



Table of Contents

Introduction	xxv
Book Preview.....	xxxii
Chapter 1: Getting an Overview of Big Data	1
What is Big Data?.....	2
History of Data Management – Evolution of Big Data	5
Structuring Big Data	7
Types of Data.....	7
Elements of Big Data	12
Volume	13
Velocity.....	13
Variety	14
Veracity	14
Big Data Analytics	14
Advantages of Big Data Analytics.....	15
Careers in Big Data	18
Skills Required	20
Future of Big Data.....	20
Summary	23
Quick Revise.....	23
Multiple-Choice Questions.....	23
Subjective Questions.....	24
Chapter 2: Exploring the Use of Big Data in Business Context	27
Use of Big Data in Social Networking.....	28
Business Intelligence.....	30
Marketing.....	32

Table of Contents

Product Design and Development	33
Use of Big Data in Preventing Fraudulent Activities.....	34
Preventing Fraud Using Big Data Analytics	35
Use of Big Data in Detecting Fraudulent Activities in Insurance Sector	36
Fraud Detection Methods	37
Use of Big Data in Retail Industry	41
Use of RFID Data in Retail.....	42
Summary	44
Quick Revise.....	44
Multiple-Choice Questions.....	44
Subjective Questions.....	46
Chapter 3: Introducing Technologies for Handling Big Data	51
Distributed and Parallel Computing for Big Data.....	52
Introducing Hadoop.....	56
How does Hadoop Function?	57
Cloud Computing and Big Data	59
Features of Cloud Computing.....	60
Cloud Deployment Models	61
Cloud Delivery Models.....	64
Cloud Services for Big Data.....	66
Cloud Providers in Big Data Market.....	67
In-Memory Computing Technology for Big Data	68
Summary	69
Quick Revise.....	70
Multiple-Choice Questions.....	70
Subjective Questions.....	71
Chapter 4: Understanding Hadoop Ecosystem	77
Hadoop Ecosystem.....	78
Hadoop Distributed File System.....	82
HDFS Architecture	82
Features of HDFS.....	84
MapReduce.....	86
Features of MapReduce.....	87
Hadoop YARN.....	87

HBase.....	89
Features of HBase	89
Hive	90
Pig and Pig Latin.....	91
Sqoop.....	91
ZooKeeper.....	92
Flume.....	93
Oozie.....	93
Summary.....	96
Quick Revise.....	96
Multiple-Choice Questions.....	96
Subjective Questions.....	97
Chapter 5: Understanding MapReduce Fundamentals and HBase.....	101
The MapReduce Framework.....	102
Exploring the Features of MapReduce.....	103
Working of MapReduce.....	104
Exploring Map and Reduce Functions.....	108
Techniques to Optimize MapReduce Jobs.....	111
Hardware/Network Topology	111
Synchronization	112
File System.....	112
Uses of MapReduce.....	113
Role of HBase in Big Data Processing.....	114
Characteristics of HBase	114
Installation of HBase.....	115
Summary.....	122
Quick Revise.....	123
Multiple-Choice Questions.....	123
Subjective Questions.....	125
Chapter 6: Understanding Big Data Technology Foundations	129
Exploring the Big Data Stack.....	130
Data Sources Layer	131
Ingestion Layer.....	132
Storage Layer.....	133

Table of Contents

Physical Infrastructure Layer	134
Platform Management Layer.....	136
Security Layer.....	137
Monitoring Layer.....	138
Analytics Engine	138
Visualization Layer.....	139
Big Data Applications.....	140
Virtualization and Big Data.....	141
Virtualization Approaches	143
Server Virtualization	143
Application Virtualization.....	144
Network Virtualization.....	144
Processor and Memory Virtualization.....	144
Data and Storage Virtualization.....	144
Managing Virtualization with Hypervisor.....	145
Implementing Virtualization to Work with Big Data	146
Summary	146
Quick Revise.....	146
Multiple-Choice Questions.....	146
Subjective Questions.....	148
Chapter 7: Storing Data in Databases and Data Warehouses.....	153
RDBMS and Big Data	154
Issues with the Relational Model.....	160
Non-Relational Database	161
Issues with the Non-Relational Model.....	162
Polyglot Persistence.....	163
Integrating Big Data with Traditional Data Warehouses	164
Big Data Analysis and Data Warehouse.....	167
Changing Deployment Models in Big Data Era	171
Summary.....	172
Quick Revise.....	172
Multiple-Choice Questions.....	172
Subjective Questions.....	173

Chapter 8: Storing Data in Hadoop.....	175
Introducing HDFS.....	176
HDFS Architecture	177
Using HDFS Files.....	179
Hadoop-Specific File System Types	181
HDFS Commands.....	182
The org.apache.hadoop.io package	183
HDF	185
HDFS High Availability.....	186
Introducing HBase.....	189
HBase Architecture.....	189
Storing Big Data with HBase.....	191
Interacting with the Hadoop Ecosystem.....	191
HBase in Operation – Programming with HBase.....	191
Installation	192
Combining HBase and HDFS.....	193
Selecting the Suitable Hadoop Data Organization for Applications	199
Summary.....	200
Quick Revise.....	201
Multiple-Choice Questions.....	201
Subjective Questions.....	201
Chapter 9: Processing Your Data with MapReduce.....	205
Recollecting the Concept of MapReduce Framework.....	206
Developing Simple MapReduce Application.....	207
Building the Application.....	213
Executing the Application	214
Points to Consider while Designing MapReduce.....	218
Summary	220
Quick Revise.....	220
Multiple-Choice Questions.....	220
Subjective Questions.....	221
Chapter 10: Customizing MapReduce Execution	223
Controlling MapReduce Execution with InputFormat.....	224
InputSplit	224

Table of Contents

RecordReader	225
FileInputFormat	226
Implementing InputFormat for Compute-Intensive Applications	227
Implementing InputFormat to control the Number of Maps	230
Implementing InputFormat for Multiple HBase Tables	231
Reading Data with Custom RecordReader	234
Organizing Output Data with OutputFormats.....	237
Customizing Data with RecordWriter	238
Optimizing MapReduce Execution with Combiner.....	239
Controlling Reducer Execution with Partitioners	240
Implementing a MapReduce Program for Sorting Text Data	240
Summary	257
Quick Revise.....	257
Multiple-Choice Questions.....	257
Subjective Questions.....	259
Chapter 11: Testing and Debugging MapReduce Applications	261
Performing Unit Testing for MapReduce Applications.....	262
Unit Testing the Mapper Component.....	264
Unit Testing the Reducer Component	265
Integration Testing of the Mapper-Reducer Combination.....	266
Performing Local Application Testing with Eclipse	268
Logging for Hadoop Testing.....	270
Application Log Processing.....	271
Defensive Programming in MapReduce.....	273
Summary.....	274
Quick Revise.....	274
Multiple-Choice Questions.....	274
Subjective Questions.....	276
Chapter 12: Understanding Hadoop YARN Architecture.....	279
Background of YARN.....	280
Revisiting MapReduce	281
Limitations of MapReduce	282
Advantages of YARN.....	283
YARN Architecture	284

ResourceManager	285
ApplicationManager.....	285
Integration of ResourceManager and ApplicationManager	286
Working of YARN	287
YARN Schedulers	288
CapacityScheduler	289
FairScheduler.....	291
Backward Compatibility with YARN	294
Script Compatibility	294
Binary Compatibility.....	295
Source Compatibility.....	295
YARN Configurations.....	295
YARN Commands.....	298
Administration Commands.....	299
User Commands.....	300
Log Management in Hadoop 1	300
Log Management in YARN.....	301
Summary.....	302
Quick Revise.....	303
Multiple-Choice Questions.....	303
Subjective Questions.....	303
Chapter 13: Exploring Hive.....	307
Introducing Hive.....	308
Getting Started with Hive.....	310
Hive Variables.....	312
Hive Properties.....	312
Hive Queries.....	313
Data Types in Hive	313
Built-In Functions in Hive	314
Hive DDL.....	316
Creating Databases	317
Viewing a Database.....	317
Dropping a Database.....	317
Altering Databases.....	318
Creating Tables.....	318

Table of Contents

Creating a Table Using the Existing Schema.....	319
Dropping Tables.....	319
Altering Tables.....	319
Using Hive DDL Statements.....	320
Data Manipulation in Hive.....	322
Loading Files into Tables.....	322
Inserting Data into Tables.....	323
Update in Hive.....	325
Delete in Hive.....	325
Using Hive DML Statements.....	325
Data Retrieval Queries.....	327
Using the SELECT Command.....	327
Using the WHERE Clause.....	327
Using the GROUP BY Clause.....	327
Using the HAVING Clause.....	328
Using the LIMIT Clause.....	328
Executing HiveQL Queries.....	328
Using JOINS in Hive.....	330
Inner Joins.....	330
Outer Joins.....	331
Cartesian Product Joins.....	333
Map-Side Joins.....	333
Joining Tables.....	334
Summary.....	338
Quick Revise.....	338
Multiple-Choice Questions.....	338
Subjective Questions.....	339
Chapter 14: Analyzing Data with Pig.....	341
Introducing Pig.....	342
The Pig Architecture.....	342
Benefits of Pig.....	343
Installing Pig.....	343
Properties of Pig.....	344
Running Pig.....	346
Running Pig Programs.....	347

Getting Started with Pig Latin	349
Pig Latin Structure	350
Pig Latin Application Flow.....	351
Working with Operators in Pig.....	352
FOREACH	352
ASSERT	353
FILTER.....	353
GROUP	354
ORDER BY	355
DISTINCT	356
JOIN	357
LIMIT.....	358
SAMPLE.....	358
SPLIT	358
FLATTEN.....	359
Working with Functions in Pig.....	360
Summary	364
Quick Revise.....	364
Multiple-Choice Questions.....	364
Subjective Questions.....	365
Chapter 15: Using Oozie	369
Introducing Oozie.....	370
Main Functional Components of Oozie.....	371
Benefits of Oozie	372
Installing and Configuring Oozie.....	373
Understanding the Oozie Workflow.....	379
Execution of Asynchronous Actions in Oozie	382
Implementing the Oozie Workflow.....	383
Oozie Recovery Capabilities.....	386
Oozie Workflow Life Cycle	387
Oozie Coordinator	388
Types of Oozie Coordinator	388
Oozie Coordinator Lifecycle Operations	391
Oozie Bundle	392
Oozie Parameterization with EL.....	394

Table of Contents

Workflow Functions	395
Coordinator Functions	395
Bundle Functions	395
EL Functions	395
Oozie Job Execution Model	396
Accessing Oozie	399
Oozie SLA	399
Event Status	400
SLA Status.....	400
Oozie Activity.....	400
The Oozie SLA Subsystem.....	400
SLA Language Schema.....	401
Summary.....	402
Quick Revise.....	403
Multiple-Choice Questions.....	403
Subjective Questions.....	404
Chapter 16: NoSQL Data Management.....	407
Introduction to NoSQL	408
Characteristics of NoSQL.....	409
Evolution of Databases.....	410
Aggregate Data Models	412
Key Value Data Model	413
Document Databases	415
Relationships	416
Graph Databases.....	417
Schema-Less Databases.....	418
Materialized Views.....	419
Distribution Models.....	420
CAP Theorem.....	420
Sharding.....	421
MapReduce Partitioning and Combining.....	422
Composing MapReduce Calculations.....	424
Summary.....	427
Quick Revise.....	427
Multiple-Choice Questions.....	427
Subjective Questions.....	428

Chapter 17: Understanding Analytics and Big Data	431
Comparing Reporting and Analysis.....	432
Reporting	433
Analysis.....	434
The Analytic Process	436
Types of Analytics	437
Basic Analytics	437
Advanced Analytics	438
Operationalized Analytics.....	439
Monetized Analytics.....	439
Characteristics of Big Data Analysis	439
Points to Consider during Analysis	440
Frame the Problem Correctly	440
Statistical Significance or Business Importance?	441
Making Inferences versus Computing Statistics.....	443
Developing an Analytic Team.....	444
Skills Required for an Analyst.....	445
Convergence of IT and Analytics.....	445
Understanding Text Analytics	447
Summary.....	448
Quick Revise.....	448
Multiple-Choice Questions.....	448
Subjective Questions.....	449
 Chapter 18: Analytical Approaches and Tools to Analyze Data.....	 455
Analytical Approaches.....	456
Ensemble Methods.....	456
Text Data Analysis.....	457
History of Analytical Tools.....	459
Graphical User Interfaces.....	459
Point Solutions	460
Data Visualization Tools.....	460
Introducing Popular Analytical Tools.....	462
The R Project for Statistical Computing.....	462
IBM SPSS.....	463
SAS.....	464

Table of Contents

Comparing Various Analytical Tools.....	466
Installing R.....	469
Installing R on a Windows Computer.....	469
Installing R on a Macintosh Computer	475
Installing R on a Linux Computer	477
Installing RStudio on Windows.....	479
Installing RStudio on Linux.....	482
Summary.....	484
Quick Revise.....	484
Multiple-Choice Questions.....	484
Subjective Questions.....	486
Chapter 19: Exploring R.....	489
Exploring Basic Features of R.....	490
Statistical Features	491
Programming Features.....	492
Packages.....	492
Graphical User Interfaces.....	492
Exploring RGui	493
R Console	494
Developing a Program	495
Quitting R.....	496
Exploring RStudio.....	496
Handling Basic Expressions in R	500
Basic Arithmetic in R.....	500
Mathematical Operators	502
Variables in R.....	503
Calling Functions in R.....	504
Working with Vectors	504
Storing and Calculating Values in R	506
Creating and Using Objects.....	509
Interacting with Users.....	510
Handling Data in R Workspace	511
The ls() Function	511
The rm() Function.....	512
The getwd() Function.....	513

The save() Function	513
The load () Function.....	514
Executing Scripts.....	515
Creating Plots	518
Accessing Help and Documentation in R.....	519
Using Built-in Datasets in R.....	519
Summary	523
Quick Revise.....	523
Multiple-Choice Questions.....	523
Subjective Questions.....	525
Chapter 20: Reading Datasets and Exporting Data from R	527
Using the c() Command	528
Reading and Combining Numerical Data.....	528
Reading and Combining Text Data	533
Reading Both Numeric and Text Values in R	534
Using the scan() Command	535
Reading the Text Data Using the scan() Command	536
Using Clipboard to Create the Data	538
Reading the Data of a File from Disk	540
Reading Multiple Data Values from Large Files	543
Using the read.csv() Command	544
Using the read.table() Command.....	545
Reading Data from R Studio.....	546
Exporting Data from R.....	549
Using the write.table() Command	549
Using the write.csv() Command.....	553
Summary	554
Quick Revise.....	554
Multiple-Choice Questions.....	554
Subjective Questions.....	556
Chapter 21: Manipulating and Processing Data in R	557
Selecting the Most Appropriate Data Structure.....	558
Creating Data Subsets	559
Creating Subsets in Vectors	559

Table of Contents

Creating Subsets in Data Frames	561
Merging Datasets in R.....	563
Using the merge() Function.....	564
Using the cbind Function.....	567
Using the rbind() Function	567
Sorting Data	568
Sorting Data	569
Ordering Data.....	569
Reverse Sort.....	570
Putting Your Data into Shape	572
Transposing Data.....	572
Converting Data to Wide or Long Formats.....	573
Melting Data to Long Format.....	573
Casting Data to Wide Format.....	575
Managing Data in R Using Matrices	576
Reshaping a Vector into a Matrix	576
Accessing Matrix and Subsetting the Data.....	577
Managing Data in R Using Data Frames	578
Creating Data Frames.....	578
Accessing Data Frames	579
Merging Data Frames.....	580
Performing Operations on Data Frames.....	583
Summary.....	583
Quick Revise.....	584
Multiple-Choice Questions.....	584
Subjective Questions.....	585
Chapter 22: Working with Functions and Packages in R.....	587
Using Functions Instead of Scripts.....	588
Transforming an R Script into a Function	588
Returning Results in R.....	591
Reducing the Number of Lines in an R Function	593
Assigning the Function Objects	594
Writing Function Without Braces.....	595
Using Arguments in Functions.....	596
Using Dot Argument in Function.....	598

Passing Functions as Arguments.....	599
Anonymous Functions.....	599
Local and Global Environment of Functions.....	600
Built-in Functions in R.....	601
Numeric Functions.....	602
Character Functions.....	602
Statistical Probability Functions.....	603
Miscellaneous Functions.....	604
Introducing Packages.....	604
Working with Packages.....	605
Summary.....	610
Quick Revise.....	610
Multiple-Choice Questions.....	610
Subjective Questions.....	611
Chapter 23: Performing Graphical Analysis in R.....	613
Using Plots.....	614
Using Plots for a Single Variable.....	615
Using Plots for Two Variables.....	632
Using Plots for Multiple Variables.....	636
Designing Special Plots.....	637
Saving Graphs to External Files.....	639
Summary.....	639
Quick Revise.....	640
Multiple-Choice Questions.....	640
Subjective Questions.....	640
Chapter 24: Integrating R and Hadoop and Understanding Hive.....	643
RHadoop—An Integration of R and Hadoop.....	644
Installing RHadoop.....	645
Using RHadoop.....	653
Text Mining in RHadoop.....	654
Data Analysis Using the MapReduce Technique in RHadoop.....	657
Data Mining in Hive.....	659
Summary.....	660
Quick Revise.....	660

Table of Contents

Multiple-Choice Questions.....	660
Subjective Questions.....	661
Chapter 25: Data Visualization-I	663
Introducing Data Visualization.....	664
Techniques Used for Visual Data Representation.....	665
Types of Data Visualization	670
Applications of Data Visualization.....	672
Visualizing Big Data.....	672
Deriving Business Solutions.....	673
Turning Data into Information	673
Tools Used in Data Visualization	674
Proprietary Data Visualization Tools.....	678
Open-Source Data Visualization Tools.....	678
Analytical Techniques Used in Big Data Visualization	679
Tableau Products	679
Installation of Tableau Public.....	680
Summary.....	685
Quick Revise.....	686
Multiple-Choice Questions.....	686
Subjective Questions.....	686
Chapter 26: Data Visualization with Tableau (Data Visualization-II)	689
Introduction to Tableau Software.....	690
Tableau Desktop Workspace.....	695
Operations on Data.....	705
Data Analytics in Tableau Public.....	705
Using Visual Controls in Tableau Public.....	715
Summary.....	721
Quick Revise.....	721
Multiple-Choice Questions.....	721
Subjective Questions.....	722
Chapter 27: Social Media Analytics and Text Mining	725
Introducing Social Media.....	726
Introducing Key Elements of Social Media	728
Introducing Text Mining.....	729

Understanding Text Mining Process.....	732
Sentiment Analysis.....	734
Performing Social Media Analytics and Opinion Mining on Tweets.....	735
Online Social Media Analysis	739
Summary.....	742
Quick Revise.....	742
Multiple-Choice Questions.....	742
Subjective Questions.....	743
Chapter 28: Mobile Analytics.....	747
Introducing Mobile Analytics	748
Define Mobile Analytics.....	750
Mobile Analytics and Web Analytics.....	751
Types of Results from Mobile Analytics.....	751
Types of Applications for Mobile Analytics.....	752
Introducing Mobile Analytics Tools.....	754
Location-based Tracking Tools	756
Real-time Analytics Tools.....	756
User Behavior Tracking Tools.....	758
Performing Mobile Analytics	759
Challenges of Mobile Analytics	774
Summary.....	775
Quick Revise.....	775
Multiple-Choice Questions.....	775
Subjective Questions.....	776
Chapter 29: Finding a Job in the Big Data Market	779
Importance and Scope of Big Data Jobs.....	781
Big Data Opportunities	783
Skill Assessment for Big Data Jobs	784
Roles and Responsibilities in Big Data Jobs	789
Business Analyst for Big Data.....	789
Big Data Scientist	790
Software Developer for Big Data	791
Gaining a Foothold in the Big Data Market	791
Take Your Time.....	792

Table of Contents

Preparing the Big Data Skill Learning and Testing Mechanism.....	793
Basic Educational Requirements for Big Data Jobs	793
Basic Technological Requirements for Big Data Jobs.....	795
Tools Supporting Big Data.....	796
Consultants and In-House Specialists in Big Data	797
Big Data Consultants.....	797
In-House Big Data Experts	799
Tactics for Searching Big Data Jobs	800
Network Building Tactics.....	800
Resume Building Tactics.....	801
Preparing for Interviews.....	802
Obtaining Big Data Jobs through Social Media	804
Summary.....	805
Quick Revise.....	806
Multiple Choice Questions	806
Subjective Type Questions.....	806
Big Data Practical.....	817
Appendix A: Cassandra	887
Index.....	895