
	Keep Everything	216
	Code First (Schema Later)	217
	Model the Value	218
	Throw Compute at the Problem	218
	Introducing Parallel Data Warehouse (PDW)	220
	What Is PDW?	221
	Why Is PDW Important?	222
	How PDW Works	224
	Project Polybase	235
	Polybase Architecture	235
	Business Use Cases for Polybase Today	249
	Speculating on the Future for Polybase	251
	Summary	255

Chapter 11	Visualizing Big Data with Microsoft BI	257
	An Ecosystem of Tools	258
	Excel	258
	PowerPivot	258
	Power View	259
	Power Map	261
	Reporting Services	261
	Self-service Big Data with PowerPivot	263
	Setting Up the ODBC Driver	263
	Loading Data	265
	Updating the Model	272
	Adding Measures	273
	Creating Pivot Tables	274
	Rapid Big Data Exploration with Power View	277
	Spatial Exploration with Power Map	281
	Summary	283
Chapter 12	Big Data Analytics	285
	Data Science, Data Mining, and Predictive Analytics	286
	Data Mining	286
	Predictive Analytics	287
	Introduction to Mahout	288
	Building a Recommendation Engine	289
	Getting Started	291
	Running a User-to-user Recommendation Job	292
	Running an Item-to-item Recommendation Job	295
	Summary	296
Chapter 13	Big Data and the Cloud	297
	Defining the Cloud	298
	Exploring Big Data Cloud Providers	299
	Amazon	299
	Microsoft	300
	Setting Up a Big Data Sandbox in the Cloud	300
	Getting Started with Amazon EMR	301
	Getting Started with HDInsight	307
	Storing Your Data in the Cloud	315
	Storing Data	316
	Uploading Your Data	317
	Exploring Big Data Storage Tools	318
	Integrating Cloud Data	319
	Other Cloud Data Sources	321
	Summary	321
Chapter 14	Big Data in the Real World	323
	Common Industry Analytics	324
	Telco	324
	Energy	325

	Retail	325
	Data Services	326
	IT/Hosting Optimization	326
	Marketing Social Sentiment	327
	Operational Analytics	327
	Failing Fast	328
	A New Ecosystem of Technologies	328
	User Audiences	330
	Summary	333
Part VI	Moving Your Big Data Forward	335
Chapter 15	Building and Executing Your Big Data Plan	337
	Gaining Sponsor and Stakeholder Buy-In	338
	Problem Definition	338
	Scope Management	339
	Stakeholder Expectations	341
	Defining the Criteria for Success	342
	Identifying Technical Challenges	342
	Environmental Challenges	342
	Challenges in Skillset	344
	Identifying Operational Challenges	345
	Planning for Setup/Configuration	345
	Planning for Ongoing Maintenance	347
	Going Forward	348
	The HandOff to Operations	348
	After Deployment	349
	Summary	350
Chapter 16	Operational Big Data Management	351
	Hybrid Big Data Environments: Cloud and On-Premise Solutions Working Together	352
	Ongoing Data Integration with Cloud and On-Premise Solutions	353
	Integration Thoughts for Big Data	354
	Backups and High Availability in Your Big Data Environment	356
	High Availability	356
	Disaster Recovery	358
	Big Data Solution Governance	359
	Creating Operational Analytics	360
	System Center Operations Manager for HDP	361
	Installing the Ambari SCOM Management Pack	362
	Monitoring with the Ambari SCOM Management Pack	371
	Summary	377
Index		379