



Integrated Cloud Applications & Platform Services

# Oracle Big Data Fundamentals

Student Guide - Volume I

D86898GC20

Edition 2.0 | May 2017 | D100381

Learn more from Oracle University at [education.oracle.com](https://education.oracle.com)

The Oracle logo is displayed in white text on a red background. The word "ORACLE" is in a large, bold, sans-serif font, with a registered trademark symbol (®) to its upper right.

## Authors

Lauran K. Serhal  
Brian Pottle

## Technical Contributors and Reviewers

Marty Gubar  
Melliyal Annamalai  
Jean-Pierre Dijcks  
Frederick Kush  
Ben Gelernter  
Gail Risdal  
Susan Jang  
Salome Clement  
Ashwin garwal  
Marcos Arancibia  
Mark Hornick  
Charlie Berger  
S. Matt Taylor  
Suresh Mohan  
Robert Stanoch  
Alexey Filanovskiy  
John Wyant  
William Beauregard  
Jean Ihm  
Michael Schulman  
Siva Ravada  
Albert Godfrind  
Alan Wu  
Javier De La Torre Medina  
Thomas Vengal  
Anand Chandak  
Ashok Joshi  
Dimpi Sarmah  
Lakshmi Narapareddi  
Drishya Tm  
Josh Spiegel  
Sharath Bhujani  
Sharon Stephen

## Graphic Editors

Prakash Dharmalingam  
Maheshwari Krishnamurthy  
Kavya Bellur

Copyright © 2017, Oracle and/or its affiliates. All rights reserved.

## Disclaimer

This document contains proprietary information and is protected by copyright and other intellectual property laws. You may copy and print this document solely for your own use in an Oracle training course. The document may not be modified or altered in any way. Except where your use constitutes "fair use" under copyright law, you may not use, share, download, upload, copy, print, display, perform, reproduce, publish, license, post, transmit, or distribute this document in whole or in part without the express authorization of Oracle.

The information contained in this document is subject to change without notice. If you find any problems in the document, please report them in writing to: Oracle University, 500 Oracle Parkway, Redwood Shores, California 94065 USA. This document is not warranted to be error-free.

## Restricted Rights Notice

If this documentation is delivered to the United States Government or anyone using the documentation on behalf of the United States Government, the following notice is applicable:

### U.S. GOVERNMENT RIGHTS

The U.S. Government's rights to use, modify, reproduce, release, perform, display, or disclose these training materials are restricted by the terms of the applicable Oracle license agreement and/or the applicable U.S. Government contract.

## Trademark Notice

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

## Editors

Smita Kommuni  
Vijayalakshmi Narasimhan

## Publishers

Jayanthi Keshavamurthy  
Raghunath M  
Veena Narasimhan

# Contents

## 1 Introduction

Objectives	1-2
Questions About You	1-3
Course Objectives	1-4
Course Road Map	1-5
Course Road Map: Module 1 – Big Data Fundamentals	1-6
Course Road Map: Module 2 – Data Acquisition and Storage	1-7
Course Road Map: Module 3 – Data Access and Processing	1-8
Course Road Map: Module 4 – Data Unification	1-9
Course Road Map: Module 5 – Data Analysis	1-10
Course Road Map: Module 6 – Oracle Big Data Deployment Options	1-11
Oracle Big Data Lite (BDLite) Virtual Machine (VM) Home Page	1-12
Connecting to the Practice Environment	1-13
Starting the Oracle BDLite VM Used in this Course	1-14
Starting the Oracle BDLite VM Used in This Course	1-15
Accessing the Getting Started Page from the Oracle BDLite VM	1-16
Accessing the Practice Files	1-17
Accessing the /home/oracle/exercises Directory	1-18
Accessing the /home/oracle/movie Directory	1-19
Oracle MoviePlex Demo on Oracle Big Data Lite Landing Page	1-20
Appendixes	1-21
Oracle Big Data Appliance Help Center Documentation	1-22
Oracle Big Data Appliance Documentation	1-23
Additional Resources: Oracle Big Data Tutorials in the Oracle Learning Library (OLL)	1-24
Oracle Big Data Tutorials in the OLL	1-25
Product Libraries	1-26
Oracle Big Data Landing Page	1-27
Oracle Big Data Administration Series	1-28
Oracle University Courses	1-29
Oracle University Courses: Oracle Big Data Fundamentals	1-30
Practice 1-1: Overview	1-32
Summary	1-33

## **2 Introducing Oracle Big Data Strategy**

- Course Road Map 2-2
- Lesson Objectives 2-3
- Big Data: A Strategic IM Perspective 2-4
- Big Data 2-5
- Characteristics of Big Data 2-6
- Big Data Opportunities: Examples 2-8
- Publishing: Customer 360 View 2-9
- Mobile Phone Service: Increased Productivity 2-10
- Mobile Phone Provider: Fraud Prevention 2-11
- Transportation Equipment: Analytic Performance 2-12
- Banking: Increasing Sales 2-13
- Gaming: Increasing Revenue 2-14
- Retail and CPG: Improved Customer Response 2-15
- Oracle Big Data Momentum: Customers Around the World 2-16
- Big Data Challenges 2-17
- Information Management Landscape 2-19
- Extending the Boundaries of Information Management with Big Data 2-20
- Integrating Data: Unstructured, Semi-structured, and Structured 2-21
- Big Data Processing Engines: Comparison Preview 2-22
- Big Data Integration Considerations 2-23
- Summary 2-24

## **3 Using Oracle Big Data Lite Virtual Machine**

- Course Road Map 3-2
- Agenda 3-3
- Lesson Objectives 3-4
- Oracle Big Data Lite VM Used in This Course 3-5
- Oracle Big Data Lite VM Home Page Sections 3-6
- Components of the Oracle Big Data Lite VM 3-7
- Downloading, Installing, and Using the Oracle Big Data Lite VM 3-8
  - 1. Reviewing the Deployment Guide 3-9
  - 2. Downloading Oracle VM VirtualBox Plus and Its Extension Pack 3-10
  - 3. Downloading the Required 7-zip Files 3-11
  - 4. Installing Oracle VM VirtualBox and Its Extension Pack 3-12
  - 5. Running the 7-zip Extractor on BigDataLite4701.7z.001 File 3-13
  - 6. Importing BigDataLite4701.ova 3-14
  - 7. Starting BigDataLite-4.7.0.1 3-15
  - 8. Logging In as oracle/welcome1 3-16
- Reviewing the Important “Start Here” Page 3-17
- Starting and Stopping Services 3-18

- Big Data Lite Samples 3-19
- Available Tools on the Browser's Toolbar 3-20
- Agenda 3-21
- Oracle MoviePlex Case Study: Introduction 3-22
- Big Data Challenge 3-23
- Deriving Value from Big Data 3-24
- Oracle MoviePlex: Goal 3-25
- Oracle MoviePlex: Big Data Challenges 3-26
- Oracle MoviePlex: Architecture 3-27
- Oracle MoviePlex: Data Generation Format 3-29
- Oracle MoviePlex Application 3-30
- Summary 3-31

#### **4 Introduction to the Big Data Ecosystem**

- Course Road Map 4-2
- Objectives 4-3
- Computer Clusters 4-4
- Distributed Computing 4-5
- Apache Hadoop 4-6
- Types of Analyses That Use Hadoop 4-7
- Types of Data Generated 4-8
- Apache Hadoop Core Components 4-9
- Apache Hadoop Core Components: HDFS 4-10
- Apache Hadoop Core Components: MapReduce Framework (MRv1) 4-11
- Running Applications Before Hadoop 2.x with MapReduce 1 (MR 1) 4-12
- Apache Hadoop Core Components: YARN (MR2) 4-13
- Running Applications Starting with Hadoop 2.x With YARN (MR 2) 4-14
- Apache Hadoop Ecosystem 4-15
- Additional Resources: Cloudera Distribution 4-16
- Oracle Big Data Appliance (BDA) 4-17
- Additional Resources: Apache Hadoop 4-18
- CDH Architecture 4-19
- CDH Components 4-20
- CDH Architecture 4-21
- CDH Components 4-23
- Hadoop Major Timelines at a Glance 4-24
- Where to Go for More Information 4-25
- Summary 4-26

#### **5 Introduction to the Hadoop Distributed File System (HDFS)**

- Course Road Map 5-2

Objectives	5-3
Agenda	5-4
HDFS Design Principles and Characteristics	5-5
HDFS Key Definitions	5-6
HDFS Deployments: High Availability (HA) and Non-HA	5-7
Sample Hadoop High Availability (HA) Cluster	5-8
HDFS Files and Blocks	5-9
Blocks are Replicated in the Cluster Upon Ingestion into HDFS	5-10
Active and Standby NameNodes Daemons	5-11
DataNodes Daemons	5-12
Functions of the NameNode	5-13
Functions of DataNodes	5-14
Writing a File to HDFS: Example	5-15
Writing a File to HDFS: File is “Chunked” into Blocks – Example	5-16
Writing a File to HDFS: Pipeline Created, Block A – Example	5-17
Writing a File to HDFS: Pipeline Created, Block B – Example	5-18
Writing a File to HDFS: Pipeline Created, Block C – Example	5-19
Writing a File to HDFS: Example	5-20
HDFS High Availability (HA) Using the Quorum Journal Manager (QJM)	5-21
HDFS High Availability (HA) Using the Quorum Journal Manager (QJM) Feature	5-22
Enabling HDFS HA	5-24
Data Replication Rack-Awareness in HDFS	5-25
Data Replication Process	5-26
Accessing HDFS	5-27
Agenda	5-28
Using Cloudera Hue to Interact with HDFS	5-29
Using Hadoop Client to Batch Load Data	5-30
HDFS Commands	5-31
HDFS File System (FS) Shell Interface	5-32
HDFS FS (File System) Shell Interface	5-33
FS Shell Commands	5-34
Sample FS Shell Commands	5-35
ls Command	5-36
mkdir and copyFromLocal Commands	5-37
rm and cat Commands	5-38
Using the hdfs fsck Command: Example	5-39
Agenda	5-40
Loading Data with WebHDFS or HttpFS	5-41
hadoop fs -ls and LISTSTATUS	5-42
Uploading a Local File to an HDFS Directory with hadoop fs	5-43

Creating an HDFS Directory with WebHDFS 5-44  
Uploading a Local File to HDFS with WebHDFS 5-45  
Creating an HDFS Directory and Loading Data by Using HttpFS 5-46  
Summary 5-47  
Practice 5: Overview 5-48

## **6 Acquiring Data by Using CLI, Fuse DFS, Flume, and Kafka**

Course Road Map 6-2  
Objectives 6-3  
Reviewing the Command-Line Interface (CLI) 6-4  
Viewing File System Contents by Using the CLI 6-5  
Loading Data by Using the CLI 6-6  
What Is Fuse DFS? 6-7  
Enabling Fuse DFS on Big Data Lite 6-8  
Using Fuse DFS 6-9  
What Is Flume? 6-10  
Flume: Architecture 6-11  
Flume Sources (Consume Events) 6-12  
Flume Channels (Hold Events) 6-13  
Flume Sinks (Deliver Events) 6-14  
Flume: Data Flows 6-15  
Configuring Flume 6-16  
Exploring a flume\*.conf File 6-17  
What Is Apache Kafka? 6-18  
Activity Log Processing in Real Time 6-19  
Apache Kafka: Features 6-20  
Apache Kafka: Basic Concepts 6-21  
Using Apache Kafka: Oracle MoviePlex Examples 6-22  
Additional Resources 6-23  
Summary 6-24  
Practice 6: Overview 6-25

## **7 Acquiring and Accessing Data by Using Oracle NoSQL Database**

Course Road Map 7-2  
Objectives 7-3  
What Is a NoSQL Database? 7-4  
NoSQL Key-Value Data Model 7-5  
What Is Oracle NoSQL Database? 7-7  
Oracle NoSQL Supported Data Types 7-8  
When to Use Oracle NoSQL? 7-9  
Oracle MoviePlex Application: Data in Oracle NoSQL Database 7-10

Using Oracle NoSQL Database with MoviePlex 7-11  
Acquiring and Accessing Data in a Key-Value Store 7-12  
Accessing the KVStore 7-13  
Startup KVLite: Single Process Version of Oracle NoSQL Database 7-14  
Access Methods 7-15  
Creating Tables Using the JAVA API 7-16  
Oracle NoSQL Command-Line Interfaces 7-17  
Data Definition Language (DDL) Commands 7-18  
Using CREATE TABLE 7-19  
Executing a DDL Command 7-21  
Viewing Table Descriptions 7-22  
Recommendation: Using Scripts 7-23  
Loading Data Into Tables Using the JAVA API 7-24  
Introducing the TableAPI 7-25  
Write Operations: put() Methods 7-26  
Writing Rows to Tables: Steps 7-27  
Constructing a Handle 7-28  
Creating Row Object, Adding Fields, and Writing Record 7-29  
Reading Data from Tables Using the JAVA API 7-30  
Read Operations: get() Methods 7-31  
Retrieving Table Data: Steps 7-32  
Retrieving a Single Row 7-33  
Retrieving Multiple Rows 7-34  
Removing Data from Tables Using the JAVA API 7-35  
Delete Operations: Three Table APIs 7-36  
Deleting Row(s) From a Table: Steps 7-37  
Additional Resources 7-38  
Summary 7-39  
Practice 7: Overview 7-40

## **8 Introduction to MapReduce and YARN Processing Frameworks**

Course Road Map 8-2  
Agenda 8-3  
Objectives 8-4  
Apache Hadoop Core Components 8-5  
MapReduce Framework: Features 8-6  
MapReduce Job 8-7  
Benefits of MapReduce 8-8  
MapReduce Jobs 8-9  
Parallel Processing with MapReduce 8-10  
Word Count Process: Example 1 8-11



MapReduce Mechanics: Deck of Cards Example 8-12  
MapReduce Mechanics Example: Assumptions 8-13  
MapReduce Mechanics: Map Phase 8-14  
MapReduce Mechanics: Shuffle and Sort Phase 8-15  
MapReduce Mechanics: Reduce Phase (Result) 8-16  
Interacting with MapReduce 8-17  
Data Locality Optimization in Hadoop 8-18  
Submitting a MapReduce Job 8-19  
Submitting a WordCount MapReduce Job: Reviewing the Input Data Files 8-20  
Submitting the WordCount MapReduce Job 8-21  
Monitoring MapReduce Jobs by Using the YARN Resource Manager Web UI 8-22  
Monitoring MapReduce Jobs by Using the JobHistory Server Web UI 8-23  
Viewing the WordCount.java Program Output 8-24  
Agenda 8-25  
Running Applications Starting with Hadoop 2.x with YARN (MR 2) 8-26  
YARN Architecture 8-27  
YARN: Features 8-28  
YARN Daemons 8-29  
YARN Architecture 8-30  
YARN (MRv2) Architecture 8-31  
YARN Application Workflow 8-32  
Hadoop Basic Cluster (MRv1): Example 8-34  
Hadoop Basic Cluster YARN (MRv2): Example 8-35  
Summary 8-36  
Practice 8: Overview 8-37

## **9 Resource Management Using YARN**

Course Road Map 9-2  
Objectives 9-3  
Agenda 9-4  
Job Scheduling in YARN 9-5  
YARN Fair Scheduler 9-6  
Cloudera Manager Resource Management: Features 9-8  
Static Service Pools 9-9  
Working with the Fair Scheduler 9-10  
Cloudera Manager Dynamic Resource Management: Example 9-11  
Submitting a Job to hrpool by the lucy User from the hr Group 9-17  
Monitoring the Status of the Submitted MapReduce Job 9-18  
Examining marketingpool 9-19  
Submitting a Job to marketingpool by the lucy User from the hr Group 9-20  
Monitoring the Status of the Submitted MapReduce Job 9-21

Submitting a Job to marketingpool by the bob User from the marketing Group 9-22  
Monitoring the Status of the Submitted MapReduce Job 9-23  
Delay Scheduling 9-24  
Agenda 9-25  
YARN application Command 9-26  
YARN application Command: Example 9-27  
Monitoring an Application by Using the ResourceManager Web UI 9-29  
Scheduler: BDA Example 9-30  
Summary 9-31  
Practice 9 9-32

## **10 Overview of Spark**

Course Road Map 10-2  
Objectives 10-3  
What Is Apache Spark? 10-4  
Benefits of Using Spark 10-5  
Spark Versus MapReduce 10-6  
How is Spark Used? 10-7  
Spark Architecture 10-8  
Spark Application Components 10-10  
Spark Driver 10-11  
Spark Master 10-13  
Cluster Manager 10-14  
Executors 10-15  
Running a Spark Application on YARN (yarn-cluster Mode) 10-16  
Running a Spark Application on YARN (yarn-client Mode) 10-17  
Spark on YARN 10-18  
Resilient Distributed Dataset (RDD) 10-19  
Resilient Distributed Dataset (RDD) 10-20  
RDDs 10-21  
RDD Operations 10-23  
Commonly Used Transformations 10-24  
Sample Actions 10-25  
RDD Example 10-26  
RDD Transformation: Filter() example 10-27  
RDD Action: Count() Example 10-28  
Spark Interactive Shells 10-29  
Scala Language: Overview 10-30  
Starting Interactive Spark-Shell for Scala 10-31  
Using the ResourceManager UI to Monitor a Spark Application 10-32  
Word Count Example by Using Interactive Scala 10-33

Word Count Example by Using Interactive Scala: Output 10-34  
Monitoring Spark Jobs by Using YARN's ResourceManager Web UI 10-35  
Scala Program: Word Count Example 10-36  
Definitions 10-37  
Launching Spark Applications Using the spark-submit Command 10-38  
Summary 10-39  
Practice 10: Overview 10-40

## **11 Overview of Hive**

Course Road Map 11-2  
Objectives 11-3  
Hive 11-4  
Use Case: Storing Clickstream Data 11-5  
Hadoop Architecture 11-6  
How Is Data Stored in HDFS? 11-7  
How Is Data Stored in HDFS 11-8  
Organizing and Describing Data with Hive 11-9  
How Does Hive Read ANY Data? 11-10  
How Does Hive Read Data FASTER? 11-11  
Big Data SQL on Top of "Hive" Data 11-12  
Defining Tables Over HDFS 11-13  
Defining Tables over HDFS 11-14  
Hive: Data Units 11-15  
Hive Metastore Database 11-16  
Hive Framework 11-17  
Creating a Hive Database 11-18  
Data Manipulation in Hive 11-19  
Data Manipulation in Hive: Nested Queries 11-20  
Steps in a Hive Query 11-21  
Hive-Based Applications 11-22  
Hive: Limitations 11-23  
Summary 11-24  
Practice 11: Overview 11-25

## **12 Overview of Cloudera Impala**

Course Road Map 12-2  
Objectives 12-3  
Hadoop: Some Data Access/Processing Options 12-4  
Cloudera Impala 12-5  
Cloudera Impala: Key Features 12-6  
Cloudera Impala: Supported Data Formats 12-7

Cloudera Impala: Programming Interfaces 12-8  
How Impala Fits Into the Hadoop Ecosystem 12-9  
How Impala Works with Hive 12-10  
How Impala Works with HDFS and HBase 12-11  
Summary of Cloudera Impala Benefits 12-12  
Impala and Hadoop: Limitations 12-13  
Summary 12-14

## **13 Using Oracle XQuery for Hadoop**

Course Road Map 13-2  
Objectives 13-3  
XML 13-4  
Simple XML Document: Example 13-5  
XML Elements 13-6  
Markup Rules for Elements 13-7  
XML Attributes 13-8  
XML Path Language 13-9  
XPath Terminology: Node Types 13-10  
XPath Terminology: Family Relationships 13-11  
XPath Expressions 13-12  
Location Path Expression: Example 13-13  
XQuery: Review 13-14  
XQuery Terminology 13-15  
XQuery Review: books.xml Document Example 13-16  
FLWOR Expressions: Review 13-17  
Oracle XQuery for Hadoop (OXH) 13-18  
OXH Features 13-19  
Oracle XQuery for Hadoop Data Flow 13-20  
Using OXH 13-21  
OXH Installation 13-22  
OXH Functions 13-23  
OXH Adapters 13-24  
Running a Query: Syntax 13-25  
OXH: Configuration Properties 13-26  
XQuery Transformation and Basic Filtering: Example 13-27  
Viewing the Completed Application in YARN 13-30  
Calling Custom Java Functions from XQuery 13-31  
Additional Resources 13-32  
Summary 13-33  
Practice 13: Overview 13-34

## **14 Overview of Solr**

- Course Road Map 14-2
- Objectives 14-3
- Apache Solr (Cloudera Search) 14-4
- Cloudera Search: Features 14-5
- Cloudera Search Requirements 14-7
- Indexing Data for Cloudera Search 14-8
- Indexing Types 14-9
- What Is the Schema.XML File? 14-11
- solrctl Command: Options and Example 14-12
- Creating a Solr Collection 14-13
- Using Oracle XQuery with Solr 14-14
- Using Solr with Hue 14-15
- Summary 14-16
- Practice 14: Overview 14-17

## **15 Integrating Your Big Data**

- Course Road Map 15-2
- Objectives 15-3
- Unifying Data: Typical Requirement 15-4
- Integrating Data of Different Usage Patterns 15-5
- Introducing Data Unification Technologies 15-6
- Data Unification: Comparing Big Data Processing Engines 15-7
- Comparing Hadoop, NoSQL, and RDBMS 15-8
- Comparing Data Ingest 15-9
- Data Ingest: Hadoop 15-10
- Data Ingest: NoSQL 15-11
- Data Ingest: RDBMS 15-12
- Data Ingest: Summary 15-13
- Comparing Disaster Recovery 15-14
- Disaster Recovery: Hadoop 15-15
- Disaster Recovery: NoSQL 15-16
- Disaster Recovery: RDBMS 15-17
- Big Data Appliance Including Cloudera BDR 15-18
- Expanded Summary 15-19
- Comparing Data Access 15-20
- Data Sets and Analytical Queries: Hadoop 15-21
- Data Sets and Analytical Queries: NoSQL 15-22
- Data Sets and Analytical Queries: RDBMS 15-23
- Expanded Summary 15-24
- Comparing Performance 15-25

Data Ingest Performance 15-26  
Query Performance 15-27  
Comparing Cost 15-28  
Cost Consideration: RDBMS to Hadoop? 15-29  
Big Data Processing Engines: Comparative Conclusions 15-30  
Data Unification: Batch Loading 15-31  
Sqoop 15-32  
Oracle Loader for Hadoop (OLH) 15-33  
Copy to Hadoop 15-34  
Data Unification: Batch and Dynamic Loading 15-35  
Oracle SQL Connector for Hadoop 15-36  
Data Unification: ETL and Synchronization 15-37  
Oracle Data Integrator for Big Data - Heterogeneous Integration with Hadoop  
Environments 15-38  
Oracle GoldenGate for Big Data 15-39  
Data Unification: Data Virtualization 15-40  
What Is Oracle Big Data SQL? 15-41  
Hadoop Architectural View 15-42  
Big Data SQL: Another Hadoop Processing Engine 15-43  
When to Use Different Oracle Technologies 15-44  
Summary 15-45

## **16 Batch Loading Options**

Course Road Map 16-2  
Objectives 16-3  
Apache Sqoop 16-4  
Sqoop Components 16-5  
Sqoop Features 16-6  
Sqoop: Connectors 16-7  
Importing Data into Hive 16-8  
Sqoop: Summary 16-9  
Oracle Loader for Hadoop (OLH) 16-10  
OLH Specifications 16-11  
OLH: Online Database Mode 16-12  
Running an OLH Job 16-13  
OLH Use Cases 16-14  
Load Balancing in OLH 16-15  
Data Input Formats 16-16  
OLH: Offline Database Mode 16-17  
Offline Load Advantages in OLH 16-18  
OLH Versus Sqoop 16-19

- OLH: Summary 16-20
- Copy to Hadoop 16-21
- Copy to Hadoop: Functional Steps 16-22
- Step 1: Identify the Target Directory 16-23
- Step 2: Create an External Table 16-24
- Step 3: Copy Files to HDFS 16-25
- Step 4: Create a Hive External Table 16-26
- Oracle to Hive Data Type Conversions 16-27
- Querying the Data in Hive 16-28
- Summary 16-29
- Practice 16: Overview 16-30

## **17 Using Oracle SQL Connector for HDFS**

- Course Road Map 17-2
- Objectives 17-3
- Oracle SQL Connector for HDFS 17-4
- OSCH Architecture 17-5
- Using OSCH: Primary Steps 17-6
- Using OSCH: Creating Directory Object 17-7
- Using OSCH: Database Objects and Grants 17-8
- Using OSCH: Supported Data Input Formats 17-9
- Using OSCH: HDFS Text File Support 17-10
- Using OSCH: Hive Table Support 17-12
- Using OSCH: Partitioned Hive Table Support 17-14
- OSCH: Features 17-15
- Parallelism and Performance 17-16
- OSCH: Performance Tuning 17-17
- OSCH: Key Benefits 17-18
- Loading Data to Oracle: Choosing a Connector 17-19
- Summary 17-20
- Practice 17: Overview 17-21

## **18 Using Oracle Data Integrator and Oracle GoldenGate for Big Data**

- Course Road Map 18-2
- Objectives 18-3
- Oracle Data Integrator 18-4
- ODI's Declarative Design 18-5
- ODI Knowledge Modules (KMs) Simpler Physical Design / Shorter Implementation Time 18-6
- Using ODI for Big Data Heterogeneous Integration with Hadoop Environments 18-7
- ODI Studio 18-8

ODI Studio Components	18-9
ODI Studio: Big Data Knowledge Modules	18-10
Oracle GoldenGate for Big Data	18-11
Oracle GoldenGate Studio	18-12
Oracle GoldenGate: Flexible Deployment	18-13
Using OGG for Big Data: Hive Example	18-14
Using OGG for Big Data: Kafka Example	18-15
Resources	18-16
Summary	18-17
Practice 18: Overview	18-18

## **19 Using Oracle Big Data SQL**

Course Road Map	19-2
Objectives	19-3
The New Normal for Oracle Customers	19-4
Big Data SQL: A Single SQL Interface for All Big Data	19-5
Using Oracle Big Data SQL: Overview	19-6
Using Big Data SQL: Agenda	19-7
Demonstration of Big Data SQL	19-8
Using Big Data SQL: Agenda	19-27
Extending Oracle External Tables for Metadata Support	19-28
Making External Tables More Like Internal Tables	19-29
Creating External Tables Over HDFS Data	19-30
Using Access Parameters with oracle_hdfs	19-31
Creating External Tables to Leverage the Hive Metastore	19-32
Using Access Parameters with oracle_hive	19-33
Automating External Table Creation	19-35
Using Big Data SQL: Agenda	19-36
Big Data SQL Security Benefits	19-37
Applying Oracle Database Security Policies	19-38
Viewing the Results	19-39
Applying Redaction Policies to Data in Hadoop	19-40
Viewing Results from the Hive (Avro) Source	19-41
Viewing the Results from Joined RDBMS and HDFS Data	19-42
Using Big Data SQL: Agenda	19-43
Anatomy of a Big Data SQL Cell: Convert Data	19-44
Anatomy of a Big Data SQL Cell: Add SmartScan	19-45
Big Data SQL Performance Features	19-46
Using Big Data SQL: Agenda	19-47
Big Data SQL: Expanding Deployment Options	19-48
Big Data SQL: Summary of Key Features	19-49



Available Resources 19-50  
Summary 19-51  
Practice 19: Overview 19-52

## **20 Using Oracle Big Data Spatial and Graph**

Course Road Map 20-2  
Objectives 20-3  
Role of Relationships in Spatial and Graph Analysis 20-4  
What Is Most Important? 20-5  
Motivations for Big Data Spatial and Graph Analysis 20-6  
Oracle Big Data Spatial and Graph Components 20-7  
Modeling and Analyzing the Internet of Things 20-8  
Oracle Big Data Spatial and Graph: Property Graph 20-9  
Property Graph Analysis: Common Use Cases 20-10  
Property Graph Data Model 20-12  
Example: Property Graph Image 20-13  
39 Built-in Graph Functions 20-14  
Key Feature: In-Memory Analyst Provides the following 20-15  
Multiple Interfaces for Many Kinds of Users 20-16  
Integration with a Variety of Visualization Tools 20-17  
Graph Property Differentiators 20-18  
Spatial Analysis Use Case: Insurance Industry Linking Information by  
Location 20-19  
Data Harmonization Example: Linking information by location 20-20  
What Does Big Data Spatial Analysis Provide? 20-21  
Oracle Big Data Spatial Features 20-22  
Vector Spatial Analysis: Binning Example 20-24  
Raster Spatial Analysis: Example 20-25  
Value Proposition for Big Data Spatial Features 20-26  
Oracle Big Data Spatial and Graph: Benefits 20-27  
Oracle Big Data Spatial and Graph: Multimedia Analytics Framework 20-28  
Multiple Deployment Options for Big Data Spatial and Graph 20-29  
Resources 20-30  
Summary 20-31  
Practices for Lesson 20: Overview 20-32

## **21 Using Oracle Advanced Analytics: Oracle Data Mining and Oracle R Enterprise**

Course Road Map	21-2
Objectives	21-3
Oracle Advanced Analytics (OAA)	21-4
OAA: Oracle Data Mining	21-5
What Is Data Mining?	21-6
Common Uses of Data Mining	21-7
Defining Key Data Mining Properties	21-8
Data Mining Categories	21-10
Supervised Data Mining Techniques	21-11
Supervised Data Mining Algorithms	21-12
Unsupervised Data Mining Techniques	21-13
Unsupervised Data Mining Algorithms	21-14
Oracle Data Mining: Overview	21-15
Oracle Data Miner GUI	21-16
ODM SQL Interface	21-17
Oracle Data Miner Big Data Support	21-18
Example Workflow Using JSON Query Node	21-19
ODM Resources	21-20
OAA: Oracle R Enterprise	21-21
What Is R?	21-22
Who Uses R?	21-23
Why Statisticians, Data Analysts, and Data Scientists Use R	21-24
Limitations of R	21-25
Oracle's Strategy for the R Community	21-26
Oracle R Enterprise	21-27
ORE: Software Features	21-28
ORE Packages	21-29
Functions for Interacting with Oracle Database	21-30
ORE: Target Environment	21-31
ORE: Data Sources	21-32
ORE and Hadoop	21-33
ORAAH: Architecture	21-34
ORAAH Package	21-35
HDFS Connectivity and Interaction	21-36
ORAAH Functions for HDFS Interaction	21-37
ORAAH Functions for Predictive Algorithms	21-38
Hadoop Connectivity and Interaction	21-39
Word Count: Example Without ORAAH	21-40
Word Count: Example with ORAAH	21-41
ORE Resources	21-42

Practice 21: Overview 21-43

Summary 21-44

## **22 Introduction to the Oracle Big Data Appliance (BDA)**

Objectives 22-2

Course Road Map 22-3

Agenda 22-4

Big Data Management System 22-5

Oracle BDA 22-6

Core Design Principles for BDA 22-7

Configuring and Installing the Oracle BDA: Road Map 22-8

Configuring and Installing the Oracle BDA: Key Players 22-9

Key Definitions 22-10

Accessing the Big Data Documentation on Oracle Help Center 22-11

Completing the Oracle BDA Site Checklists 22-12

Completing the Oracle BDA Site Checklists Before Configuring and Installing the Software 22-13

Using the Oracle BDA Configuration Generation Utility 22-14

Oracle BDA Configuration Generation Utility Pages 22-15

Downloading and Extracting the Configuration Generation Utility 22-16

Running the Oracle BDA Configuration Generation Utility 22-17

Welcome Screen 22-18

Selecting the Last Option on the Welcome Screen 22-19

Customer Details 22-20

Hardware Selection 22-21

Rack Details and Networking Pages 22-22

Define Clusters (Cluster 1: CDH Cluster) 22-23

Define Clusters (Cluster 2: NoSQL Cluster) 22-24

Cluster 1 22-25

Cluster 2 22-26

Client and InfiniBand Network and Complete Pages 22-27

Complete Deployment Assistant and Generate Files 22-28

Viewing the Generated Directories and Files 22-29

Agenda 22-31

Oracle BDA Provides Flexible Cluster Configurations 22-32

BDA X6-2 Hardware 22-33

BDA X6-2 Integrated and Optional Software 22-34

Oracle BDA Mammoth Software Deployment Bundle 22-35

Using the Oracle BDA mammoth Utility 22-36

Downloading the Mammoth Software Deployment Bundle: Overview 22-38

Searching for Document ID 1445745.2 on MOS 22-39

Viewing Document ID 1445745.2 22-40  
Oracle BDA Patch Set Master Note Page 22-41  
Oracle BDA Mammoth Software Deployment Bundle Installation Document 22-42  
BDA CDH Cluster Service Layout After Deployment 22-43  
Successful Big Data Systems Grow: Example 22-44  
BDA Service Layout: CDH Cluster Only 22-45  
Successful Big Data Systems Grow 22-46  
Automatic Failover of the NameNode 22-47  
Restoring HA and Reinstating a Node 22-48  
Agenda 22-49  
Monitoring the Health of Oracle BDA: Management Utilities 22-50  
Monitoring Oracle BDA 22-51  
Administering the Oracle BDA: Overview 22-52  
What Is Cloudera Manager? 22-53  
Cloudera Manager Web Console (Main Page) 22-54  
Oracle Enterprise Manager BDA Plug-in 22-55  
BDA Network Page: Software Overview 22-56  
BDA Network Page: Schematic Hardware Overview 22-57  
Viewing the BDA Hardware and Software Components 22-58  
Deployment of Secured Clusters on Oracle BDA 22-59  
BDA Secure Installation 22-60  
Resources 22-61  
Summary 22-62

## **23 Introduction to Oracle Big Data Cloud Service**

Objectives 23-2  
Course Road Map 23-3  
Oracle Big Data Cloud Service 23-4  
Oracle Big Data Cloud Service: Key Features 23-5  
Oracle Big Data Cloud Service: Benefits 23-6  
Elasticity: Dedicated Compute Bursting 23-7  
Automated Service 23-8  
Security Made Easy 23-9  
Comprehensive Analytics Toolset Included 23-10  
Comprehensive Data Integration Toolset Included 23-11  
Oracle “Cloud at Customer” 23-12  
Big Data Deployment Models: Choices 23-13  
Before Creating a New CDH Cluster 23-14  
Using Oracle Big Data Cloud Service: Typical Workflow 23-15  
Cluster Nodes 23-16  
Permanent Hadoop Node 23-17

Edge Nodes	23-18
Cluster Compute Nodes	23-19
Creating a Service Instance	23-20
Creating a Cluster	23-21
Performing Other Administrative Operations on a Cluster	23-23
Managing Oracle Big Data Cloud Service	23-24
Appendix B: Pages for Administering Oracle Big Data Cloud Service	23-25
Resources	23-26
Summary	23-27

## **24 Introduction to Oracle Big Data Cloud Service – Compute Edition**

Objectives	24-2
Course Road Map	24-3
What Is Oracle Big Data Cloud Service – Compute Edition?	24-4
Benefits	24-5
Features	24-6
Requirements Before Creating a Cluster	24-7
Starting With Oracle Big Data Cloud Service – Computer Edition	24-8
Typical Oracle Big Data Cloud Service – Computer Edition: Workflow	24-10
Accessing Oracle Big Data Cloud Service – Compute Edition	24-11
Managing Oracle Big Data Cloud Service – Compute Edition	24-13
Managing Network Access	24-14
Managing Credentials	24-15
Managing Data	24-16
Managing Apache Spark Jobs	24-17
Working with Notebook	24-18
Available Interpreters for Oracle Big Data Cloud Service – Compute Edition	24-19
Appendix A: Using Oracle Big Data Cloud Service – Compute Edition	24-20
Resources	24-21
Summary	24-22

## **25 Securing Your Data**

Course Road Map	25-2
Objectives	25-3
Security Trends	25-4
Security Levels	25-5
Outline	25-6
Relaxed Security	25-7
Authentication with Relaxed Security	25-8
Authorization	25-9
HDFS ACLs	25-10

Changing Access Privileges	25-11
Relaxed Security Summary	25-12
Challenges with Relaxed Security	25-13
BDA and BDCS Secure Installation	25-14
Kerberos: Key Definitions	25-15
Strong Authentication with Kerberos	25-16
Snapshot of Principals in KDC	25-17
Authentication with Kerberos	25-18
User Authentication: Examples	25-19
Service Authentication and Keytabs	25-20
TGT Cache: Review	25-21
Ticket Renewal	25-22
Adding a New User	25-23
Adding a New User: Example	25-24
Adding a User to Hue: Example	25-25
Authorization	25-26
Sentry Authorization Features	25-27
Sentry Configuration	25-28
Users, Groups, and Roles	25-29
Sentry Example: Overview	25-30
Users, Roles, and Groups: Example	25-31
1. Creating Roles	25-32
2. Assigning Roles to Groups	25-33
Show Roles for a Group	25-34
Create Databases (in Hive)	25-35
Privileges on Source Data for Tables	25-36
Granting Privileges on Source Data for Tables	25-38
Creating the Table and Loading the Data	25-39
Attempting to Query the Table Without Privileges	25-40
Granting and Revoking Access to the Table	25-41
Sentry Key Configuration Tasks	25-42
Oracle Database Access to HDFS	25-43
Oracle Connection to Hadoop	25-44
Virtual Private Database Policies Restrict Data Access	25-45
Oracle Data Redaction Protects Sensitive Data	25-46
Auditing: Overview	25-47
Auditing	25-48
Cloudera Navigator	25-49
Cloudera Navigator Reporting	25-50
Cloudera Navigator Lineage Analysis	25-51
Encryption	25-52

Network Encryption 25-53  
Data at Rest Encryption 25-54  
Summary 25-55

## **A Glossary**

## **B Resources**

## **C Balancing MapReduce Jobs**

Objectives C-2  
Ideal World: Neatly Balanced MapReduce Jobs C-3  
Real World: Skewed Data and Unbalanced Jobs C-4  
Data Skew C-5  
Data Skew Can Slow Down the Entire Hadoop Job C-6  
Perfect Balance C-7  
How Does Perfect Balance Work? C-8  
Using Perfect Balance C-9  
Perfect Balance Java APIs C-10  
Application Requirements for Using Perfect Balance C-11  
Perfect Balance: Benefits C-12  
Using Job Analyzer C-13  
Getting Started with Perfect Balance C-14  
Using Job Analyzer C-16  
Environmental Setup for Perfect Balance and Job Analyzer C-17  
Running Job Analyzer as a Stand-alone Utility to Measure Data Skew in Unbalanced Jobs C-18  
Using Job Analyzer as a Stand-Alone Utility: Example with a YARN Cluster C-19  
Configuring Perfect Balance C-20  
Using Perfect Balance to Run a Balanced MapReduce Job C-21  
Perfect Balance–Generated Reports C-23  
Job Analyzer Reports: Structure of the Job Output Directory C-24  
Reading Job Analyzer Reports C-25  
Reading the Job Analyzer Report in HDFS Using a Web Browser C-26  
Reading the Job Analyzer Report in the Local File System in a Web Browser C-27  
Looking for Skew Indicators in Job Analyzer Reports C-28  
Job Analyzer Sample Reports C-29  
Using Data from Additional Metrics C-30  
Chopping C-31

Disabling Chopping	C-32
Troubleshooting Jobs Running with Perfect Balance	C-33
Perfect Balance Examples Available with Installation	C-34
Summary	C-35