

Single Node Setup Apache Hadoop

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.

If you are an undergraduate or graduate student, a beginner to algorithmic development and research,

or a software developer in the financial industry who is interested in using Python for quantitative methods in finance, this is the book for you. It would be helpful to have a bit of familiarity with basic Python usage, but no prior experience is required.

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

This book is an example-based tutorial that deals

with Optimizing Hadoop for MapReduce job performance. If you are a Hadoop administrator, developer, MapReduce user, or beginner, this book is the best choice available if you wish to optimize your clusters and applications. Having prior knowledge of creating MapReduce applications is not necessary, but will help you better understand the concepts and snippets of MapReduce class template code.

This book gathers selected papers presented at the 7th International Conference on Innovations in Electronics and Communication Engineering, held at Guru Nanak Institutions in Hyderabad, India. It highlights contributions by researchers, technocrats and experts regarding the latest technologies in electronic and communication engineering, and addresses various aspects of communication engineering, including signal processing, VLSI design, embedded systems, wireless communications, and electronics and communications in general. Covering cutting-edge technologies, the book offers a valuable resource, especially for young researchers.

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.

Summary Mahout in Action is a hands-on introduction to machine learning with Apache Mahout. Following real-world examples, the book presents practical use cases and then illustrates how Mahout can be applied to solve them. Includes a free audio- and video-enhanced ebook. About the Technology A computer system that learns and adapts as it collects data can be really powerful. Mahout, Apache's open source machine learning project, captures the core algorithms of recommendation systems, classification, and clustering in ready-to-use, scalable libraries. With Mahout, you can immediately apply to your own projects the machine learning techniques that drive Amazon, Netflix, and others. About this Book This book covers machine learning using Apache Mahout. Based on experience with real-world applications, it introduces practical use cases and

illustrates how Mahout can be applied to solve them. It places particular focus on issues of scalability and how to apply these techniques against large data sets using the Apache Hadoop framework. This book is written for developers familiar with Java -- no prior experience with Mahout is assumed. Owners of a Manning pBook purchased anywhere in the world can download a free eBook from manning.com at any time. They can do so multiple times and in any or all formats available (PDF, ePub or Kindle). To do so, customers must register their printed copy on Manning's site by creating a user account and then following instructions printed on the pBook registration insert at the front of the book.

What's Inside

- Use group data to make individual recommendations
- Find logical clusters within your data
- Filter and refine with on-the-fly classification
- Free audio and video extras
- Table of Contents
- Meet Apache Mahout

PART 1 RECOMMENDATIONS

- Introducing recommenders
- Representing recommender data
- Making recommendations
- Taking recommenders to production
- Distributing recommendation computations

PART 2 CLUSTERING

- Introduction to clustering
- Representing data
- Clustering algorithms in Mahout
- Evaluating and improving clustering quality
- Taking clustering to production
- Real-world applications of clustering

PART 3 CLASSIFICATION

- Introduction to classification
- Training a classifier
- Evaluating and

tuning a classifier Deploying a classifier Case study:
Shop It To Me

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

If your organization is about to enter the world of big data, you not only need to decide whether Apache Hadoop is the right platform to use, but also which of

its many components are best suited to your task. This field guide makes the exercise manageable by breaking down the Hadoop ecosystem into short, digestible sections. You'll quickly understand how Hadoop's projects, subprojects, and related technologies work together. Each chapter introduces a different topic—such as core technologies or data transfer—and explains why certain components may or may not be useful for particular needs. When it comes to data, Hadoop is a whole new ballgame, but with this handy reference, you'll have a good grasp of the playing field. Topics include: Core technologies—Hadoop Distributed File System (HDFS), MapReduce, YARN, and Spark Database and data management—Cassandra, HBase, MongoDB, and Hive Serialization—Avro, JSON, and Parquet Management and monitoring—Puppet, Chef, Zookeeper, and Oozie Analytic helpers—Pig, Mahout, and MLLib Data transfer—Scoop, Flume, distcp, and Storm Security, access control, auditing—Sentry, Kerberos, and Knox Cloud computing and virtualization—Serengeti, Docker, and Whirr

This book is a step-by-step tutorial filled with practical examples which will show you how to build and manage a Hadoop cluster along with its intricacies. This book is ideal for database administrators, data engineers, and system administrators, and it will act as an invaluable reference if you are planning to use the Hadoop platform in your organization. It is expected that you have basic Linux skills since all the examples in this book use this operating system. It is also

Read Book Single Node Setup Apache Hadoop

useful if you have access to test hardware or virtual machines to be able to follow the examples in the book.

Ready to use statistical and machine-learning techniques across large data sets? This practical guide shows you why the Hadoop ecosystem is perfect for the job. Instead of deployment, operations, or software development usually associated with distributed computing, you'll focus on particular analyses you can build, the data warehousing techniques that Hadoop provides, and higher order data workflows this framework can produce. Data scientists and analysts will learn how to perform a wide range of techniques, from writing MapReduce and Spark applications with Python to using advanced modeling and data management with Spark MLlib, Hive, and HBase. You'll also learn about the analytical processes and data systems available to build and empower data products that can handle—and actually require—huge amounts of data. Understand core concepts behind Hadoop and cluster computing Use design patterns and parallel analytical algorithms to create distributed data analysis jobs Learn about data management, mining, and warehousing in a distributed context using Apache Hive and HBase Use Sqoop and Apache Flume to ingest data from relational databases Program complex Hadoop and Spark applications with Apache Pig and Spark DataFrames Perform machine learning techniques such as classification, clustering, and collaborative filtering with Spark's MLlib This book is aimed at developers, designers, and architects who would like to build big data enterprise search solutions for their customers or organizations. No prior knowledge of Apache Hadoop and Apache Solr/Lucene technologies is required.

This book constitutes the refereed proceedings of the 5th International Conference on Dependability in Sensor, Cloud, and Big Data Systems and Applications, DependSys, held in

Read Book Single Node Setup Apache Hadoop

Guangzhou, China, in November 2019. The volume presents 39 full papers, which were carefully reviewed and selected from 112 submissions. The papers are organized in topical sections on ?dependability and security fundamentals and technologies; dependable and secure systems; dependable and secure applications; dependability and security measures and assessments; explainable artificial intelligence for cyberspace.

This book constitutes the refereed proceedings of the 9th International Conference on Distributed Computing and Internet Technology, ICDCIT 2013, held in Bhubaneswar, India, in February 2013. The 40 full papers presented together with 5 invited talks in this volume were carefully reviewed and selected from 164 submissions. The papers cover various research aspects in distributed computing, internet technology, computer networks, and machine learning.

Hadoop Operations A Guide for Developers and Administrators "O'Reilly Media, Inc."

You can choose several data access frameworks when building Java enterprise applications that work with relational databases. But what about big data? This hands-on introduction shows you how Spring Data makes it relatively easy to build applications across a wide range of new data access technologies such as NoSQL and Hadoop. Through several sample projects, you'll learn how Spring Data provides a consistent programming model that retains NoSQL-specific features and capabilities, and helps you develop Hadoop applications across a wide range of use-cases such as data analysis, event stream processing, and workflow. You'll also discover the features Spring Data adds to Spring's existing JPA and JDBC support for writing RDBMS-based data access layers. Learn about Spring's template helper classes to simplify the use of database-specific

Read Book Single Node Setup Apache Hadoop

functionality Explore Spring Data's repository abstraction and advanced query functionality Use Spring Data with Redis (key/value store), HBase (column-family), MongoDB (document database), and Neo4j (graph database) Discover the GemFire distributed data grid solution Export Spring Data JPA-managed entities to the Web as RESTful web services Simplify the development of HBase applications, using a lightweight object-mapping framework Build example big-data pipelines with Spring Batch and Spring Integration

Summary Hadoop in Practice, Second Edition provides over 100 tested, instantly useful techniques that will help you conquer big data, using Hadoop. This revised new edition covers changes and new features in the Hadoop core architecture, including MapReduce 2. Brand new chapters cover YARN and integrating Kafka, Impala, and Spark SQL with Hadoop. You'll also get new and updated techniques for Flume, Sqoop, and Mahout, all of which have seen major new versions recently. In short, this is the most practical, up-to-date coverage of Hadoop available anywhere. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Book It's always a good time to upgrade your Hadoop skills! Hadoop in Practice, Second Edition provides a collection of 104 tested, instantly useful techniques for analyzing real-time streams, moving data securely, machine learning, managing large-scale clusters, and taming big data using Hadoop. This completely revised edition covers changes and new features in Hadoop core, including MapReduce 2 and YARN. You'll pick up hands-on best practices for integrating Spark, Kafka, and Impala with Hadoop, and get new and updated techniques for the latest versions of Flume, Sqoop, and Mahout. In short, this is the most practical, up-to-date coverage of Hadoop available. Readers need to know a programming language like Java and have basic familiarity

Read Book Single Node Setup Apache Hadoop

with Hadoop. What's Inside Thoroughly updated for Hadoop 2
How to write YARN applications Integrate real-time
technologies like Storm, Impala, and Spark Predictive
analytics using Mahout and RR Readers need to know a
programming language like Java and have basic familiarity
with Hadoop. About the Author Alex Holmes works on tough
big-data problems. He is a software engineer, author,
speaker, and blogger specializing in large-scale Hadoop
projects. Table of Contents PART 1 BACKGROUND AND
FUNDAMENTALS Hadoop in a heartbeat Introduction to
YARN PART 2 DATA LOGISTICS Data serialization—working
with text and beyond Organizing and optimizing data in HDFS
Moving data into and out of Hadoop PART 3 BIG DATA
PATTERNS Applying MapReduce patterns to big data
Utilizing data structures and algorithms at scale Tuning,
debugging, and testing PART 4 BEYOND MAPREDUCE SQL
on Hadoop Writing a YARN application

This volume includes 74 papers presented at ICTIS 2017:
Second International Conference on Information and
Communication Technology for Intelligent Systems. The
conference was held on 25th and 26th March 2017, in
Ahmedabad, India and organized jointly by the Associated
Chambers of Commerce and Industry of India (ASSOCHAM)
Gujarat Chapter, the G R Foundation, the Association of
Computer Machinery, Ahmedabad Chapter and supported by
the Computer Society of India Division IV – Communication
and Division V – Education and Research. The papers
featured mainly focus on information and communications
technology (ICT) for computation, algorithms and data
analytics. The fundamentals of various data analytics and
algorithms discussed are useful to researchers in the field.
This book presents high-quality, original
contributions (both theoretical and experimental) on

software engineering, cloud computing, computer networks & internet technologies, artificial intelligence, information security, and database and distributed computing. It gathers papers presented at ICRIC 2019, the 2nd International Conference on Recent Innovations in Computing, which was held in Jammu, India, in March 2019. This conference series represents a targeted response to the growing need for research that reports on and assesses the practical implications of IoT and network technologies, AI and machine learning, cloud-based e-Learning and big data, security and privacy, image processing and computer vision, and next-generation computing technologies.

This IBM® Redbooks® publication provides topics to help the technical community take advantage of the resilience, scalability, and performance of the IBM Power Systems™ 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.

From Visual Surveillance to Internet of Things: Technology and Applications is an invaluable resource for students, academicians and researchers to explore the utilization of Internet of Things with visual surveillance and its underlying technologies in different application areas. Using a series of present and future applications – business insights, indoor-outdoor securities, smart grids, human detection and tracking, intelligent traffic monitoring, e-health department and many more – this book will support readers to obtain a deeper knowledge in implementing IoT with visual surveillance. The book offers comprehensive coverage of the most essential topics, including: The

rise of machines and communications to IoT (3G, 5G) Tools and technologies of IoT with visual surveillance IoT with visual surveillance for real-time applications IoT architectures Challenging issues and novel solutions for realistic applications Mining and tracking of motion-based object data Image processing and analysis into the unified framework to understand both IOT and computer vision applications This book will be an ideal resource for IT professionals, researchers, under- or post-graduate students, practitioners, and technology developers who are interested in gaining a deeper knowledge in implementing IoT with visual surveillance, critical applications domains, technologies, and solutions to handle relevant challenges. Dr. Lavanya Sharma is an Assistant Professor in the Amity Institute of Information Technology at Amity University UP, Noida, India. She is a recipient of several prestigious awards during her academic career. She is an active nationally-recognized researcher who has published numerous papers in her field. She has contributed as an Organizing Committee member and session chair at Springer and IEEE conferences. Prof. Pradeep K. Garg worked as a Vice Chancellor, Uttarakhand Technical University, Dehradun. Presently he is working in the department of Civil Engineering, IIT Roorkee as a professor. Prof. Garg has published more than 300 technical papers in national and

international conferences and journals. He has completed 26 research projects funded by various government agencies, guided 27 PhD candidates, and provided technical services to 84 consultancy projects on various aspects of Civil Engineering.

Combine the power of Apache Spark and Python to build effective big data applications

Key Features

Perform effective data processing, machine learning, and analytics using PySpark

Overcome challenges in developing and deploying Spark solutions using Python

Explore recipes for efficiently combining Python and Apache Spark to process data

Book Description

Apache Spark is an open source framework for efficient cluster computing with a strong interface for data parallelism and fault tolerance. The PySpark Cookbook presents effective and time-saving recipes for leveraging the power of Python and putting it to use in the Spark ecosystem. You'll start by learning the Apache Spark architecture and how to set up a Python environment for Spark. You'll then get familiar with the modules available in PySpark and start using them effortlessly. In addition to this, you'll discover how to abstract data with RDDs and DataFrames, and understand the streaming capabilities of PySpark. You'll then move on to using ML and MLlib in order to solve any problems related to the machine learning capabilities of PySpark and use GraphFrames to solve graph-processing problems.

Finally, you will explore how to deploy your applications to the cloud using the `spark-submit` command. By the end of this book, you will be able to use the Python API for Apache Spark to solve any problems associated with building data-intensive applications. What you will learn

- Configure a local instance of PySpark in a virtual environment
- Install and configure Jupyter in local and multi-node environments
- Create DataFrames from JSON and a dictionary using `pyspark.sql`
- Explore regression and clustering models available in the ML module
- Use DataFrames to transform data used for modeling
- Connect to PubNub and perform aggregations on streams

Who this book is for

The *PySpark Cookbook* is for you if you are a Python developer looking for hands-on recipes for using the Apache Spark 2.x ecosystem in the best possible way. A thorough understanding of Python (and some familiarity with Spark) will help you get the best out of the book. The book proposes new technologies and discusses future solutions for design infrastructure for ICT. The book contains high quality submissions presented at Second International Conference on Information and Communication Technology for Sustainable Development (ICT4SD - 2016) held at Goa, India during 1 - 2 July, 2016. The conference stimulates the cutting-edge research discussions among many academic pioneering researchers, scientists, industrial engineers, and students from all around

the world. The topics covered in this book also focus on innovative issues at international level by bringing together the experts from different countries.

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.

If you are a Hadoop programmer who wants to learn about Flume to be able to move datasets into Hadoop in a timely and replicable manner, then this book is ideal for you. No prior knowledge about Apache Flume is necessary, but a basic knowledge of Hadoop and the Hadoop File System (HDFS) is assumed.

"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 Hadoop™ 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

A comprehensive end-to-end guide that gives hands-on practice in big data and Artificial Intelligence Key Features Learn to build and run a big data application with sample code Explore examples to implement activities that a big data architect performs Use Machine Learning and AI for structured and unstructured data Book Description The big data architects are the “masters” of data, and hold high value in today’s market. Handling big data, be it of good or bad quality, is not an easy task. The prime job for any big data architect is to build an end-to-end big data solution that integrates data from different sources and analyzes it to find useful, hidden insights. Big Data Architect’s Handbook takes you through developing a complete, end-to-end big data pipeline, which will lay the foundation for you and provide the necessary knowledge required to be an architect in big data. Right from understanding the design considerations to

implementing a solid, efficient, and scalable data pipeline, this book walks you through all the essential aspects of big data. It also gives you an overview of how you can leverage the power of various big data tools such as Apache Hadoop and ElasticSearch in order to bring them together and build an efficient big data solution. By the end of this book, you will be able to build your own design system which integrates, maintains, visualizes, and monitors your data. In addition, you will have a smooth design flow in each process, putting insights in action. What you will learn

- Learn Hadoop Ecosystem and Apache projects
- Understand, compare NoSQL database and essential software architecture
- Cloud infrastructure design considerations for big data
- Explore application scenario of big data tools for daily activities
- Learn to analyze and visualize results to uncover valuable insights
- Build and run a big data application with sample code from end to end
- Apply Machine Learning and AI to perform big data intelligence
- Practice the daily activities performed by big data architects

Who this book is for

Big Data Architect's Handbook is for you if you are an aspiring data professional, developer, or IT enthusiast who aims to be an all-round architect in big data. This book is your one-stop solution to enhance your knowledge and carry out easy to complex activities required to become a big data architect.

Read Book Single Node Setup Apache Hadoop

Work with Apache Spark using Scala to deploy and set up single-node, multi-node, and high-availability clusters. This book discusses various components of Spark such as Spark Core, DataFrames, Datasets and SQL, Spark Streaming, Spark MLib, and R on Spark with the help of practical code snippets for each topic. Practical Apache Spark also covers the integration of Apache Spark with Kafka with examples. You'll follow a learn-to-do-by-yourself approach to learning – learn the concepts, practice the code snippets in Scala, and complete the assignments given to get an overall exposure. On completion, you'll have knowledge of the functional programming aspects of Scala, and hands-on expertise in various Spark components. You'll also become familiar with machine learning algorithms with real-time usage. What You Will Learn Discover the functional programming features of Scala Understand the complete architecture of Spark and its components Integrate Apache Spark with Hive and Kafka Use Spark SQL, DataFrames, and Datasets to process data using traditional SQL queries Work with different machine learning concepts and libraries using Spark's MLlib packages Who This Book Is For Developers and professionals who deal with batch and stream data processing.

This is the eBook of the printed book and may not include any media, website access codes, or print supplements that may come packaged with the bound book. The Comprehensive, Up-to-Date Apache Hadoop Administration Handbook and Reference “Sam Alapati has worked with production Hadoop clusters for six years. His unique depth of experience has enabled him to write the go-to resource for all administrators looking to spec, size, expand, and secure production Hadoop clusters of any size.” —Paul Dix, Series Editor In Expert Hadoop® Administration, leading Hadoop administrator Sam R. Alapati brings together authoritative knowledge for creating, configuring, securing, managing, and optimizing

production Hadoop clusters in any environment. Drawing on his experience with large-scale Hadoop administration, Alapati integrates action-oriented advice with carefully researched explanations of both problems and solutions. He covers an unmatched range of topics and offers an unparalleled collection of realistic examples. Alapati demystifies complex Hadoop environments, helping you understand exactly what happens behind the scenes when you administer your cluster. You'll gain unprecedented insight as you walk through building clusters from scratch and configuring high availability, performance, security, encryption, and other key attributes. The high-value administration skills you learn here will be indispensable no matter what Hadoop distribution you use or what Hadoop applications you run. Understand Hadoop's architecture from an administrator's standpoint Create simple and fully distributed clusters Run MapReduce and Spark applications in a Hadoop cluster Manage and protect Hadoop data and high availability Work with HDFS commands, file permissions, and storage management Move data, and use YARN to allocate resources and schedule jobs Manage job workflows with Oozie and Hue Secure, monitor, log, and optimize Hadoop Benchmark and troubleshoot Hadoop

Due to the increasing availability of affordable internet services, the number of users, and the need for a wider range of multimedia-based applications, internet usage is on the rise. With so many users and such a large amount of data, the requirements of analyzing large data sets leads to the need for further advancements to information processing. Big Data Processing With Hadoop is an essential reference source that discusses possible solutions for millions of users working with a variety of data applications, who expect fast turnaround responses, but encounter issues with processing data at the rate it comes in. Featuring research on topics

Read Book Single Node Setup Apache Hadoop

such as market basket analytics, scheduler load simulator, and writing YARN applications, this book is ideally designed for IoT professionals, students, and engineers seeking coverage on many of the real-world challenges regarding big data.

Ready to unlock the power of your data? With this comprehensive guide, you'll learn how to build and maintain reliable, scalable, distributed systems with Apache Hadoop. This book is ideal for programmers looking to analyze datasets of any size, and for administrators who want to set up and run Hadoop clusters. You'll find illuminating case studies that demonstrate how Hadoop is used to solve specific problems. This third edition covers recent changes to Hadoop, including material on the new MapReduce API, as well as MapReduce 2 and its more flexible execution model (YARN). Store large datasets with the Hadoop Distributed File System (HDFS) Run distributed computations with MapReduce Use Hadoop's data and I/O building blocks for compression, data integrity, serialization (including Avro), and persistence Discover common pitfalls and advanced features for writing real-world MapReduce programs Design, build, and administer a dedicated Hadoop cluster—or run Hadoop in the cloud Load data from relational databases into HDFS, using Sqoop Perform large-scale data processing with the Pig query language Analyze datasets with Hive, Hadoop's data warehousing system Take advantage of HBase for structured and semi-structured data, and ZooKeeper for building distributed systems

R and Data Mining introduces researchers, post-graduate students, and analysts to data mining using R, a free software environment for statistical computing and graphics. The book provides practical methods for using R in applications from academia to industry to extract knowledge from vast amounts of data. Readers will find this book a valuable guide to the

Read Book Single Node Setup Apache Hadoop

use of R in tasks such as classification and prediction, clustering, outlier detection, association rules, sequence analysis, text mining, social network analysis, sentiment analysis, and more. Data mining techniques are growing in popularity in a broad range of areas, from banking to insurance, retail, telecom, medicine, research, and government. This book focuses on the modeling phase of the data mining process, also addressing data exploration and model evaluation. With three in-depth case studies, a quick reference guide, bibliography, and links to a wealth of online resources, R and Data Mining is a valuable, practical guide to a powerful method of analysis. Presents an introduction into using R for data mining applications, covering most popular data mining techniques Provides code examples and data so that readers can easily learn the techniques Features case studies in real-world applications to help readers apply the techniques in their work

Get a solid grounding in Apache Oozie, the workflow scheduler system for managing Hadoop jobs. With this hands-on guide, two experienced Hadoop practitioners walk you through the intricacies of this powerful and flexible platform, with numerous examples and real-world use cases. Once you set up your Oozie server, you'll dive into techniques for writing and coordinating workflows, and learn how to write complex data pipelines. Advanced topics show you how to handle shared libraries in Oozie, as well as how to implement and manage Oozie's security capabilities. Install and configure an Oozie server, and get an overview of basic concepts Journey through the world of writing and configuring workflows Learn how the Oozie coordinator schedules and executes workflows based on triggers Understand how Oozie manages data dependencies Use Oozie bundles to package several coordinator apps into a data pipeline Learn about security features and shared library management Implement

Read Book Single Node Setup Apache Hadoop

custom extensions and write your own EL functions and actions Debug workflows and manage Oozie's operational details

A hands-on guide to leveraging NoSQL databases NoSQL databases are an efficient and powerful tool for storing and manipulating vast quantities of data. Most NoSQL databases scale well as data grows. In addition, they are often malleable and flexible enough to accommodate semi-structured and sparse data sets. This comprehensive hands-on guide presents fundamental concepts and practical solutions for getting you ready to use NoSQL databases. Expert author Shashank Tiwari begins with a helpful introduction on the subject of NoSQL, explains its characteristics and typical uses, and looks at where it fits in the application stack.

Unique insights help you choose which NoSQL solutions are best for solving your specific data storage needs.

Professional NoSQL: Demystifies the concepts that relate to NoSQL databases, including column-family oriented stores, key/value databases, and document databases. Delves into installing and configuring a number of NoSQL products and the Hadoop family of products. Explains ways of storing, accessing, and querying data in NoSQL databases through examples that use MongoDB, HBase, Cassandra, Redis, CouchDB, Google App Engine Datastore and more. Looks at architecture and internals. Provides guidelines for optimal usage, performance tuning, and scalable configurations. Presents a number of tools and utilities relating to NoSQL, distributed platforms, and scalable processing, including Hive, Pig, RRDtool, Nagios, and more.

Beginning Apache Cassandra Development introduces you to one of the most robust and best-performing NoSQL database platforms on the planet. Apache Cassandra is a document database following the JSON document model. It is specifically designed to manage large amounts of data across

Read Book Single Node Setup Apache Hadoop

many commodity servers without there being any single point of failure. This design approach makes Apache Cassandra a robust and easy-to-implement platform when high availability is needed. Apache Cassandra can be used by developers in Java, PHP, Python, and JavaScript—the primary and most commonly used languages. In *Beginning Apache Cassandra Development*, author and Cassandra expert Vivek Mishra takes you through using Apache Cassandra from each of these primary languages. Mishra also covers the Cassandra Query Language (CQL), the Apache Cassandra analog to SQL. You'll learn to develop applications sourcing data from Cassandra, query that data, and deliver it at speed to your application's users. Cassandra is one of the leading NoSQL databases, meaning you get unparalleled throughput and performance without the sort of processing overhead that comes with traditional proprietary databases. *Beginning Apache Cassandra Development* will therefore help you create applications that generate search results quickly, stand up to high levels of demand, scale as your user base grows, ensure operational simplicity, and—not least—provide delightful user experiences.

Over 100 practical recipes to help you become an expert Hadoop administrator
About This Book Become an expert Hadoop administrator and perform tasks to optimize your Hadoop Cluster
Import and export data into Hive and use Oozie to manage workflow. Practical recipes will help you plan and secure your Hadoop cluster, and make it highly available
Who This Book Is For If you are a system administrator with a basic understanding of Hadoop and you want to get into Hadoop administration, this book is for you. It's also ideal if you are a Hadoop administrator who wants a quick reference guide to all the Hadoop administration-related tasks and solutions to commonly occurring problems
What You Will Learn Set up the Hadoop architecture to run a

Read Book Single Node Setup Apache Hadoop

Hadoop cluster smoothly Maintain a Hadoop cluster on HDFS, YARN, and MapReduce Understand high availability with Zookeeper and Journal Node Configure Flume for data ingestion and Oozie to run various workflows Tune the Hadoop cluster for optimal performance Schedule jobs on a Hadoop cluster using the Fair and Capacity scheduler Secure your cluster and troubleshoot it for various common pain points In Detail Hadoop enables the distributed storage and processing of large datasets across clusters of computers. Learning how to administer Hadoop is crucial to exploit its unique features. With this book, you will be able to overcome common problems encountered in Hadoop administration. The book begins with laying the foundation by showing you the steps needed to set up a Hadoop cluster and its various nodes. You will get a better understanding of how to maintain Hadoop cluster, especially on the HDFS layer and using YARN and MapReduce. Further on, you will explore durability and high availability of a Hadoop cluster. You'll get a better understanding of the schedulers in Hadoop and how to configure and use them for your tasks. You will also get hands-on experience with the backup and recovery options and the performance tuning aspects of Hadoop. Finally, you will get a better understanding of troubleshooting, diagnostics, and best practices in Hadoop administration. By the end of this book, you will have a proper understanding of working with Hadoop clusters and will also be able to secure, encrypt it, and configure auditing for your Hadoop clusters.

Style and approach This book contains short recipes that will help you run a Hadoop cluster efficiently. The recipes are solutions to real-life problems that administrators encounter while working with a Hadoop cluster

The two volumes of this book collect high-quality peer-reviewed research papers presented in the International

Conference on ICT for Sustainable Development (ICT4SD 2015) held at Ahmedabad, India during 3 – 4 July 2015. The book discusses all areas of Information and Communication Technologies and its applications in field for engineering and management. The main focus of the volumes are on applications of ICT for Infrastructure, e-Governance, and contemporary technologies advancements on Data Mining, Security, Computer Graphics, etc. The objective of this International Conference is to provide an opportunity for the researchers, academicians, industry persons and students to interact and exchange ideas, experience and expertise in the current trend and strategies for Information and Communication Technologies. 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

Read Book Single Node Setup Apache Hadoop

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

How can you get your data from frontend servers to Hadoop in near real time? With this complete reference guide, you'll learn Flume's rich set of features for collecting, aggregating, and writing large amounts of streaming data to the Hadoop Distributed File System (HDFS), Apache HBase, SolrCloud, Elastic Search, and other systems. Using Flume shows operations engineers how to configure, deploy, and monitor a Flume cluster, and teaches developers how to write Flume plugins and custom components for their specific use-cases. You'll learn about Flume's design and implementation, as well as various features that make it highly scalable, flexible, and reliable. Code examples and exercises are available on GitHub. Learn how Flume provides a steady rate of flow by acting as a buffer between data producers and consumers Dive into key Flume components, including sources that accept data and sinks that write and deliver it Write custom plugins to customize the way Flume receives, modifies, formats, and writes data Explore APIs for sending data to Flume agents from your own applications Plan and deploy Flume in a scalable and flexible way—and monitor your cluster once it's running Let Hadoop For Dummies help harness the power of your data and rein in the information overload Big data has become big business, and companies and organizations of all sizes are struggling to find ways to retrieve valuable information from their massive data sets with becoming overwhelmed. Enter Hadoop and this

Read Book Single Node Setup Apache Hadoop

easy-to-understand For Dummies guide. Hadoop For Dummies helps readers understand the value of big data, make a business case for using Hadoop, navigate the Hadoop ecosystem, and build and manage Hadoop applications and clusters. Explains the origins of Hadoop, its economic benefits, and its functionality and practical applications Helps you find your way around the Hadoop ecosystem, program MapReduce, utilize design patterns, and get your Hadoop cluster up and running quickly and easily Details how to use Hadoop applications for data mining, web analytics and personalization, large-scale text processing, data science, and problem-solving Shows you how to improve the value of your Hadoop cluster, maximize your investment in Hadoop, and avoid common pitfalls when building your Hadoop cluster From programmers challenged with building and maintaining affordable, scalable data systems to administrators who must deal with huge volumes of information effectively and efficiently, this how-to has something to help you with Hadoop.

Get Started Fast with Apache Hadoop® 2, YARN, and Today's Hadoop Ecosystem With Hadoop 2.x and YARN, Hadoop moves beyond MapReduce to become practical for virtually any type of data processing. Hadoop 2.x and the Data Lake concept represent a radical shift away from conventional approaches to data usage and storage. Hadoop 2.x installations offer unmatched scalability and breakthrough extensibility that supports new and existing Big Data analytics processing methods and models. Hadoop® 2 Quick-Start Guide is

Read Book Single Node Setup Apache Hadoop

the first easy, accessible guide to Apache Hadoop 2.x, YARN, and the modern Hadoop ecosystem. Building on his unsurpassed experience teaching Hadoop and Big Data, author Douglas Eadline covers all the basics you need to know to install and use Hadoop 2 on personal computers or servers, and to navigate the powerful technologies that complement it. Eadline concisely introduces and explains every key Hadoop 2 concept, tool, and service, illustrating each with a simple “beginning-to-end” example and identifying trustworthy, up-to-date resources for learning more. This guide is ideal if you want to learn about Hadoop 2 without getting mired in technical details. Douglas Eadline will bring you up to speed quickly, whether you’re a user, admin, devops specialist, programmer, architect, analyst, or data scientist. Coverage Includes Understanding what Hadoop 2 and YARN do, and how they improve on Hadoop 1 with MapReduce Understanding Hadoop-based Data Lakes versus RDBMS Data Warehouses Installing Hadoop 2 and core services on Linux machines, virtualized sandboxes, or clusters Exploring the Hadoop Distributed File System (HDFS) Understanding the essentials of MapReduce and YARN application programming Simplifying programming and data movement with Apache Pig, Hive, Sqoop, Flume, Oozie, and HBase Observing application progress, controlling jobs, and managing workflows Managing Hadoop efficiently with Apache Ambari—including recipes for HDFS to NFSv3 gateway, HDFS snapshots, and YARN configuration Learning basic Hadoop 2 troubleshooting, and installing Apache Hue and Apache

Read Book Single Node Setup Apache Hadoop

Spark

[Copyright: ed486ff1197c1254be3fdc8583c3ead9](#)