Apache Spark Tutorial Tutorialspoint

Information Systems Design and Intelligent ApplicationsProceedings of Fourth International Conference INDIA 2017Springer

Many applications process high volumes of streaming data, among them Internet traffic analysis, financial tickers, and transaction log mining. In general, a data stream is an unbounded data set that is produced incrementally over time, rather than being available in full before its processing begins. In this lecture, we give an overview of recent research in stream processing, ranging from answering simple queries on high-speed streams to loading real-time data feeds into a streaming warehouse for off-line analysis. We will discuss two types of systems for end-to-end stream processing: Data Stream Management Systems (DSMSs) and Streaming Data Warehouses (SDWs). A traditional database management system typically processes a stream of ad-hoc queries over relatively static data. In contrast, a DSMS evaluates static (long-running) queries on streaming data, making a single pass over the data and using limited working memory. In the first part of this lecture, we will discuss research problems in DSMSs, such as continuous query languages, non-blocking query operators that continually react to new data, and continuous query optimization. The second part covers SDWs, which combine the real-time response of a DSMS by loading new data as soon as they arrive with a data warehouse's ability to manage Terabytes of historical data on secondary storage. Table of Contents: Introduction / Data Stream Management Systems / Streaming Data Warehouses / Conclusions

This book has a two-fold mission: to explain and facilitate digital transition in business organizations using information and communications technology and to address the associated growing threat of cyber crime and the challenge of creating and maintaining effective cyber protection. The book begins with a section on Digital Business Transformation, which includes chapters on tools for integrated marketing communications, human resource workplace digitalization, the integration of the Internet of Things in the workplace, Big Data, and more. The technologies discussed aim to help businesses and entrepreneurs transform themselves to align with today’s modern digital climate. The Evolution of Business in the Cyber Age: Digital Transformation, Threats, and Security provides a wealth of information for those involved in the development and management of conducting business online as well as for those responsible for cyber protection and security. Faculty and students, researchers, and industry professionals will find much of value in this volume.

This volume comprises the select proceedings of the annual convention of the Computer Society of India. Divided into 10 topical volumes, the proceedings present papers on state-of-the-art research, surveys, and succinct reviews. The volumes cover diverse topics ranging from communications networks to big data analytics, and from system architecture to cyber security. This volume focuses on Big Data Analytics. The contents of this book will be useful to researchers and students alike.

The big data era is upon us: data are being generated, analyzed, and used at an unprecedented scale, and data-driven decision making is sweeping through all aspects of society. Since the value of data explodes when it can be linked and fused with other data, addressing the big data integration (BDI) challenge is critical to realizing the promise of big data. BDI differs from traditional data integration along the dimensions of volume, velocity, variety, and veracity. First, not only can data sources contain a huge volume of data, but also the number of data sources is now in the millions. Second, because of the rate at which newly collected data are made available, many of the data sources are very dynamic, and the number of data sources is also rapidly exploding. Third, data sources are extremely heterogeneous in their structure and content, exhibiting considerable variety even for substantially similar entities. Fourth, the data sources are of widely differing qualities, with significant differences in the coverage, accuracy and timeliness of data provided. This book explores the progress that has been made by the data integration community on the topics of schema alignment, record linkage and data fusion in addressing these novel challenges faced by big data integration. Each of these topics is covered in a systematic way: first starting with a quick tour of the topic in the context of traditional data integration, followed by a detailed, example-driven exposition of recent innovative techniques that have been proposed to address the BDI challenges of volume, velocity, variety, and veracity. Finally, it presents merging topics and opportunities that are specific to BDI, identifying promising directions for the data integration community.

Search is an important component of problem solving in artificial intelligence (AI) and, more generally, in computer science, engineering and operations research. Combinatorial optimization, decision analysis, game playing, learning, planning, pattern recognition, robotics and theorem proving are some of the areas in which search algbrithms playa key role. Less than a decade ago the conventional wisdom in artificial intelligence was that the best search algorithms had already been invented and the likelihood of finding new results in this area was very small. Since then many new insights and results have been obtained. For example, new algorithms for state space, AND/OR graph, and game tree search were discovered. Articles on new theoretical developments and experimental results on backtracking, heuristic search and constraint propaga tion were published. The relationships among various search and combinatorial algorithms in AI, Operations Research, and other fields were clarified. This volume brings together some of this recent work in a manner designed to be accessible to students and professionals interested in these new insights and developments.

Get complete instructions for manipulating, processing, cleaning, and crunching datasets in Python. Updated for Python 3.6, the second edition of this hands-on guide is packed with practical case studies that show you how to solve a broad set of data analysis problems effectively. You’ll learn the latest versions of pandas, NumPy, IPython, and Jupyter in the process. Written by Wes McKinney, the creator of the Python pandas project, this book is a practical, modern introduction to data science tools in Python. It’s ideal for analysts new to Python and for Python programmers new to data science and scientific computing. Data files and related material are available on GitHub. Use the IPython shell and Jupyter notebook for exploratory computing Learn basic and advanced features in NumPy (Numerical Python) Get started with data analysis tools in the pandas library Use flexible tools to load, clean, transform, merge, and reshape data Create informative visualizations with matplotlib Apply the pandas groupby facility to slice, dice, and summarize datasets Analyze and manipulate regular and irregular time series data Learn how to solve real-world data analysis problems with thorough, detailed examples

As the first volume of World Scientific Encyclopedia with Semantic Computing and Robotic Intelligence, this volume is designed to lay the foundation for the understanding of the Semantic Computing (SC), as a core concept to study Robotic Intelligence in the subsequent volumes. This volume aims to provide a reference to the development of Semantic Computing, in the terms

of "meaning", "context", and "intention". It brings together a series of technical notes, in average, no longer than 10 pages in length, each focuses on one topic in Semantic Computing; being review article or research paper, to explain the fundamental concepts, models or algorithms, and possible applications of the technology concerned. This volume will address three core areas in Semantic Computing: Understanding the (possibly naturally-expressed) intentions (semantics) of users and expressing them in a machine-processable format: Semantics description languages, ontology integration, interoperabilityUnderstanding the meanings (semantics) of computational content (of various sorts, including, but is not limited to, text, video, audio, process, network, software and hardware) and expressing them in a machine-processable format in Multimedia, IoT, SDN, wearable computing, interfacable with mobile computing, search engines, question answering, web services, to support applications in biomedicine, healthcare, manufacturing, engineering, education, finance, entertainment, business, science and humanityMapping the semantics of the user in context for content retrieval, management, creation in the form of structured data, image and video, audio and speech, big data, natural language, deep learning.

Build data-intensive applications locally and deploy at scale using the combined powers of Python and Spark 2.0 About This Book Learn why and how you can efficiently use Python to process data and build machine learning models in Apache Spark 2.0 Develop and deploy efficient, scalable real-time Spark solutions Take your understanding of using Spark with Python to the next level with this jump start guide Who This Book Is For If you are a Python developer who wants to learn about the Apache Spark 2.0 ecosystem, this book is for you. A firm understanding of Python is expected to get the best out of the book. Familiarity with Spark would be useful, but is not mandatory. What You Will Learn Learn about Apache Spark and the Spark 2.0 architecture Build and interact with Spark DataFrames using Spark SQL Learn how to solve graph and deep learning problems using GraphFrames and TensorFrames respectively Read, transform, and understand data and use it to train machine learning models Build machine learning models with MLlib and ML Learn how to submit your applications programmatically using spark-submit Deploy locally built applications to a cluster In Detail Apache Spark is an open source framework for efficient cluster computing with a strong interface for data parallelism and fault tolerance. This book will show you how to leverage the power of Python and put it to use in the Spark ecosystem. You will start by getting a firm understanding of the Spark 2.0 architecture and how to set up a Python environment for Spark. You will get familiar with the modules available in PySpark. You will learn how to abstract data with RDDs and DataFrames and understand the streaming capabilities of PySpark. Also, you will get a thorough overview of machine learning capabilities of PySpark using ML and MLlib, graph processing using GraphFrames, and polyglot persistence using Blaze. Finally, you will learn how to deploy your applications to the cloud using the spark-submit command. By the end of this book, you will have established a firm understanding of the Spark Python API and how it can be used to build data-intensive applications. Style and approach This book takes a very comprehensive, step-by-step approach so you understand how the Spark ecosystem can be used with Python to develop efficient, scalable solutions. Every chapter is standalone and written in a very easy-to-understand manner, with a focus on both the hows and the whys of each concept.

This book is intended for developers and operators who want to build and run scalable and fault-tolerant applications leveraging Apache Mesos. A basic knowledge of programming with some fundamentals of Linux is a prerequisite.

If you’re like most R users, you have deep knowledge and love for statistics. But as your organization continues to collect huge amounts of data, adding tools such as Apache Spark makes a lot of sense. With this practical book, data scientists and professionals working with large-scale data applications will learn how to use Spark from R to tackle big data and big compute problems. Authors Javier Luraschi, Kevin Kuo, and Edgar Ruiz show you how to use R with Spark to solve different data analysis problems. This book covers relevant data science topics, cluster computing, and issues that should interest even the most advanced users. Analyze, explore, transform, and visualize data in Apache Spark with R Create statistical models to extract information and predict outcomes; automate the process in production-ready workflows Perform analysis and modeling across many machines using distributed computing techniques Use large-scale data from multiple sources and different formats with ease from within Spark Learn about alternative modeling frameworks for graph processing, geospatial analysis, and genomics at scale Dive into advanced topics including custom transformations, real-time data processing, and creating custom Spark extensions

Docker containers offer simpler, faster, and more robust methods for developing, distributing, and running software than previously available. With this hands-on guide, you’ll learn why containers are so important, what you’ll gain by adopting Docker, and how to make it part of your development process. Ideal for developers, operations engineers, and system administrators—especially those keen to embrace a DevOps approach—Using Docker will take you from Docker and container basics to running dozens of containers on a multi-host system with networking and scheduling. The core of the book walks you through the steps needed to develop, test, and deploy a web application with Docker. Get started with Docker by building and deploying a simple web application Use Continuous Deployment techniques to push your application to production multiple times a day Learn various options and techniques for logging and monitoring multiple containers Examine networking and service discovery: how do containers find each other and how do you connect them? Orchestrate and cluster containers to address load-balancing, scaling, failover, and scheduling Secure your system by following the principles of defense-in-depth and least privilege

The fast and easy way to learn Python programming and statistics Python is a general-purpose programming language created in the late 1980s—and named after Monty Python—that's used by thousands of people to do things from testing microchips at Intel, to powering Instagram, to building video games with the PyGame library. Python For Data Science For Dummies is written for people who are new to data analysis, and discusses the basics of Python data analysis programming and statistics. The book also discusses Google Colab, which makes it possible to write Python code in the cloud. Get started with data science and Python Visualize information Wrangle data Learn from data The book provides the statistical background needed to get started in data science programming, including probability, random distributions, hypothesis testing, confidence intervals, and building regression models for prediction.

The book is a collection of high-quality peer-reviewed research papers presented at International Conference on Information System Design and Intelligent Applications (INDIA 2017) held at Duy Tan University, Da Nang, Vietnam during 15-17 June 2017. The book covers a wide range of topics of computer science and information technology discipline ranging from image processing, database application, data mining, grid and cloud computing, bioinformatics and many others. The various intelligent tools like swarm intelligence, artificial intelligence, evolutionary algorithms, bio-inspired algorithms have been well applied in different domains for solving various challenging problems.

Microsoft Azure Essentials from Microsoft Press is a series of free ebooks designed to help you advance your technical skills with Microsoft Azure. The first ebook in the series, Microsoft Azure Essentials: Fundamentals of Azure, introduces developers and IT professionals to the wide range of capabilities in Azure. The authors - both Microsoft MVPs in Azure - present both conceptual and how-to content for key areas, including: Azure Websites and Azure Cloud Services Azure Virtual Machines Azure Storage Azure Virtual Networks Databases Azure Active Directory Management tools Business scenarios Watch Microsoft Press’s blog and Twitter (@MicrosoftPress) to learn about other free ebooks in the “Microsoft Azure Essentials” series.

The sixth edition of this most trusted book on JAVA for beginners is here with some essential updates. Retaining its quintessential style of concept explanation with exhaustive programs, solved examples, and illustrations, this test takes the journey of understanding JAVA to slightly higher level. The book introduces readers to some of the Core JAVA topics like JDBC, Java Servlets, Java Beans, Lambada Expression and much more. Practical real-life projects will give a better understanding of JAVA usage and make students industry-ready.

JSON is becoming the backbone for meaningful data interchange over the internet. This format is now supported by an entire ecosystem of standards, tools, and technologies for building truly elegant, useful, and efficient applications. With this hands-on guide, author and architect Tom Marrs shows you how to build enterprise-class applications and services by leveraging JSON tooling and message/document design. JSON at Work provides application architects and developers with guidelines, best practices, and use cases, along with lots of real-world examples and code samples. You'll start with a comprehensive JSON overview, explore the JSON ecosystem, and then dive into JSON's use in the enterprise. Get acquainted with JSON basics and learn how to model JSON data Learn how to use JSON with Node.js, Ruby on Rails, and Java Structure JSON documents with JSON Schema to design and test APIs Search the contents of JSON documents with JSON Search tools Convert JSON documents to other data formats with JSON Transform tools Compare JSON-based hypermedia formats, including HAL and jsonapi Leverage MongoDB to store and access JSON documents Use Apache Kafka to exchange JSON-based messages between services

This book gathers papers addressing state-of-the-art research in all areas of information and communication technologies and their applications in intelligent computing, cloud storage, data mining and software analysis. It presents the outcomes of the Fourth International Conference on Information and Communication Technology for Intelligent Systems, which was held in Ahmedabad, India. Divided into two volumes, the book discusses the fundamentals of various data analysis techniques and algorithms, making it a valuable resource for researchers and practitioners alike.

Learn how to write scalable and concurrent programs in Scala, a language that grows with you. Key Features Get a grip on the functional features of the Scala programming language Understand and develop optimal applications using object-oriented and functional Scala constructs Learn reactive principles with Scala and work with the Akka framework Book Description Scala is a general-purpose programming language that supports both functional and object-oriented programming paradigms. Due to its concise design and versatility, Scala's applications have been extended to a wide variety of fields such as data science and cluster computing. You will learn to write highly scalable, concurrent, and testable programs to meet everyday software requirements. We will begin by understanding the language basics, syntax, core data types, literals, variables, and more. From here you will be introduced to data structures with Scala and you will learn to work with higher-order functions. Scala's powerful collections framework will help you get the best out of immutable data structures and utilize them effectively. You will then be introduced to concepts such as pattern matching, case classes, and functional programming features. From here, you will learn to work with Scala's object-oriented features. Going forward, you will learn about asynchronous and reactive programming with Scala, where you will be introduced to the Akka framework. Finally, you will learn the interoperability of Scala and Java. After reading this book, you'll be well versed with this language and its features, and you will be able to write scalable, concurrent, and reactive programs in Scala. What you will learn Get to know the reasons for choosing Scala: its use and the advantages it provides over other languages Bring together functional and object-oriented programming constructs to make a manageable application Master basic to advanced Scala constructs Test your applications using advanced testing methodologies such as TDD Select preferred language constructs from the wide variety of constructs provided by Scala Make the transition from the object-oriented paradigm to the functional programming paradigm Write clean, concise, and powerful code with a functional mindset Create concurrent, scalable, and reactive applications utilizing the advantages of Scala Who this book is for This book is for programmers who choose to get a grip over Scala to write concurrent, scalable, and reactive programs. No prior experience with any programming language is required to learn the concepts explained in this book. Knowledge of any programming language would help the reader understanding concepts faster though.

If you are a data analyst, developer, or simply someone who wants to use Hive to explore and analyze data in Hadoop, this is the book for you. Whether you are new to big data or an expert, with this book, you will be able to master both the basic and the advanced features of Hive. Since Hive is an SQL-like language, some previous experience with the SQL language and databases is useful to have a better understanding of this book.

From the Foreword: "This book lays out much of what we've learned at AT&T about SDN and NFV. Some of the smartest network experts in the industry have drawn a map to help you navigate this journey. Their goal isn't to predict the future but to help you design and build a network that will be ready for whatever that future holds. Because if there's one thing the last decade has taught us, it's that network demand will always exceed expectations. This book will help you get ready." —Randall Stephenson, Chairman, CEO, and President of AT&T "Software is changing the world, and networks too. In this in-depth book, AT&T's top networking experts discuss how they're moving software-defined networking from concept to practice, and why it's a business imperative to do this rapidly." —Urs Hölzle, SVP Cloud Infrastructure, Google "Telecom operators face a continuous challenge for more agility to serve their customers with a better customer experience and a lower cost. This book is a very inspiring and vivid testimony of the huge transformation this means, not only for the networks but for the entire companies, and how AT&T is leading it. It provides a lot of very deep insights about the technical challenges telecom engineers are facing today. Beyond AT&T, I'm sure this book will be extremely helpful to the whole industry." —Alain Maloberti, Group Chief Network Officer, Orange Labs Networks "This new book should be read by any organization faced with a future driven by a "shift to software." It is a holistic view of how AT&T has transformed its core infrastructure from hardware based to largely software based to lower costs and speed innovation. To do so, AT&T had to redefine their technology supply chain, retrain their workforce, and move toward open source user-driven innovation; all while managing one of the biggest networks in the world. It is an amazing feat that will put AT&T in a leading position for years to come." —Jim Zemlin, Executive Director, The Linux Foundation This book is based on the lessons learned from AT&T's software transformation journey starting in 2012 when rampant traffic growth necessitated a change in network architecture and design. Using new technologies such as NFV, SDN, Cloud, and Big Data, AT&T's engineers outlined and implemented a radical network transformation program that dramatically reduced capital and operating expenditures. This book describes the transformation in substantial detail. The subject matter is of great interest to telecom professionals worldwide, as well as academic researchers looking to apply the latest techniques in computer science to solving telecom's big problems around scalability, resilience, and survivability.

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 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

"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

This comprehensive book focuses on better big-data security for healthcare organizations. Following an extensive introduction to the Internet of Things (IoT) in healthcare including challenging topics and scenarios, it offers an in-depth analysis of medical body area networks with the 5th generation of IoT communication technology along with its nanotechnology. It also describes a novel strategic framework and computationally intelligent model to measure possible security vulnerabilities in the context of e-health. Moreover, the book addresses healthcare systems that handle large volumes of data driven by patients' records and health/personal information, including big-data-based knowledge management systems to support clinical decisions. Several of the issues faced in storing/processing big data are presented along with the available tools, technologies and algorithms to deal with those problems as well as a case study in healthcare analytics. Addressing trust, privacy, and security issues as well as the IoT and big-data challenges, the book highlights the advances in the field to guide engineers developing different IoT devices and evaluating the performance of different IoT techniques. Additionally, it explores the impact of such technologies on public, private, community, and hybrid scenarios in healthcare. This book offers professionals, scientists and engineers the latest technologies, techniques, and strategies for IoT and big data.

Would you like to use a consistent visual notation for drawing integration solutions? "Look inside the front cover." Do you want to harness the power of asynchronous systems without getting caught in the pitfalls? "See "Thinking Asynchronously" in the Introduction." Do you want to know which style of application integration is best for your purposes? "See Chapter 2, Integration Styles." Do you want to learn techniques for processing messages concurrently? "See Chapter 10, Competing Consumers and Message Dispatcher." Do you want to learn how you can track asynchronous messages as they flow across distributed systems? "See Chapter 11, Message History and Message Store." Do you want to understand how a system designed using integration patterns can be implemented using Java Web services, .NET message queuing, and a TIBCO-based publish-subscribe architecture? "See Chapter 9, Interlude: Composed Messaging." Utilizing years of practical experience, seasoned experts Gregor Hohpe and Bobby Woolf show how asynchronous messaging has proven to be the best strategy for enterprise integration success. However, building and deploying messaging solutions presents a number of problems for developers. " Enterprise Integration Patterns " provides an invaluable catalog of sixty-five patterns, with real-world solutions that demonstrate the formidable of messaging and help you to design effective messaging solutions for your enterprise. The authors also include examples covering a variety of different integration technologies, such as JMS, MSMQ, TIBCO ActiveEnterprise, Microsoft BizTalk, SOAP, and XSL. A case study describing a bond trading system illustrates the patterns in practice, and the book offers a look at emerging standards, as well as insights into what the future of enterprise integration might hold. This book provides a consistent vocabulary and visual notation framework to describe large-scale integration solutions across many technologies. It also explores in detail the advantages and limitations of asynchronous messaging architectures. The authors present practical advice on designing code that connects an application to a messaging system, and provide extensive information to help you determine when to send a message, how to route it to the proper destination, and how to monitor the health of a messaging system. If you want to know how to manage, monitor, and maintain a messaging system once it is in use, get this book. 0321200683B09122003

Hadoop in Action teaches readers how to use Hadoop and write MapReduce programs. The intended readers are programmers, architects, and project managers who have to process large amounts of data offline. Hadoop in Action will lead the reader from obtaining a copy of Hadoop to setting it up in a cluster and writing data analytic programs. The book begins by making the basic idea of Hadoop and MapReduce easier to grasp by applying the default Hadoop installation to a few easy-to-follow tasks, such as analyzing changes in word frequency across a body of documents. The book continues through the basic concepts of MapReduce applications developed using Hadoop, including a close look at framework components, use of Hadoop for a variety of data analysis tasks, and numerous examples of Hadoop in action. Hadoop in Action will explain how to use Hadoop and present design patterns and practices of programming MapReduce. MapReduce is a complex idea both conceptually and in its implementation, and Hadoop users are challenged to learn all the knobs and levers for running Hadoop. This book takes you beyond the mechanics of running Hadoop, teaching you to write meaningful programs in a MapReduce framework. This book assumes the reader will have a basic familiarity with Java, as most code examples will be written in Java. Familiarity with basic statistical concepts (e.g. histogram, correlation) will help the reader appreciate the more advanced data processing examples. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book.

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

Quickly find solutions to common programming problems encountered while processing big data. Content is presented in the popular problem-solution format. Look up the programming problem that you want to solve. Read the solution. Apply the solution directly in your own code. Problem solved! PySpark Recipes covers Hadoop and its shortcomings. The architecture of Spark, PySpark, and RDD are presented. You will learn to apply RDD to solve day-to-day big data problems. Python and NumPy are included and make it easy for new learners of PySpark to understand and adopt the model. What You Will Learn Understand the advanced features of PySpark2 and SparkSQL Optimize your code Program SparkSQL with Python Use Spark Streaming and Spark MLlib with Python Perform graph analysis with GraphFrames Who This Book Is For Data analysts, Python programmers, big data enthusiasts

It started with a spontaneous awakening of the chakras, although Katie didn't know exactly what was happening at the time. She felt an explosion of creativity, with spiritual awareness, insight and psychic abilities. She saw that reality was actually a dream state. These experiences were so powerful, Katie felt compelled to follow the spiritual path in her quest to hang on to the light that filled her.

Deep learning is often viewed as the exclusive domain of math PhDs and big tech companies. But as this hands-on guide demonstrates, programmers comfortable with Python can achieve impressive results in deep learning with little math background, small amounts of data, and minimal code. How? With fastai, the first library to provide a consistent interface to the most frequently used deep learning applications. Authors Jeremy Howard and Sylvain Gugger, the creators of fastai, show you how to train a model on a wide range of tasks using fastai and PyTorch. You'll also dive progressively further into deep learning theory to gain a complete understanding of the algorithms behind the scenes. Train models in computer vision, natural language processing, tabular data, and collaborative filtering Learn the latest deep learning techniques that matter most in practice Improve accuracy, speed, and reliability by understanding how deep learning models work Discover how to turn your models into web applications Implement deep learning algorithms from scratch Consider the ethical implications of your work Gain insight from the foreword by PyTorch cofounder, Soumith Chintala

This book will teach you how to move quickly from business questions to machine learning models in production. Using real-world examples implemented with Python and Jupyter notebooks, you'll learn about many the features and APIs of Amazon SageMaker on a wide spectrum of use cases: tabular data, computer vision, and natural language processing.

One of Mark Cuban's top reads for better understanding A.I. (inc.com, 2021) Your comprehensive entry-level guide to machine learning While machine learning expertise doesn't quite mean you can create your own Turing Test-proof android—as in the movie Ex Machina—it is a form of artificial intelligence and one of the most exciting technological means of identifying opportunities and solving problems fast and on a large scale. Anyone who masters the principles of machine learning is mastering a big part of our tech future and opening up incredible new directions in careers that include fraud detection, optimizing search results, serving real-time ads, credit-scoring, building accurate and sophisticated pricing models—and way, way more. Unlike most machine learning books, the fully updated 2nd Edition of Machine Learning For Dummies doesn't assume you have years of experience using programming languages such as Python (R source is also included in a downloadable form with comments and explanations), but lets you in on the ground floor, covering the entry-level materials that will get you up and running building models you need to perform practical tasks. It takes a look at the underlying—and fascinating—math principles that power machine learning but also shows that you don't need to be a math whiz to build fun new tools and apply them to your work and study. Understand the history of AI and machine learning Work with Python 3.8 and TensorFlow 2.x (and R as a download) Build and test your own models Use the latest datasets, rather than the worn out data found in other books Apply machine learning to real problems Whether you want to learn for college or to enhance your business or career performance, this friendly beginner's guide is your best introduction to machine learning, allowing you to become quickly confident using this amazing and fast-developing technology that's impacting lives for the better all over the world.

Summary Kafka Streams in Action teaches you everything you need to know to implement stream processing on data flowing into your Kafka platform, allowing you to focus on getting more from your data without sacrificing time or effort. Foreword by Neha Narkhede, Cocreator of Apache Kafka Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Not all stream-based applications require a dedicated processing cluster. The lightweight Kafka Streams library provides exactly the power and simplicity you need for message

handling in microservices and real-time event processing. With the Kafka Streams API, you filter and transform data streams with just Kafka and your application. About the Book Kafka Streams in Action teaches you to implement stream processing within the Kafka platform. In this easy-to-follow book, you'll explore real-world examples to collect, transform, and aggregate data, work with multiple processors, and handle real-time events. You'll even dive into streaming SQL with KSQL! Practical to the very end, it finishes with testing and operational aspects, such as monitoring and debugging. What's inside Using the KStreams API Filtering, transforming, and splitting data Working with the Processor API Integrating with external systems About the Reader Assumes some experience with distributed systems. No knowledge of Kafka or streaming applications required. About the Author Bill Bejeck is a Kafka Streams contributor and Confluent engineer with over 15 years of software development experience. Table of Contents PART 1 - GETTING STARTED WITH KAFKA STREAMS Welcome to Kafka Streams Kafka quicklyPART 2 - KAFKA STREAMS DEVELOPMENT Developing Kafka Streams Streams and state The KTable API The Processor APIPART 3 - ADMINISTERING KAFKA STREAMS Monitoring and performance Testing a Kafka Streams applicationPART 4 - ADVANCED CONCEPTS WITH KAFKA STREAMS Advanced applications with Kafka StreamsAPPENDIXES Appendix A - Additional configuration information Appendix B - Exactly once semantics

Every enterprise application creates data, whether it’s log messages, metrics, user activity, outgoing messages, or something else. And how to move all of this data becomes nearly as important as the data itself. If you’re an application architect, developer, or production engineer new to Apache Kafka, this practical guide shows you how to use this open source streaming platform to handle real-time data feeds. Engineers from Confluent and LinkedIn who are responsible for developing Kafka explain how to deploy production Kafka clusters, write reliable event-driven microservices, and build scalable stream-processing applications with this platform. Through detailed examples, you’ll learn Kafka’s design principles, reliability guarantees, key APIs, and architecture details, including the replication protocol, the controller, and the storage layer. Understand publish-subscribe messaging and how it fits in the big data ecosystem. Explore Kafka producers and consumers for writing and reading messages Understand Kafka patterns and use-case requirements to ensure reliable data delivery Get best practices for building data pipelines and applications with Kafka Manage Kafka in production, and learn to perform monitoring, tuning, and maintenance tasks Learn the most critical metrics among Kafka’s operational measurements Explore how Kafka’s stream delivery capabilities make it a perfect source for stream processing systems

This textbook presents an end-to-end Internet of Things (IoT) architecture that comprises of devices, network, compute, storage, platform, applications along with management and security components with focus on the missing functionality in the current state of the art. As with the first edition, it is organized into six main parts: an IoT reference model; Fog computing and the drivers; IoT management and applications ranging from smart homes to manufacturing and energy conservation solutions; Smart Services in IoT; IoT standards; and case studies. The textbook edition features a new chapter entitled The Blockchain in IoT, updates based on latest standards and technologies, and new slide ware for professors. It features a full suite of classroom material for easy adoption.

“Big data” has become a commonly used term to describe large-scale and complex data sets which are difficult to manage and analyze using standard data management methodologies. With applications across sectors and fields of study, the implementation and possible uses of big data are limitless. Effective Big Data Management and Opportunities for Implementation explores emerging research on the ever-growing field of big data and facilitates further knowledge development on methods for handling and interpreting large data sets. Providing multi-disciplinary perspectives fueled by international research, this publication is designed for use by data analysts, IT professionals, researchers, and graduate-level students interested in learning about the latest trends and concepts in big data.

Run queries and analysis on big data clusters across relational and non relational databases KEY FEATURES ? Connect to Hadoop, Azure, Spark, Oracle, Teradata, Cassandra, MongoDB, CosmosDB, MySQL, PostgreSQL, MariaDB, and SAP HANA. ? Numerous techniques on how to query data and troubleshoot Polybase for better data analytics. ? Exclusive coverage on Azure Synapse Analytics and building Big Data clusters. DESCRIPTION This book brings exciting coverage on establishing and managing data virtualization using polybase. This book teaches how to configure polybase on almost all relational and nonrelational databases. You will learn to set up the test environment for any tool or software instantly without hassle. You will practice how to design and build some of the high performing data warehousing solutions and that too in a few minutes of time. You will almost become an expert in connecting to all databases including hadoop, cassandra, MySQL, PostgreSQL, MariaDB and Oracle database. This book also brings exclusive coverage on how to build data clusters on Azure and using Azure Synapse Analytics. By the end of this book, you just don't administer the polybase for managing big data clusters but rather you learn to optimize and boost the performance for enabling data analytics and ease of data accessibility. WHAT YOU WILL LEARN ? Learn to configure Polybase and process Transact SQL queries with ease. ? Create a Docker container with SQL Server 2019 on Windows and Polybase. ? Establish SQL Server instance with any other software or tool using Polybase ? Connect with Cassandra, MongoDB, MySQL, PostgreSQL, MariaDB, and IBM DB2. WHO THIS BOOK IS FOR This book is for database developers and administrators familiar with the SQL language and command prompt. Managers and decision-makers will also find this book useful. No prior knowledge of any other technology or language is required. TABLE OF CONTENTS 1. What is Data Virtualization (Polybase) 2. History of Polybase 3. Polybase current state 4. Differences with other technologies 5. Usage 6. Future 7. SQL Server 8. Hadoop Cloudera and Hortonworks 9. Windows Azure Storage Blob 10. Spark 11. From Azure Synapse Analytics 12. From Big Data Clusters 13. Oracle 14. Teradata 15. Cassandra 16. MongoDB 17. CosmosDB 18. MySQL 19. PostgreSQL 20. MariaDB 21. SAP HANA 22. IBM DB2 23. Excel

Frank Kane's hands-on Spark training course, based on his bestselling Taming Big Data with Apache Spark and Python video, now available in a book. Understand and analyze large data sets using Spark on a single system or on a cluster. About This Book Understand how Spark can be distributed across computing clusters Develop and run Spark jobs efficiently using Python A hands-on tutorial by Frank Kane with over 15 real-world examples teaching you Big Data processing with Spark Who This Book Is For If you are a data scientist or data analyst who wants to learn Big Data processing using Apache Spark and Python, this book is for you. If you have some programming experience in Python, and want to learn how to process large amounts of data using Apache Spark, Frank Kane's Taming Big Data with Apache Spark and Python will also help you. What You Will Learn Find out how you can identify Big Data problems as Spark problems Install and run Apache Spark on your computer or on a cluster Analyze large data sets across many CPUs using Spark's Resilient Distributed Datasets Implement machine learning on Spark using the MLlib library Process continuous streams of data in real time using the Spark streaming module Perform complex network analysis using Spark's GraphX library Use Amazon's Elastic MapReduce service to run your Spark jobs on a cluster In Detail Frank Kane's Taming Big Data with Apache Spark and Python is your companion to learning Apache Spark in a hands-on manner. Frank will start you off by teaching you how to set up Spark on a single system or on a cluster, and you'll soon move on to analyzing large data sets using Spark RDD, and developing and running effective Spark jobs quickly using Python. Apache Spark has emerged as the next big thing in the Big Data domain – quickly rising from an ascending technology to an established superstar in just a matter of years. Spark allows you to quickly extract actionable insights from large amounts of data, on a real-time basis, making it an essential tool in many modern businesses. Frank has packed this book with over 15 interactive, fun-filled examples relevant to the real world, and he will empower you to understand the Spark ecosystem and implement production-grade real-time Spark projects with ease. Style and approach Frank Kane's Taming Big Data with Apache Spark and Python is a hands-on tutorial with over 15 real-world examples carefully explained by Frank in a step-by-step manner. The examples vary in complexity, and you can move through them at your own pace.

Summary Functional Programming in Scala is a serious tutorial for programmers looking to learn FP and apply it to the everyday business of coding. The book guides readers from basic techniques to advanced topics in a logical, concise, and clear progression. In it, you'll find concrete examples and exercises that open up the world of functional programming. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Functional programming (FP) is a style of software development emphasizing functions that don't depend on program state. Functional code is easier to test and reuse, simpler to parallelize, and less prone to bugs than other code. Scala is an emerging JVM language that offers strong support for FP. Its familiar syntax and transparent interoperability with Java make Scala a great place to start learning FP. About the Book Functional Programming in Scala is a serious tutorial for programmers looking to learn FP and apply it to their everyday work. The book guides readers from basic techniques to advanced topics in a logical, concise, and clear progression. In it, you'll find concrete examples and exercises that open up the world of functional programming. This book assumes no prior experience with functional programming. Some prior exposure to Scala or Java is helpful. What's Inside Functional programming concepts The whys and hows of FP How to write multicore programs Exercises and checks for understanding About the Authors Paul Chiusano and Rúnar Bjarnason are recognized experts in functional programming with Scala and are core contributors to the Scalaz library. Table of Contents PART 1 INTRODUCTION TO FUNCTIONAL PROGRAMMING What is functional programming? Getting started with functional programming in Scala Functional data structures Handling errors without exceptions Strictness and laziness Purely functional state PART 2 FUNCTIONAL DESIGN AND COMBINATOR LIBRARIES Purely functional parallelism Property-based testing Parser combinators PART 3 COMMON STRUCTURES IN FUNCTIONAL DESIGN Monoids Monads Applicative and traversable functors PART 4 EFFECTS AND I/O External effects and I/O Local effects and mutable state Stream processing and incremental I/O

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