Apache Hadoop 3 0 0 Hdfs Architecture

Mastering Apache Spark 2.xPackt Publishing Ltd

Construct a robust end-to-end solution for analyzing andvisualizing streaming data Real-time analytics is the hottest topic in data analyticstoday. In Real-Time Analytics: Techniques to Analyze andVisualize Streaming Data, expert Byron Ellis teaches dataanalysts technologies to build an effective real-time analyticsplatform. This platform can then be used to make sense of theconstantly changing data that is beginning to outpace traditionalbatch-based analysis platforms. The author is among a very few leading experts in the field. Hehas a prestigious background in research, development, analytics,real-time visualization, and Big Data streaming and is uniquelyqualified to help you explore this revolutionary field. Moving froma description of the overall analytic architecture of real-timeanalytics to using specific tools to obtain targeted results,Real-Time Analytics leverages open source and moderncommercial tools to construct robust, efficient systems that canprovide real-time analysis in a cost-effective manner. The bookincludes: A deep discussion of streaming data systems andarchitectures Instructions for analyzing, storing, and delivering streamingdata Tips on aggregating data and working with sets Information on data warehousing options and techniques Real-Time Analytics includes in-depth case studies forwebsite analytics, Big Data, visualizing streaming and mobile data,and mining and visualizing operational data

flows. The book's"recipe" layout lets readers quickly learn and implement differenttechniques. All of the code examples presented in the book, alongwith their related data sets, are available on the companionwebsite.

Big Data Analytics(BDA) is a rapidly evolving field that finds applications in many areas such as healthcare, medicine, advertising, marketing, and sales. This book dwells on all the aspects of Big Data Analytics and covers the subject in its entirety. It comprises several illustrations, sample codes, case studies and real-life analytics of datasets such as toys, chocolates, cars, and student's GPAs. The book will serve the interests of undergraduate and post graduate students of computer science and engineering, information technology, and related disciplines. It will also be useful to software developers. Salient Features: - Comprehensive coverage on Big Data NoSQL Column-family, Object and Graph databases, programming with open-source Big Data - Hadoop and Spark ecosystem tools, such as MapReduce, Hive, Pig, Spark, Python, Mahout, Streaming, GraphX - Inclusion of latest topics machine learning, K-NN, predictive-analytics, similar and frequent item sets, clustering, decision-tree, classifiers recommenders, real-time streaming data analytics, graph networks, text, web structure, web-links, social network analytics. - Web supplement includes instructional PPT's, solution of exercises, analysis using open source datasets of a car company, and topics for advanced learning.

This book includes 46 scientific papers presented at the conference and reflecting the

latest research in the fields of data mining, machine learning and decision-making. The international scientific conference “Intellectual Systems of Decision-Making and Problems of Computational Intelligence” was held in the Kherson region, Ukraine, from May 25 to 29, 2020. The papers are divided into three sections: “Analysis and Modeling of Complex Systems and Processes,” “Theoretical and Applied Aspects of Decision-Making Systems” and “Computational Intelligence and Inductive Modeling.” The book will be of interest to scientists and developers specialized in the fields of data mining, machine learning and decision-making systems.

Bigdata is one of the most demanding markets in the IT sector. If you are an administrator or a have a passion for knowing the internal configurations of Hadoop, then this book is for you. This book enables a professional to learn about Hadoop in terms of installation, configuration, and management. This book will help the reader to jumpstart with Hadoop frameworks, its eco-system components and slowly progress towards learning the administration part of Hadoop. The level of this book goes from beginner to intermediate with 70% hands-on exercises. Some of the techniques that you will learn include, • Installation and configuration of Hadoop cluster • Performing Hadoop Cluster Upgrade • Understanding and implementing HDFS Federation • Understanding and Implementing High Availability • Implementing HA on a Federated Cluster • Zookeeper CLI • Apache Hive Installation and Security • HBase Multi-master setup • Oozie installation, configuration and job submission • Setting up HDFS Quotas

• Setting up HDFS NFS gateway • Understanding and implementing rolling upgrade and much more.

Web Scraping techniques are getting more popular, since data is as valuable as oil in 21st century. Through this book get some key knowledge about using XPath, regEX; web scraping libraries for R like rvest and RSelenium technologies. Key Features Techniques, tools and frameworks for web scraping with R Scrape data effortlessly from a variety of websites Learn how to selectively choose the data to scrape, and build your dataset Book Description Web scraping is a technique to extract data from websites. It simulates the behavior of a website user to turn the website itself into a web service to retrieve or introduce new data. This book gives you all you need to get started with scraping web pages using R programming. You will learn about the rules of RegEx and Xpath, key components for scraping website data. We will show you web scraping techniques, methodologies, and frameworks. With this book's guidance, you will become comfortable with the tools to write and test RegEx and XPath rules. We will focus on examples of dynamic websites for scraping data and how to implement the techniques learned. You will learn how to collect URLs and then create XPath rules for your first web scraping script using rvest library. From the data you collect, you will be able to calculate the statistics and create R plots to visualize them. Finally, you will discover how to use Selenium drivers with R for more sophisticated scraping. You will create AWS instances and use R to connect a PostgreSQL database hosted on AWS.

By the end of the book, you will be sufficiently confident to create end-to-end web scraping systems using R. What you will learn Write and create regEX rules Write XPath rules to query your data Learn how web scraping methods work Use rvest to crawl web pages Store data retrieved from the web Learn the key uses of Rselenium to scrape data Who this book is for This book is for R programmers who want to get started quickly with web scraping, as well as data analysts who want to learn scraping using R. Basic knowledge of R is all you need to get started with this book.

"Apache Hadoop is helping drive the Big Data revolution. Now, its data processing has been completely overhauled: Apache Hadoop YARN provides resource management at data center scale and easier ways to create distributed applications that process petabytes of data. And now in Apache HadoopTM YARN, two Hadoop technical leaders show you how to develop new applications and adapt existing code to fully leverage these revolutionary advances." -- From the Amazon

Cloud computing is becoming the next revolution in the IT industry; providing central storage for internet data and services that have the potential to bring data transmission performance, security and privacy, data deluge, and inefficient architecture to the next level. Enabling the New Era of Cloud Computing: Data Security, Transfer, and Management discusses cloud computing as an emerging technology and its critical role in the IT industry upgrade and economic development in the future. This book is an essential resource for business decision makers, technology investors, architects and

engineers, and cloud consumers interested in the cloud computing future.
Navigate the world of data analysis, visualization, and machine learning with over 100 hands-on Scala recipes About This Book Implement Scala in your data analysis using features from Spark, Breeze, and Zeppelin Scale up your data anlytics infrastructure with practical recipes for Scala machine learning Recipes for every stage of the data analysis process, from reading and collecting data to distributed analytics Who This Book Is For This book shows data scientists and analysts how to leverage their existing knowledge of Scala for quality and scalable data analysis. What You Will Learn Familiarize and set up the Breeze and Spark libraries and use data structures Import data from a host of possible sources and create dataframes from CSV Clean, validate and transform data using Scala to pre-process numerical and string data Integrate quintessential machine learning algorithms using Scala stack Bundle and scale up Spark jobs by deploying them into a variety of cluster managers Run streaming and graph analytics in Spark to visualize data, enabling exploratory analysis In Detail This book will introduce you to the most popular Scala tools, libraries, and frameworks through practical recipes around loading, manipulating, and preparing your data. It will also help you explore and make sense of your data using stunning and insightfulvisualizations, and machine learning toolkits. Starting with introductory recipes on utilizing the Breeze and Spark libraries, get to grips withhow to import data from a host of possible sources and how to pre-process numerical, string, and date data. Next,

you'll get an understanding of concepts that will help you visualize data using the Apache Zeppelin and Bokeh bindings in Scala, enabling exploratory data analysis. iscover how to program quintessential machine learning algorithms using Spark ML library. Work through steps to scale your machine learning models and deploy them into a standalone cluster, EC2, YARN, and Mesos. Finally dip into the powerful options presented by Spark Streaming, and machine learning for streaming data, as well as utilizing Spark GraphX. Style and approach This book contains a rich set of recipes that covers the full spectrum of interesting data analysis tasks and will help you revolutionize your data analysis skills using Scala and Spark.

This book offers a comprehensible overview of Big Data Preprocessing, which includes a formal description of each problem. It also focuses on the most relevant proposed solutions. This book illustrates actual implementations of algorithms that helps the reader deal with these problems. This book stresses the gap that exists between big, raw data and the requirements of quality data that businesses are demanding. This is called Smart Data, and to achieve Smart Data the preprocessing is a key step, where the imperfections, integration tasks and other processes are carried out to eliminate superfluous information. The authors present the concept of Smart Data through data preprocessing in Big Data scenarios and connect it with the emerging paradigms of IoT and edge computing, where the end points generate Smart Data without completely relying on the cloud. Finally, this book provides some novel areas of study that are gathering a deeper attention on the Big Data preprocessing. Specifically, it considers the relation with Deep Learning (as of a technique that also relies in large volumes of

data), the difficulty of finding the appropriate selection and concatenation of preprocessing techniques applied and some other open problems. Practitioners and data scientists who work in this field, and want to introduce themselves to preprocessing in large data volume scenarios will want to purchase this book. Researchers that work in this field, who want to know which algorithms are currently implemented to help their investigations, may also be interested in this book.

Explore big data concepts, platforms, analytics, and their applications using the power of Hadoop 3 Key Features Learn Hadoop 3 to build effective big data analytics solutions on-premise and on cloud Integrate Hadoop with other big data tools such as R, Python, Apache Spark, and Apache Flink Exploit big data using Hadoop 3 with real-world examples Book Description Apache Hadoop is the most popular platform for big data processing, and can be combined with a host of other big data tools to build powerful analytics solutions. Big Data Analytics with Hadoop 3 shows you how to do just that, by providing insights into the software as well as its benefits with the help of practical examples. Once you have taken a tour of Hadoop 3's latest features, you will get an overview of HDFS, MapReduce, and YARN, and how they enable faster, more efficient big data processing. You will then move on to learning how to integrate Hadoop with the open source tools, such as Python and R, to analyze and visualize data and perform statistical computing on big data. As you get acquainted with all this, you will explore how to use Hadoop 3 with Apache Spark and Apache Flink for real-time data analytics and stream processing. In addition to this, you will understand how to use Hadoop to build analytics solutions on the cloud and an end-to-end pipeline to perform big data analysis using practical use cases. By the end of this book, you will be well-versed with the

analytical capabilities of the Hadoop ecosystem. You will be able to build powerful solutions to perform big data analytics and get insight effortlessly. What you will learn Explore the new features of Hadoop 3 along with HDFS, YARN, and MapReduce Get well-versed with the analytical capabilities of Hadoop ecosystem using practical examples Integrate Hadoop with R and Python for more efficient big data processing Learn to use Hadoop with Apache Spark and Apache Flink for real-time data analytics Set up a Hadoop cluster on AWS cloud Perform big data analytics on AWS using Elastic Map Reduce Who this book is for Big Data Analytics with Hadoop 3 is for you if you are looking to build high-performance analytics solutions for your enterprise or business using Hadoop 3's powerful features, or you're new to big data analytics. A basic understanding of the Java programming language is required.
Advanced analytics on your Big Data with latest Apache Spark 2.x About This Book An advanced guide with a combination of instructions and practical examples to extend the most up-to date Spark functionalities. Extend your data processing capabilities to process huge chunk of data in minimum time using advanced concepts in Spark. Master the art of real-time processing with the help of Apache Spark 2.x Who This Book Is For If you are a developer with some experience with Spark and want to strengthen your knowledge of how to get around in the world of Spark, then this book is ideal for you. Basic knowledge of Linux, Hadoop and Spark is assumed. Reasonable knowledge of Scala is expected. What You Will Learn Examine Advanced Machine Learning and DeepLearning with MLlib, SparkML, SystemML, H2O and DeepLearning4J Study highly optimised unified batch and real-time data processing using SparkSQL and Structured Streaming Evaluate large-scale Graph Processing and Analysis using GraphX and GraphFrames Apply Apache Spark in Elastic deployments using Jupyter

and Zeppelin Notebooks, Docker, Kubernetes and the IBM Cloud Understand internal details of cost based optimizers used in Catalyst, SystemML and GraphFrames Learn how specific parameter settings affect overall performance of an Apache Spark cluster Leverage Scala, R and python for your data science projects In Detail Apache Spark is an in-memory cluster-based parallel processing system that provides a wide range of functionalities such as graph processing, machine learning, stream processing, and SQL. This book aims to take your knowledge of Spark to the next level by teaching you how to expand Spark's functionality and implement your data flows and machine/deep learning programs on top of the platform. The book commences with an overview of the Spark ecosystem. It will introduce you to Project Tungsten and Catalyst, two of the major advancements of Apache Spark 2.x. You will understand how memory management and binary processing, cache-aware computation, and code generation are used to speed things up dramatically. The book extends to show how to incorporate H20, SystemML, and Deeplearning4j for machine learning, and Jupyter Notebooks and Kubernetes/Docker for cloud-based Spark. During the course of the book, you will learn about the latest enhancements to Apache Spark 2.x, such as interactive querying of live data and unifying DataFrames and Datasets. You will also learn about the updates on the APIs and how DataFrames and Datasets affect SQL, machine learning, graph processing, and streaming. You will learn to use Spark as a big data operating system, understand how to implement advanced analytics on the new APIs, and explore how easy it is to use Spark in day-to-day tasks. Style and approach This book is an extensive guide to Apache Spark modules and tools and shows how Spark's functionality can be extended for real-time processing and storage with worked examples.

Your software needs to leverage multiple cores, handle thousands of users and terabytes of data, and continue working in the face of both hardware and software failure. Concurrency and parallelism are the keys, and Seven Concurrency Models in Seven Weeks equips you for this new world. See how emerging technologies such as actors and functional programming address issues with traditional threads and locks development. Learn how to exploit the parallelism in your computer's GPU and leverage clusters of machines with MapReduce and Stream Processing. And do it all with the confidence that comes from using tools that help you write crystal clear, high-quality code. This book will show you how to exploit different parallel architectures to improve your code's performance, scalability, and resilience. You'll learn about seven concurrency models: threads and locks, functional programming, separating identity and state, actors, sequential processes, data parallelism, and the lambda architecture. Learn about the perils of traditional threads and locks programming and how to overcome them through careful design and by working with the standard library. See how actors enable software running on geographically distributed computers to collaborate, handle failure, and create systems that stay up 24/7/365. Understand why shared mutable state is the enemy of robust concurrent code, and see how functional programming together with technologies such as Software Transactional Memory (STM) and automatic parallelism help you tame it. You'll learn about the untapped potential within every GPU and how GPGPU software can unleash it. You'll see how to use MapReduce to harness massive clusters to solve previously intractable problems, and how, in concert with Stream Processing, big data can be tamed. With an understanding of the strengths and weaknesses of each of the different models and hardware architectures, you'll be empowered to tackle any problem with confidence. What You Need:

The example code can be compiled and executed on *nix, OS X, or Windows. Instructions on how to download the supporting build systems are given in each chapter.

This IBM® Redbooks® publication provides topics to help the technical community take advantage of the resilience, scalability, and performance of the IBM Power SystemsTM platform to implement or integrate an IBM Data Engine for Hadoop and Spark solution for analytics solutions to access, manage, and analyze data sets to improve business outcomes. This book documents topics to demonstrate and take advantage of the analytics strengths of the IBM POWER8® platform, the IBM analytics software portfolio, and selected third-party tools to help solve customer's data analytic workload requirements. This book describes how to plan, prepare, install, integrate, manage, and show how to use the IBM Data Engine for Hadoop and Spark solution to run analytic workloads on IBM POWER8. In addition, this publication delivers documentation to complement available IBM analytics solutions to help your data analytic needs. This publication strengthens the position of IBM analytics and big data solutions with a well-defined and documented deployment model within an IBM POWER8 virtualized environment so that customers have a planned foundation for security, scaling, capacity, resilience, and optimization for analytics workloads. This book is targeted at technical professionals (analytics consultants, technical support staff, IT Architects, and IT Specialists) that are responsible for delivering analytics solutions and support on IBM Power Systems.

Data is bigger, arrives faster, and comes in a variety of formats—and it all needs to be processed at scale for analytics or machine learning. But how can you process such varied workloads efficiently? Enter Apache Spark. Updated to include Spark 3.0, this second edition shows data engineers and data scientists why structure and unification in Spark matters.

Specifically, this book explains how to perform simple and complex data analytics and employ machine learning algorithms. Through step-by-step walk-throughs, code snippets, and notebooks, you'll be able to: Learn Python, SQL, Scala, or Java high-level Structured APIs Understand Spark operations and SQL Engine Inspect, tune, and debug Spark operations with Spark configurations and Spark UI Connect to data sources: JSON, Parquet, CSV, Avro, ORC, Hive, S3, or Kafka Perform analytics on batch and streaming data using Structured Streaming Build reliable data pipelines with open source Delta Lake and Spark Develop machine learning pipelines with MLlib and productionize models using MLflow

If you've been asked to maintain large and complex Hadoop clusters, this book is a must. Demand for operations-specific material has skyrocketed now that Hadoop is becoming the de facto standard for truly large-scale data processing in the data center. Eric Sammer, Principal Solution Architect at Cloudera, shows you the particulars of running Hadoop in production, from planning, installing, and configuring the system to providing ongoing maintenance. Rather than run through all possible scenarios, this pragmatic operations guide calls out what works, as demonstrated in critical deployments. Get a high-level overview of HDFS and MapReduce: why they exist and how they work Plan a Hadoop deployment, from hardware and OS selection to network requirements Learn setup and configuration details with a list of critical properties Manage resources by sharing a cluster across multiple groups Get a runbook of the most common cluster maintenance tasks Monitor Hadoop clusters—and learn troubleshooting with the help of real-world war stories Use basic tools and techniques to handle backup and catastrophic failure

Over 70 recipes to help you use Apache Spark as your single big data computing platform and

master its libraries About This Book This book contains recipes on how to use Apache Spark as a unified compute engine Cover how to connect various source systems to Apache Spark Covers various parts of machine learning including supervised/unsupervised learning & recommendation engines Who This Book Is For This book is for data engineers, data scientists, and those who want to implement Spark for real-time data processing. Anyone who is using Spark (or is planning to) will benefit from this book. The book assumes you have a basic knowledge of Scala as a programming language. What You Will Learn Install and configure Apache Spark with various cluster managers & on AWS Set up a development environment for Apache Spark including Databricks Cloud notebook Find out how to operate on data in Spark with schemas Get to grips with real-time streaming analytics using Spark Streaming & Structured Streaming Master supervised learning and unsupervised learning using MLlib Build a recommendation engine using MLlib Graph processing using GraphX and GraphFrames libraries Develop a set of common applications or project types, and solutions that solve complex big data problems In Detail While Apache Spark 1.x gained a lot of traction and adoption in the early years, Spark 2.x delivers notable improvements in the areas of API, schema awareness, Performance, Structured Streaming, and simplifying building blocks to build better, faster, smarter, and more accessible big data applications. This book uncovers all these features in the form of structured recipes to analyze and mature large and complex sets of data. Starting with installing and configuring Apache Spark with various cluster managers, you will learn to set up development environments. Further on, you will be introduced to working with RDDs, DataFrames and Datasets to operate on schema aware data, and real-time streaming with various sources such as Twitter Stream and Apache Kafka. You will also

work through recipes on machine learning, including supervised learning, unsupervised learning & recommendation engines in Spark. Last but not least, the final few chapters delve deeper into the concepts of graph processing using GraphX, securing your implementations, cluster optimization, and troubleshooting. Style and approach This book is packed with intuitive recipes supported with line-by-line explanations to help you understand Spark 2.x's real-time processing capabilities and deploy scalable big data solutions. This is a valuable resource for data scientists and those working on large-scale data projects.

Over 170 advanced recipes to search, analyze, deploy, manage, and monitor data effectively with Elasticsearch 5.x About This Book Deploy and manage simple Elasticsearch nodes as well as complex cluster topologies Write native plugins to extend the functionalities of Elasticsearch 5.x to boost your business Packed with clear, step-by-step recipes to walk you through the capabilities of Elasticsearch 5.x Who This Book Is For If you are a developer who wants to get the most out of Elasticsearch for advanced search and analytics, this is the book for you. Some understanding of JSON is expected. If you want to extend Elasticsearch, understanding of Java and related technologies is also required. What You Will Learn Choose the best Elasticsearch cloud topology to deploy and power it up with external plugins Develop tailored mapping to take full control of index steps Build complex queries through managing indices and documents Optimize search results through executing analytics aggregations Monitor the performance of the cluster and nodes Install Kibana to monitor cluster and extend Kibana for plugins Integrate Elasticsearch in Java, Scala, Python and Big Data applications In Detail Elasticsearch is a Lucene-based distributed search server that allows users to index and search unstructured content with petabytes of data. This book is your one-stop guide to master

the complete Elasticsearch ecosystem. We'll guide you through comprehensive recipes on what's new in Elasticsearch 5.x, showing you how to create complex queries and analytics, and perform index mapping, aggregation, and scripting. Further on, you will explore the modules of Cluster and Node monitoring and see ways to back up and restore a snapshot of an index. You will understand how to install Kibana to monitor a cluster and also to extend Kibana for plugins. Finally, you will also see how you can integrate your Java, Scala, Python, and Big Data applications such as Apache Spark and Pig with Elasticsearch, and add enhanced functionalities with custom plugins. By the end of this book, you will have an in-depth knowledge of the implementation of the Elasticsearch architecture and will be able to manage data efficiently and effectively with Elasticsearch. Style and approach This book follows a problem-solution approach to effectively use and manage Elasticsearch. Each recipe focuses on a particular task at hand, and is explained in a very simple, easy to understand manner.
This book constitutes the proceedings of the 15th IFIP TC8 International Conference on Computer Information Systems and Industrial Management, CISIM 2016, held in Vilnius, Lithuania, in September 2016. The 63 regular papers presented together with 1 inivted paper and 5 keynotes in this volume were carefully reviewed and selected from about 89 submissions. The main topics covered are rough set methods for big data analytics; images, visualization, classification; optimization, tuning; scheduling in manufacturing and other applications; algorithms; decisions; intelligent distributed systems; and biometrics, identification, security.
Over 90 hands-on recipes to help you learn and master the intricacies of Apache Hadoop 2.X, YARN, Hive, Pig, Oozie, Flume, Sqoop, Apache Spark, and Mahout About This Book

Implement outstanding Machine Learning use cases on your own analytics models and processes. Solutions to common problems when working with the Hadoop ecosystem. Step-by-step implementation of end-to-end big data use cases. Who This Book Is For Readers who have a basic knowledge of big data systems and want to advance their knowledge with hands-on recipes. What You Will Learn Installing and maintaining Hadoop 2.X cluster and its ecosystem. Write advanced Map Reduce programs and understand design patterns. Advanced Data Analysis using the Hive, Pig, and Map Reduce programs. Import and export data from various sources using Sqoop and Flume. Data storage in various file formats such as Text, Sequential, Parquet, ORC, and RC Files. Machine learning principles with libraries such as Mahout Batch and Stream data processing using Apache Spark In Detail Big data is the current requirement. Most organizations produce huge amount of data every day. With the arrival of Hadoop-like tools, it has become easier for everyone to solve big data problems with great efficiency and at minimal cost. Grasping Machine Learning techniques will help you greatly in building predictive models and using this data to make the right decisions for your organization. Hadoop Real World Solutions Cookbook gives readers insights into learning and mastering big data via recipes. The book not only clarifies most big data tools in the market but also provides best practices for using them. The book provides recipes that are based on the latest versions of Apache Hadoop 2.X, YARN, Hive, Pig, Sqoop, Flume, Apache Spark, Mahout and many more such ecosystem tools. This real-world-solution cookbook is packed with handy recipes you can apply to your own everyday issues. Each chapter provides in-depth recipes that can be referenced easily. This book provides detailed practices on the latest technologies such as YARN and Apache Spark. Readers will be able to consider themselves

as big data experts on completion of this book. This guide is an invaluable tutorial if you are planning to implement a big data warehouse for your business. Style and approach An easy-to-follow guide that walks you through world of big data. Each tool in the Hadoop ecosystem is explained in detail and the recipes are placed in such a manner that readers can implement them sequentially. Plenty of reference links are provided for advanced reading.

If you are ready to dive into the MapReduce framework for processing large datasets, this practical book takes you step by step through the algorithms and tools you need to build distributed MapReduce applications with Apache Hadoop or Apache Spark. Each chapter provides a recipe for solving a massive computational problem, such as building a recommendation system. You'll learn how to implement the appropriate MapReduce solution with code that you can use in your projects. Dr. Mahmoud Parsian covers basic design patterns, optimization techniques, and data mining and machine learning solutions for problems in bioinformatics, genomics, statistics, and social network analysis. This book also includes an overview of MapReduce, Hadoop, and Spark. Topics include: Market basket analysis for a large set of transactions Data mining algorithms (K-means, KNN, and Naive Bayes) Using huge genomic data to sequence DNA and RNA Naive Bayes theorem and Markov chains for data and market prediction Recommendation algorithms and pairwise document similarity Linear regression, Cox regression, and Pearson correlation Allelic frequency and mining DNA Social network analysis (recommendation systems, counting triangles, sentiment analysis)

Pro Microsoft HDInsight is a complete guide to deploying and using Apache Hadoop on the Microsoft Windows Azure Platforms. The information in this book enables you to process

enormous volumes of structured as well as non-structured data easily using HDInsight, which is Microsoft's own distribution of Apache Hadoop. Furthermore, the blend of Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) offerings available through Windows Azure lets you take advantage of Hadoop's processing power without the worry of creating, configuring, maintaining, or managing your own cluster. With the data explosion that is soon to happen, the open source Apache Hadoop Framework is gaining traction, and it benefits from a huge ecosystem that has risen around the core functionalities of the Hadoop distributed file system (HDFS™) and Hadoop Map Reduce. Pro Microsoft HDInsight equips you with the knowledge, confidence, and technique to configure and manage this ecosystem on Windows Azure. The book is an excellent choice for anyone aspiring to be a data scientist or data engineer, putting you a step ahead in the data mining field. Guides you through installation and configuration of an HDInsight cluster on Windows Azure Provides clear examples of configuring and executing Map Reduce jobs Helps you consume data and diagnose errors from the Windows Azure HDInsight Service

This book takes its reader on a journey through Apache Giraph, a popular distributed graph processing platform designed to bring the power of big data processing to graph data. Designed as a step-by-step self-study guide for everyone interested in large-scale graph processing, it describes the fundamental abstractions of the system, its programming models and various techniques for using the system to process graph data at scale, including the implementation of several popular and advanced graph analytics algorithms. The book is organized as follows: Chapter 1 starts by providing a general background of the big data phenomenon and a general introduction to the Apache Giraph system, its abstraction,

programming model and design architecture. Next, chapter 2 focuses on Giraph as a platform and how to use it. Based on a sample job, even more advanced topics like monitoring the Giraph application lifecycle and different methods for monitoring Giraph jobs are explained. Chapter 3 then provides an introduction to Giraph programming, introduces the basic Giraph graph model and explains how to write Giraph programs. In turn, Chapter 4 discusses in detail the implementation of some popular graph algorithms including PageRank, connected components, shortest paths and triangle closing. Chapter 5 focuses on advanced Giraph programming, discussing common Giraph algorithmic optimizations, tunable Giraph configurations that determine the system's utilization of the underlying resources, and how to write a custom graph input and output format. Lastly, chapter 6 highlights two systems that have been introduced to tackle the challenge of large scale graph processing, GraphX and GraphLab, and explains the main commonalities and differences between these systems and Apache Giraph. This book serves as an essential reference guide for students, researchers and practitioners in the domain of large scale graph processing. It offers step-by-step guidance, with several code examples and the complete source code available in the related github repository. Students will find a comprehensive introduction to and hands-on practice with tackling large scale graph processing problems using the Apache Giraph system, while researchers will discover thorough coverage of the emerging and ongoing advancements in big graph processing systems.

Many corporations are finding that the size of their data sets are outgrowing the capability of their systems to store and process them. The data is becoming too big to manage and use with traditional tools. The solution: implementing a big data system. As Big Data Made Easy: A

Working Guide to the Complete Hadoop Toolset shows, Apache Hadoop offers a scalable, fault-tolerant system for storing and processing data in parallel. It has a very rich toolset that allows for storage (Hadoop), configuration (YARN and ZooKeeper), collection (Nutch and Solr), processing (Storm, Pig, and Map Reduce), scheduling (Oozie), moving (Sqoop and Avro), monitoring (Chukwa, Ambari, and Hue), testing (Big Top), and analysis (Hive). The problem is that the Internet offers IT pros wading into big data many versions of the truth and some outright falsehoods born of ignorance. What is needed is a book just like this one: a wide-ranging but easily understood set of instructions to explain where to get Hadoop tools, what they can do, how to install them, how to configure them, how to integrate them, and how to use them successfully. And you need an expert who has worked in this area for a decade—someone just like author and big data expert Mike Frampton. Big Data Made Easy approaches the problem of managing massive data sets from a systems perspective, and it explains the roles for each project (like architect and tester, for example) and shows how the Hadoop toolset can be used at each system stage. It explains, in an easily understood manner and through numerous examples, how to use each tool. The book also explains the sliding scale of tools available depending upon data size and when and how to use them. Big Data Made Easy shows developers and architects, as well as testers and project managers, how to: Store big data Configure big data Process big data Schedule processes Move data among SQL and NoSQL systems Monitor data Perform big data analytics Report on big data processes and projects Test big data systems Big Data Made Easy also explains the best part, which is that this toolset is free. Anyone can download it and—with the help of this book—start to use it within a day. With the skills this book will teach you under your belt, you will add value to

your company or client immediately, not to mention your career.

This book constitutes the proceedings of the 14th International Conference on Parallel Computing Technologies, PaCT 2017, held in Nizhny Novgorod, Russia, in September 2017. The 25 full papers and 24 short papers presented were carefully reviewed and selected from 93 submissions. The papers are organized in topical sections on mainstream parallel computing, parallel models and algorithms in numerical computation, cellular automata and discrete event systems, organization of parallel computation, parallel computing applications.

Data in all domains is getting bigger. How can you work with it efficiently? Recently updated for Spark 1.3, this book introduces Apache Spark, the open source cluster computing system that makes data analytics fast to write and fast to run. With Spark, you can tackle big datasets quickly through simple APIs in Python, Java, and Scala. This edition includes new information on Spark SQL, Spark Streaming, setup, and Maven coordinates. Written by the developers of Spark, this book will have data scientists and engineers up and running in no time. You'll learn how to express parallel jobs with just a few lines of code, and cover applications from simple batch jobs to stream processing and machine learning. Quickly dive into Spark capabilities such as distributed datasets, in-memory caching, and the interactive shell Leverage Spark's powerful built-in libraries, including Spark SQL, Spark Streaming, and MLlib Use one programming paradigm instead of mixing and matching tools like Hive, Hadoop, Mahout, and Storm Learn how to deploy interactive, batch, and streaming applications Connect to data sources including HDFS, Hive, JSON, and S3 Master advanced topics like data partitioning and shared variables

Hadoop has changed the way large data sets are analyzed, stored, transferred, and

processed. At such low cost, it provides benefits like supports partial failure, fault tolerance, consistency, scalability, flexible schema, and so on. It also supports cloud computing. More and more number of individuals are looking forward to mastering their Hadoop skills. While initiating with Hadoop, most users are unsure about how to proceed with Hadoop. They are not aware of what are the pre-requisite or data structure they should be familiar with. Or How to make the most efficient use of Hadoop and its ecosystem. To help them with all these queries and other issues this e-book is designed. The book gives insights into many of Hadoop libraries and packages that are not known to many Big data Analysts and Architects. The e-book also tells you about Hadoop MapReduce and HDFS. The example in the e-book is well chosen and demonstrates how to control Hadoop ecosystem through various shell commands. With this book, users will gain expertise in Hadoop technology and its related components. The book leverages you with the best Hadoop content with the lowest price range. After going through this book, you will also acquire knowledge on Hadoop Security required for Hadoop Certifications like CCAH and CCDH. It is a definite guide to Hadoop. Table Of Content Chapter 1: What Is Big Data 1. Examples Of 'Big Data' 2. Categories Of 'Big Data' 3. Characteristics Of 'Big Data' 4. Advantages Of Big Data Processing Chapter 2: Introduction to Hadoop 1. Components of Hadoop 2. Features Of 'Hadoop' 3. Network Topology In Hadoop Chapter 3: Hadoop Installation Chapter 4: HDFS 1. Read Operation 2. Write Operation 3. Access HDFS using JAVA API 4. Access HDFS Using COMMAND-LINE INTERFACE Chapter 5: Mapreduce 1. How MapReduce works 2. How MapReduce Organizes Work? Chapter 6: First Program 1. Understanding MapReducer Code 2. Explanation of SalesMapper Class 3. Explanation of SalesCountryReducer Class 4. Explanation of SalesCountryDriver Class Chapter 7: Counters

& Joins In MapReduce 1. Two types of counters 2. MapReduce Join Chapter 8: MapReduce Hadoop Program To Join Data Chapter 9: Flume and Sqoop 1. What is SQOOP in Hadoop? 2. What is FLUME in Hadoop? 3. Some Important features of FLUME Chapter 10: Pig 1. Introduction to PIG 2. Create your First PIG Program 3. PART 1) Pig Installation 4. PART 2) Pig Demo Chapter 11: OOZIE 1. What is OOZIE? 2. How does OOZIE work? 3. Example Workflow Diagram 4. Oozie workflow application 5. Why use Oozie? 6. FEATURES OF OOZIE

Over 150 recipes to design and optimize large scale Apache Cassandra deployments.

Due to increasing globalization and the explosion of media available on the Internet, computer techniques to organize, classify, and find desired media are becoming more and more relevant. One such technique to extract semantic information from multimedia data sources is Multimedia Information Retrieval (MMIR or MIR). MIR is a broad area covering both structural issues and intelligent content analysis and retrieval. These aspects must be integrated into a seamless whole, which involves expertise from a wide variety of fields. This book presents recent applications of MIR for content-based image retrieval, bioinformation analysis and processing, forensic multimedia retrieval techniques, and audio and music classification.

Scala will be a valuable tool to have on hand during your data science journey for everything from data cleaning to cutting-edge machine learning About This Book Build data science and data engineering solutions with ease An in-depth look at each stage of the data analysis process — from reading and collecting data to distributed analytics Explore a broad variety of data processing, machine learning, and genetic algorithms

through diagrams, mathematical formulations, and source code Who This Book Is For This learning path is perfect for those who are comfortable with Scala programming and now want to enter the field of data science. Some knowledge of statistics is expected. What You Will Learn Transfer and filter tabular data to extract features for machine learning Read, clean, transform, and write data to both SQL and NoSQL databases Create Scala web applications that couple with JavaScript libraries such as D3 to create compelling interactive visualizations Load data from HDFS and HIVE with ease Run streaming and graph analytics in Spark for exploratory analysis Bundle and scale up Spark jobs by deploying them into a variety of cluster managers Build dynamic workflows for scientific computing Leverage open source libraries to extract patterns from time series Master probabilistic models for sequential data In Detail Scala is especially good for analyzing large sets of data as the scale of the task doesn't have any significant impact on performance. Scala's powerful functional libraries can interact with databases and build scalable frameworks — resulting in the creation of robust data pipelines. The first module introduces you to Scala libraries to ingest, store, manipulate, process, and visualize data. Using real world examples, you will learn how to design scalable architecture to process and model data — starting from simple concurrency constructs and progressing to actor systems and Apache Spark. After this, you will also learn how to build interactive visualizations with web frameworks. Once you have become familiar with all the tasks involved in data science, you will explore data

analytics with Scala in the second module. You'll see how Scala can be used to make sense of data through easy to follow recipes. You will learn about Bokeh bindings for exploratory data analysis and quintessential machine learning with algorithms with Spark ML library. You'll get a sufficient understanding of Spark streaming, machine learning for streaming data, and Spark graphX. Armed with a firm understanding of data analysis, you will be ready to explore the most cutting-edge aspect of data science — machine learning. The final module teaches you the A to Z of machine learning with Scala. You'll explore Scala for dependency injections and implicits, which are used to write machine learning algorithms. You'll also explore machine learning topics such as clustering, dimentionality reduction, Naive Bayes, Regression models, SVMs, neural networks, and more. This learning path combines some of the best that Packt has to offer into one complete, curated package. It includes content from the following Packt products: Scala for Data Science, Pascal Bugnion Scala Data Analysis Cookbook, Arun Manivannan Scala for Machine Learning, Patrick R. Nicolas Style and approach A complete package with all the information necessary to start building useful data engineering and data science solutions straight away. It contains a diverse set of recipes that cover the full spectrum of interesting data analysis tasks and will help you revolutionize your data analysis skills using Scala.

Dig deep into the data with a hands-on guide to machinelearning Machine Learning: Hands-On for Developers and TechnicalProfessionals provides hands-on instruction

and fully-codedworking examples for the most common machine learning techniquesused by developers and technical professionals. The book contains abreakdown of each ML variant, explaining how it works and how it isused within certain industries, allowing readers to incorporate thepresented techniques into their own work as they follow along. Acore tenant of machine learning is a strong focus on datapreparation, and a full exploration of the various types oflearning algorithms illustrates how the proper tools can help anydeveloper extract information and insights from existing data. Thebook includes a full complement of Instructor's Materials tofacilitate use in the classroom, making this resource useful forstudents and as a professional reference. At its core, machine learning is a mathematical, algorithm-basedtechnology that forms the basis of historical data mining andmodern big data science. Scientific analysis of big data requires aworking knowledge of machine learning, which forms predictionsbased on known properties learned from training data. MachineLearning is an accessible, comprehensive guide for thenon-mathematician, providing clear guidance that allows readersto: Learn the languages of machine learning including Hadoop,Mahout, and Weka Understand decision trees, Bayesian networks, and artificialneural networks Implement Association Rule, Real Time, and Batch learning Develop a strategic plan for safe, effective, and efficientmachine learning By learning to construct a system that can learn from data,readers can increase their utility across industries. Machinelearning sits at the core of deep dive

data analysis andvisualization, which is increasingly in demand as companiesdiscover the goldmine hiding in their existing data. For the techprofessional involved in data science, Machine Learning:Hands-On for Developers and Technical Professionals providesthe skills and techniques required to dig deeper.

A comprehensive guide to mastering the most advanced Hadoop 3 concepts Key Features Get to grips with the newly introduced features and capabilities of Hadoop 3 Crunch and process data using MapReduce, YARN, and a host of tools within the Hadoop ecosystem Sharpen your Hadoop skills with real-world case studies and code Book Description Apache Hadoop is one of the most popular big data solutions for distributed storage and for processing large chunks of data. With Hadoop 3, Apache promises to provide a high-performance, more fault-tolerant, and highly efficient big data processing platform, with a focus on improved scalability and increased efficiency. With this guide, you'll understand advanced concepts of the Hadoop ecosystem tool. You'll learn how Hadoop works internally, study advanced concepts of different ecosystem tools, discover solutions to real-world use cases, and understand how to secure your cluster. It will then walk you through HDFS, YARN, MapReduce, and Hadoop 3 concepts. You'll be able to address common challenges like using Kafka efficiently, designing low latency, reliable message delivery Kafka systems, and handling high data volumes. As you advance, you'll discover how to address major challenges when building an enterprise-grade messaging system, and how to use

different stream processing systems along with Kafka to fulfil your enterprise goals. By the end of this book, you'll have a complete understanding of how components in the Hadoop ecosystem are effectively integrated to implement a fast and reliable data pipeline, and you'll be equipped to tackle a range of real-world problems in data pipelines. What you will learn Gain an in-depth understanding of distributed computing using Hadoop 3 Develop enterprise-grade applications using Apache Spark, Flink, and more Build scalable and high-performance Hadoop data pipelines with security, monitoring, and data governance Explore batch data processing patterns and how to model data in Hadoop Master best practices for enterprises using, or planning to use, Hadoop 3 as a data platform Understand security aspects of Hadoop, including authorization and authentication Who this book is for If you want to become a big data professional by mastering the advanced concepts of Hadoop, this book is for you. You'll also find this book useful if you're a Hadoop professional looking to strengthen your knowledge of the Hadoop ecosystem. Fundamental knowledge of the Java programming language and basics of Hadoop is necessary to get started with this book. A fast paced guide that will help you learn about Apache Hadoop 3 and its ecosystem Key Features Set up, configure and get started with Hadoop to get useful insights from large data sets Work with the different components of Hadoop such as MapReduce, HDFS and YARN Learn about the new features introduced in Hadoop 3 Book Description Apache Hadoop is a widely used distributed data platform. It enables large

datasets to be efficiently processed instead of using one large computer to store and process the data. This book will get you started with the Hadoop ecosystem, and introduce you to the main technical topics, including MapReduce, YARN, and HDFS. The book begins with an overview of big data and Apache Hadoop. Then, you will set up a pseudo Hadoop development environment and a multi-node enterprise Hadoop cluster. You will see how the parallel programming paradigm, such as MapReduce, can solve many complex data processing problems. The book also covers the important aspects of the big data software development lifecycle, including quality assurance and control, performance, administration, and monitoring. You will then learn about the Hadoop ecosystem, and tools such as Kafka, Sqoop, Flume, Pig, Hive, and HBase. Finally, you will look at advanced topics, including real time streaming using Apache Storm, and data analytics using Apache Spark. By the end of the book, you will be well versed with different configurations of the Hadoop 3 cluster. What you will learn Store and analyze data at scale using HDFS, MapReduce and YARN Install and configure Hadoop 3 in different modes Use Yarn effectively to run different applications on Hadoop based platform Understand and monitor how Hadoop cluster is managed Consume streaming data using Storm, and then analyze it using Spark Explore Apache Hadoop ecosystem components, such as Flume, Sqoop, HBase, Hive, and Kafka Who this book is for Aspiring Big Data professionals who want to learn the essentials of Hadoop 3 will find this book to be useful. Existing Hadoop users who want to get up to

speed with the new features introduced in Hadoop 3 will also benefit from this book. Having knowledge of Java programming will be an added advantage.

Pro Apache Hadoop, Second Edition brings you up to speed on Hadoop – the framework of big data. Revised to cover Hadoop 2.0, the book covers the very latest developments such as YARN (aka MapReduce 2.0), new HDFS high-availability features, and increased scalability in the form of HDFS Federations. All the old content has been revised too, giving the latest on the ins and outs of MapReduce, cluster design, the Hadoop Distributed File System, and more. This book covers everything you need to build your first Hadoop cluster and begin analyzing and deriving value from your business and scientific data. Learn to solve big-data problems the MapReduce way, by breaking a big problem into chunks and creating small-scale solutions that can be flung across thousands upon thousands of nodes to analyze large data volumes in a short amount of wall-clock time. Learn how to let Hadoop take care of distributing and parallelizing your software—you just focus on the code; Hadoop takes care of the rest. Covers all that is new in Hadoop 2.0 Written by a professional involved in Hadoop since day one Takes you quickly to the seasoned pro level on the hottest cloud-computing framework

Learn how to use the Apache Hadoop projects, including MapReduce, HDFS, Apache Hive, Apache HBase, Apache Kafka, Apache Mahout, and Apache Solr. From setting up the environment to running sample applications each chapter in this book is a

practical tutorial on using an Apache Hadoop ecosystem project. While several books on Apache Hadoop are available, most are based on the main projects, MapReduce and HDFS, and none discusses the other Apache Hadoop ecosystem projects and how they all work together as a cohesive big data development platform. What You Will Learn: Set up the environment in Linux for Hadoop projects using Cloudera Hadoop Distribution CDH 5 Run a MapReduce job Store data with Apache Hive, and Apache HBase Index data in HDFS with Apache Solr Develop a Kafka messaging system Stream Logs to HDFS with Apache Flume Transfer data from MySQL database to Hive, HDFS, and HBase with Sqoop Create a Hive table over Apache Solr Develop a Mahout User Recommender System Who This Book Is For: Apache Hadoop developers. Pre-requisite knowledge of Linux and some knowledge of Hadoop is required.

This book constitutes the thoroughly refereed post-conference proceedings of the 8th TPC Technology Conference, on Performance Evaluation and Benchmarking, TPCTC 2017, held in conjunction with the43rd International Conference on Very Large Databases (VLDB 2017) in August/September 2017. The 12 papers presented were carefully reviewed and selected from numeroussubmissions. The TPC remains committed to developing new benchmark standards to keep pace with these rapid changes in technology.

As more corporations turn to Hadoop to store and process their most valuable data, the risk of a potential breach of those systems increases exponentially. This practical book not only

shows Hadoop administrators and security architects how to protect Hadoop data from unauthorized access, it also shows how to limit the ability of an attacker to corrupt or modify data in the event of a security breach. Authors Ben Spivey and Joey Echeverria provide in-depth information about the security features available in Hadoop, and organize them according to common computer security concepts. You'll also get real-world examples that demonstrate how you can apply these concepts to your use cases. Understand the challenges of securing distributed systems, particularly Hadoop Use best practices for preparing Hadoop cluster hardware as securely as possible Get an overview of the Kerberos network authentication protocol Delve into authorization and accounting principles as they apply to Hadoop Learn how to use mechanisms to protect data in a Hadoop cluster, both in transit and at rest Integrate Hadoop data ingest into enterprise-wide security architecture Ensure that security architecture reaches all the way to end-user access

The authors provide an understanding of big data and MapReduce by clearly presenting the basic terminologies and concepts. They have employed over 100 illustrations and many worked-out examples to convey the concepts and methods used in big data, the inner workings of MapReduce, and single node/multi-node installation on physical/virtual machines. This book covers almost all the necessary information on Hadoop MapReduce for most online certification exams. Upon completing this book, readers will find it easy to understand other big data processing tools such as Spark, Storm, etc. Ultimately, readers will be able to: • understand what big data is and the factors that are involved • understand the inner workings of MapReduce, which is essential for certification exams • learn the features and weaknesses of MapReduce • set up Hadoop clusters with 100s of physical/virtual machines • create a

virtual machine in AWS • write MapReduce with Eclipse in a simple way • understand other big data processing tools and their applications

Pro MongoDB Development is about MongoDB, a NoSQL database based on the BSON (binary JSON) document model. The book discusses all aspects of using MongoDB in web applications: Java, PHP, Ruby, JavaScript are the most commonly used programming/scripting languages and the book discusses accessing MongoDB database with these languages. The book also discusses using Java EE frameworks Kundera and Spring Data with MongoDB. As NoSQL databases are commonly used with the Hadoop ecosystem the book also discusses using MongoDB with Apache Hive. Migration from other NoSQL databases (Apache Cassandra and Couchbase) and from relational databases (Oracle Database) is also discussed. What You'll Learn: How to use a Java client and MongoDB shell How to use MongoDB with PHP, Ruby, and Node.js as well How to migrate Apache Cassandra tables to MongoDB documents; Couchbase to MongoDB; and transferring data between Oracle and MongoDB How to use Kundera, Spring Data, and Spring XD with MongoDB How to load MongoDB data into Oracle Database and integrating MongoDB with Oracle Database in Oracle Data Integrator Audience: The target audience of the book is NoSQL database developers. Target audience includes Java, PHP and Ruby developers. The book is suitable for an intermediate level course in NoSQL database.

Build efficient data flow and machine learning programs with this flexible, multi-functional open-source cluster-computing framework Key Features Master the art of real-time big data processing and machine learning Explore a wide range of use-cases to analyze large data Discover ways to optimize your work by using many features of Spark 2.x and Scala Book

Description Apache Spark is an in-memory, cluster-based data processing system that provides a wide range of functionalities such as big data processing, analytics, machine learning, and more. With this Learning Path, you can take your knowledge of Apache Spark to the next level by learning how to expand Spark's functionality and building your own data flow and machine learning programs on this platform. You will work with the different modules in Apache Spark, such as interactive querying with Spark SQL, using DataFrames and datasets, implementing streaming analytics with Spark Streaming, and applying machine learning and deep learning techniques on Spark using MLlib and various external tools. By the end of this elaborately designed Learning Path, you will have all the knowledge you need to master Apache Spark, and build your own big data processing and analytics pipeline quickly and without any hassle. This Learning Path includes content from the following Packt products: Mastering Apache Spark 2.x by Romeo Kienzler Scala and Spark for Big Data Analytics by Md. Rezaul Karim, Sridhar Alla Apache Spark 2.x Machine Learning Cookbook by Siamak Amirghodsi, Meenakshi Rajendran, Broderick Hall, Shuen MeiCookbook What you will learn Get to grips with all the features of Apache Spark 2.x Perform highly optimized real-time big data processing Use ML and DL techniques with Spark MLlib and third-party tools Analyze structured and unstructured data using SparkSQL and GraphX Understand tuning, debugging, and monitoring of big data applications Build scalable and fault-tolerant streaming applications Develop scalable recommendation engines Who this book is for If you are an intermediate-level Spark developer looking to master the advanced capabilities and use-cases of Apache Spark 2.x, this Learning Path is ideal for you. Big data professionals who want to learn how to integrate and use the features of Apache Spark and build a strong big data pipeline will also

find this Learning Path useful. To grasp the concepts explained in this Learning Path, you must know the fundamentals of Apache Spark and Scala.

Big Data: Principles and Paradigms captures the state-of-the-art research on the architectural aspects, technologies, and applications of Big Data. The book identifies potential future directions and technologies that facilitate insight into numerous scientific, business, and consumer applications. To help realize Big Data's full potential, the book addresses numerous challenges, offering the conceptual and technological solutions for tackling them. These challenges include life-cycle data management, large-scale storage, flexible processing infrastructure, data modeling, scalable machine learning, data analysis algorithms, sampling techniques, and privacy and ethical issues. Covers computational platforms supporting Big Data applications Addresses key principles underlying Big Data computing Examines key developments supporting next generation Big Data platforms Explores the challenges in Big Data computing and ways to overcome them Contains expert contributors from both academia and industry