

[Table of contents](#)

[Foreword](#)

[Preface](#)

[Who This Book Is For](#)

[What You Should Already Know](#)

[What This Book Leaves Out](#)

[How This Book Works](#)

[Which Software Versions This Book Uses](#)

[Conventions Used in This Book](#)

[IP Addresses](#)

[Using Code Examples](#)

[Oâ€™Reilly Safari](#)

[How to Contact Us](#)

[Acknowledgments](#)

[I. Introduction to the Cloud](#)

[1. Why Hadoop in the Cloud?](#)

[What Is the Cloud?](#)

[What Does Hadoop in the Cloud Mean?](#)

[Reasons to Run Hadoop in the Cloud](#)

[Reasons to Not Run Hadoop in the Cloud](#)

[What About Security?](#)

[Hybrid Clouds](#)

[Hadoop Solutions from Cloud Providers](#)

[Elastic MapReduce](#)

[Google Cloud Dataproc](#)

[HDInsight](#)

[Hadoop-Like Services](#)

[A Spectrum of Choices](#)

[Getting Started](#)

[2. Overview and Comparison of Cloud Providers](#)

[Amazon Web Services](#)

[References](#)

[Google Cloud Platform](#)

[References](#)

[Microsoft Azure](#)

[References](#)

[Which One Should You Use?](#)

[II. Cloud Primer](#)

[3. Instances](#)

[Instance Types](#)

[Regions and Availability Zones](#)

[Instance Control](#)

[Temporary Instances](#)

[Spot Instances](#)

[Preemptible Instances](#)

[Images](#)

[No Instance Is an Island](#)

[4. Networking and Security](#)

[A Drink of CIDR](#)
[Virtual Networks](#)
[Private DNS](#)
[Public IP Addresses and DNS](#)
[Virtual Networks and Regions](#)
[Routing](#)
[Routing in AWS](#)
[Routing in Google Cloud Platform](#)
[Routing in Azure](#)
[Network Security Rules](#)
[Inbound Versus Outbound](#)
[Allow Versus Deny](#)
[Network Security Rules in AWS](#)
[Network Security Rules in Google Cloud Platform](#)
[Network Security Rules in Azure](#)
[Putting Networking and Security Together](#)
[What About the Data?](#)
[5. Storage](#)
[Block Storage](#)
[Block Storage in AWS](#)
[Block Storage in Google Cloud Platform](#)
[Block Storage in Azure](#)
[Object Storage](#)
[Buckets](#)
[Data Objects](#)
[Object Access](#)
[Object Storage in AWS](#)
[Object Storage in Google Cloud Platform](#)
[Object Storage in Azure](#)
[Cloud Relational Databases](#)
[Cloud Relational Databases in AWS](#)
[Cloud Relational Databases in Google Cloud Platform](#)
[Cloud Relational Databases in Azure](#)
[Cloud NoSQL Databases](#)
[Where to Start?](#)
[III. A Simple Cluster in the Cloud](#)
[6. Setting Up in AWS](#)
[Prerequisites](#)
[Allocating Instances](#)
[Generating a Key Pair](#)
[Launching Instances](#)
[Securing the Instances](#)
[Next Steps](#)
[7. Setting Up in Google Cloud Platform](#)
[Prerequisites](#)
[Creating a Project](#)
[Allocating Instances](#)
[SSH Keys](#)

[Creating Instances](#)
[Securing the Instances](#)
[Next Steps](#)
[8. Setting Up in Azure](#)
[Prerequisites](#)
[Creating a Resource Group](#)
[Creating Resources](#)
[SSH Keys](#)
[Creating Virtual Machines](#)
[The Manager Instance](#)
[The Worker Instances](#)
[Next Steps](#)
[9. Standing Up a Cluster](#)
[The JDK](#)
[Hadoop Accounts](#)
[Passwordless SSH](#)
[Hadoop Installation](#)
[HDFS and YARN Configuration](#)
[The Environment](#)
[XML Configuration Files](#)
[Finishing Up Configuration](#)
[Startup](#)
[SSH Tunneling](#)
[Running a Test Job](#)
[What If the Job Hangs?](#)
[Running Basic Data Loading and Analysis](#)
[Wikipedia Exports](#)
[Analyzing a Small Export](#)
[Go Bigger](#)
[IV. Enhancing Your Cluster](#)
[10. High Availability](#)
[Planning HA in the Cloud](#)
[HDFS HA](#)
[YARN HA](#)
[Installing and Configuring ZooKeeper](#)
[Adding New HDFS and YARN Daemons](#)
[The Second Manager](#)
[HDFS HA Configuration](#)
[YARN HA Configuration](#)
[Testing HA](#)
[Improving the HA Configuration](#)
[A Bigger Cluster](#)
[Complete HA](#)
[A Third Availability Zone?](#)
[Benchmarking HA](#)
[MRBench](#)
[Terasort](#)
[Grains of Salt](#)

[11. Relational Data with Apache Hive](#)

[Planning for Hive in the Cloud](#)

[Installing and Configuring Hive](#)

[Startup](#)

[Running Some Test Hive Queries](#)

[Switching to a Remote Metastore](#)

[The Remote Metastore and Stopped Clusters](#)

[Hive Control Scripts](#)

[Hive on S3](#)

[Configuring the S3 Filesystem](#)

[Adding Data to S3](#)

[Configuring S3 Authentication](#)

[Configuring the S3 Endpoint](#)

[External Table in S3](#)

[What About Google Cloud Platform and Azure?](#)

[A Step Toward Transient Clusters](#)

[A Different Means of Computation](#)

[12. Streaming in the Cloud with Apache Spark](#)

[Planning for Spark in the Cloud](#)

[Installing and Configuring Spark](#)

[Startup](#)

[Running Some Test Jobs](#)

[Configuring Hive on Spark](#)

[Add Spark Libraries to Hive](#)

[Configure Hive for Spark](#)

[Switch YARN to the Fair Scheduler](#)

[Try Out Hive on Spark on YARN](#)

[Spark Streaming from AWS Kinesis](#)

[Creating a Kinesis Stream](#)

[Populating the Stream with Data](#)

[Streaming Kinesis Data into Spark](#)

[What About Google Cloud Platform and Azure?](#)

[Building Clusters Versus Building Clusters Well](#)

[V. Care and Feeding of Hadoop in the Cloud](#)

[13. Pricing and Performance](#)

[Picking Instance Types](#)

[The Criteria](#)

[General Cluster Instance Roles](#)

[Persistent Versus Ephemeral Block Storage](#)

[Stopping and Starting Entire Clusters](#)

[Using Temporary Instances](#)

[Geographic Considerations](#)

[Regions](#)

[Availability Zones](#)

[Performance and Networking](#)

[14. Network Topologies](#)

[Public and Private Subnets](#)

[SSH Tunneling](#)

[SOCKS Proxy](#)
[VPN Access](#)
[Access from Other Subnets](#)
[Cluster Topologies](#)
[The Public Cluster](#)
[The Secured Public Cluster](#)
[Gateway Instances](#)
[The Private Cluster](#)
[Cluster Access to the Internet and Cloud Provider Services](#)
[Geographic Considerations](#)
[Regions](#)
[Availability Zones](#)
[Starting Topologies](#)
[Higher-Level Planning](#)
[15. Patterns for Cluster Usage](#)
[Long-Running or Transient?](#)
[Single-User or Multitenant?](#)
[Self-Service or Managed?](#)
[Cloud-Only or Hybrid?](#)
[Watching Cost](#)
[The Rising Need for Automation](#)
[16. Using Images for Cluster Management](#)
[The Structure of an Image](#)
[EC2 Images](#)
[GCE Images](#)
[Azure Images](#)
[Image Preparation](#)
[Wait, Iâ€™m Using That!](#)
[Image Creation](#)
[Image Creation in AWS](#)
[Image Creation in Google Cloud Platform](#)
[Image Creation in Azure](#)
[Image Use](#)
[Scripting Hadoop Configuration](#)
[Image Maintenance](#)
[Image Deletion](#)
[Image Deletion in AWS](#)
[Image Deletion in Google Cloud Platform](#)
[Image Deletion in Azure](#)
[Automated Image Creation with Packer](#)
[Automated Cloud Cluster Creation](#)
[Cloudera Director](#)
[Hortonworks Data Cloud](#)
[Qubole Data Service](#)
[General System Management Tools](#)
[Images or Tools?](#)
[More Tooling](#)
[17. Monitoring and Automation](#)

[Monitoring Choices](#)
[Cloud Provider Monitoring Services](#)
[Rolling Your Own](#)
[Cloud Provider Command-Line Interfaces](#)
[AWS CLI](#)
[Google Cloud Platform CLI](#)
[Azure CLI](#)
[Data Formatting for CLI Results](#)
[What to Monitor](#)
[Instance Existence](#)
[Instance Reachability](#)
[Hadoop Daemon Status](#)
[System Load](#)
[Putting Scripting to Use](#)
[Custom Metrics in CloudWatch](#)
[Basic Metrics](#)
[Defining a Custom Metric](#)
[Feeding Custom Metric Data to CloudWatch](#)
[Setting an Alarm on a Custom Metric](#)
[Elastic Compute Using a Custom Metric](#)
[A Custom Metric for Compute Capacity](#)
[Prerequisites for Autoscaling Compute](#)
[Triggering Autoscaling with an Alarm Action](#)
[What About Shrinking?](#)
[Other Things to Watch](#)
[Ingesting Logs into CloudWatch](#)
[Creating an IAM User for Log Streaming](#)
[Installing the CloudWatch Agent](#)
[Creating a Metric Filter](#)
[Creating an Alarm from a Metric Filter](#)
[So Much More to See and Do](#)
[18. Backup and Restoration](#)
[Patterns to Supplement Backups](#)
[Backup via Imaging](#)
[HDFS Replication](#)
[Cloud Storage Filesystems](#)
[HDFS Snapshots](#)
[Hive Metastore Replication](#)
[Logs](#)
[A General Cloud Hadoop Backup Strategy](#)
[Not So Different, But Better](#)
[To the Cloud](#)
[A. Hadoop Component Start and Stop Scripts](#)
[Apache ZooKeeper](#)
[Apache Hive](#)
[B. Hadoop Cluster Configuration Scripts](#)
[SSH Key Creation and Distribution](#)
[Configuration Update Script](#)

[New Worker Configuration Update Script](#)

[C. Monitoring Cloud Clusters with Nagios](#)

[Where Nagios Should Run](#)

[Instance Existence Through Ping](#)

[Hosts and Host Groups](#)

[Services and Service Groups](#)

[Provider CLI Integration](#)

[Index](#)