Top 50 Apache Spark Interview Questions Answers

This is likewise one of the factors by obtaining the soft documents of this **Top 50 Apache Spark Interview Questions Answers** by online. You might not require more get older to spend to go to the book establishment as without difficulty as search for them. In some cases, you likewise complete not discover the revelation Top 50 Apache Spark Interview Questions Answers that you are looking for. It will unquestionably squander the time.

However below, in the same way as you visit this web page, it will be hence very simple to acquire as skillfully as download lead Top 50 Apache Spark Interview Questions Answers

It will not agree to many era as we accustom before. You can complete it while enactment something else at home and even in your workplace. consequently easy! So, are you question? Just exercise just what we allow under as without difficulty as review **Top 50 Apache Spark Interview Questions Answers** what you afterward to read!

Python Machine Learning - Sebastian Raschka 2015-09-23
Unlock deeper insights into Machine Leaning with this vital guide to cutting-edge predictive analytics About This Book Leverage Python's most powerful open-source libraries for deep learning, data wrangling, and data visualization Learn effective strategies and best practices to improve and optimize machine learning systems and algorithms Ask - and answer - tough questions of your data with robust statistical models, built for a range of datasets Who This Book Is For If you want to find out how to use Python to start answering critical questions of your data, pick up Python Machine Learning - whether you want to get started from scratch or want to extend your data science knowledge, this is an essential and unmissable resource. What You Will Learn Explore how to use different machine learning models to ask different questions of your data Learn how to build neural networks using Keras and Theano Find out how to write clean and elegant Python code that will optimize the strength of your algorithms Discover how to embed your machine learning model in a web application for increased accessibility Predict continuous target outcomes using regression analysis Uncover hidden patterns and structures in data with clustering Organize data using effective pre-processing techniques Get to grips with sentiment analysis to delve deeper into textual and social media data In Detail Machine learning and predictive analytics are transforming the way businesses and other organizations operate. Being able to understand trends and patterns in complex data is critical to success, becoming one of the key strategies for unlocking growth in a challenging contemporary marketplace. Python can help you deliver key insights into your data - its unique capabilities as a language let you build sophisticated algorithms and statistical models that can reveal new perspectives and answer key questions that are vital for success. Python Machine Learning gives you access to the world of predictive analytics and demonstrates why Python is one of the world's leading data science languages. If you want to ask better questions of data, or need to improve and extend the capabilities of your machine learning systems, this practical data science book is invaluable. Covering a wide range of powerful Python libraries, including scikit-learn, Theano, and Keras, and

featuring guidance and tips on everything from sentiment analysis to neural networks, you'll soon be able to answer some of the most important questions facing you and your organization. Style and approach Python Machine Learning connects the fundamental theoretical principles behind machine learning to their practical application in a way that focuses you on asking and answering the right questions. It walks you through the key elements of Python and its powerful machine learning libraries, while demonstrating how to get to grips with a range of statistical models.

Learning Spark - Jules S. Damji 2020-07-16

Data is bigger, arrives faster, and comes in a variety of formats—and it all needs to be processed at scale for analytics or machine learning. But how can you process such varied workloads efficiently? Enter Apache Spark. Updated to include Spark 3.0, this second edition shows data engineers and data scientists why structure and unification in Spark matters. Specifically, this book explains how to perform simple and complex data analytics and employ machine learning algorithms. Through step-by-step walk-throughs, code snippets, and notebooks, you'll be able to: Learn Python, SQL, Scala, or Java high-level Structured APIs Understand Spark operations and SQL Engine Inspect, tune, and debug Spark operations with Spark configurations and Spark UI Connect to data sources: JSON, Parquet, CSV, Avro, ORC, Hive, S3, or Kafka Perform analytics on batch and streaming data using Structured Streaming Build reliable data pipelines with open source Delta Lake and Spark Develop machine learning pipelines with MLlib and productionize models using MLflow

Top 50 Apache Spark Interview Questions and Answers - Knowledge Powerhouse 2017-03-18

Introduction: Top 50 Apache Spark Interview Questions & Answers Apache Spark is a highly popular trend in technology world. There is a growing demand for Data Engineer jobs with Apache Spark knowledge in IT Industry. This book contains technical interview questions that an interviewer asks for Apache Spark. Each question is accompanied with an answer so that you can prepare for job interview in short time. We have compiled this list after attending dozens of technical interviews in top-notch companies like- Amazon, Netflix, Uber etc.Often, these questions and concepts are used in our daily work. There is a sample answer with each question. But try to answer these questions in your own words.After going through this book 2-3 times, you will be well prepared to face interview of Apache Spark topic for Data Engineer position. How will this book help me? By reading this book, you do not have to spend time searching the Internet for Apache Spark Data Engineer interview questions. We have already compiled the list of most popular and latest Apache Spark Data Engineer Interview questions. Are there answers in this book? Yes, in this book each question is followed by an answer. So you can save time in interview preparation. What is the best way of reading this book? You have to first do a slow reading of all the questions in this book. Once you go through them in the first pass try to go through the difficult questions. After going through this book 2-3 times, you will be well prepared to face Apache Spark Data Engineer interview in IT. What is the level of questions in this book? This book contains questions that are good for Software Engineer, Senior Software Engineer, Principal Engineer and Associate Architect level. What are the sample questions in this book? How will you minimize data transfer while working with Apache Spark? How does Spark Streaming work internally? What are the main features of Apache Spark? What is a Resilient Distribution Dataset in Apache Spark? What is a Transformation in Apache Spark? What are security options in Apache Spark? What are the two ways to create RDD in Spark? What are the main operations that can be done on a RDD in Apache Spark? What is a Shuffle operation in Spark? What are the operations that can cause a shuffle in Spark? What is purpose of Spark

SQL? What is a DataFrame in Spark SQL? What is a Parquet file in Spark? What is the difference between Apache Spark and Apache Hadoop MapReduce? What are the main languages supported by Apache Spark? What is the use of SparkContext in Apache Spark? Do we need HDFS for running Spark application? What is Spark Streaming? What is a Pipeline in Apache Spark? How does Pipeline work in Apache Spark? What is the difference between Transformer and Estimator in Apache Spark? What are the different types of Cluster Managers in Apache Spark? What is the main use of MLib in Apache Spark? What is the Checkpointing in Apache Spark? What is an Accumulator in Apache Spark? What is a Broadcast variable in Apache Spark? What is Structured Streaming in Apache Spark? What is a Property Graph? What is Neighborhood Aggregation in Spark? What are different Persistence levels in Apache Spark? How will you select the storage level in Apache Spark? What are the options in Spark to create a Graph? What are the basic Graph operators in Spark? What is the partitioning approach used in GraphX of Apache Spark? http://www.knowledgepowerhouse.com

Grokking the System Design Interview - Design Gurus 2021-12-18

This book (also available online at www.designgurus.org) by Design Gurus has helped 60k+ readers to crack their system design interview (SDI). System design questions have become a standard part of the software engineering interview process. These interviews determine your ability to work with complex systems and the position and salary you will be offered by the interviewing company. Unfortunately, SDI is difficult for most engineers, partly because they lack experience developing large-scale systems and partly because SDIs are unstructured in nature. Even engineers who've some experience building such systems aren't comfortable with these interviews, mainly due to the open-ended nature of design problems that don't have a standard answer. This book is a comprehensive guide to master SDIs. It was created by hiring managers who have worked for Google, Facebook, Microsoft, and Amazon. The book contains a carefully chosen set of questions that have been repeatedly asked at top companies. What's inside? This book is divided into two parts. The first part includes a step-by-step guide on how to answer a system design question in an interview, followed by famous system design case studies. The second part of the book includes a glossary of system design concepts. Table of Contents First Part: System Design Interviews: A step-by-step guide. Designing a URL Shortening service like TinyURL. Designing Pastebin. Designing Instagram. Designing Dropbox. Designing Facebook Messenger. Designing Twitter. Designing YouTube or Netflix. Designing Typeahead Suggestion. Designing an API Rate Limiter. Designing Twitter Search. Designing a Web Crawler. Designing Facebook's Newsfeed. Designing Yelp or Nearby Friends. Designing Uber backend. Designing Ticketmaster. Second Part: Key Characteristics of Distributed Systems. Load Balancing. Caching. Data Partitioning. Indexes. Proxies. Redundancy and Replication. SQL vs. NoSQL. CAP Theorem. PACELC Theorem. Consistent Hashing. Long-Polling vs. WebSockets vs. Server-Sent Events. Bloom Filters. Quorum. Leader and Follower. Heartbeat. Checksum. About the Authors Designed Gurus is a platform that offers online courses to help software engineers prepare for coding and system design interviews. Learn more about our courses at www.designgurus.org.

Moving Hadoop to the Cloud - Bill Havanki 2017-07-14

Until recently, Hadoop deployments existed on hardware owned and run by organizations. Now, of course, you can acquire the computing resources and network connectivity to run Hadoop clusters in the cloud. But there’s a lot more to deploying Hadoop to the public cloud than simply renting machines. This hands-on guide shows developers and systems administrators familiar with Hadoop how to install, use, and manage cloud-born clusters efficiently. You’ll learn how to architect

clusters that work with cloud-provider features—not just to avoid pitfalls, but also to take full advantage of these services. You'll also compare the Amazon, Google, and Microsoft clouds, and learn how to set up clusters in each of them. Learn how Hadoop clusters run in the cloud, the problems they can help you solve, and their potential drawbacks Examine the common concepts of cloud providers, including compute capabilities, networking and security, and storage Build a functional Hadoop cluster on cloud infrastructure, and learn what the major providers require Explore use cases for high availability, relational data with Hive, and complex analytics with Spark Get patterns and practices for running cloud clusters, from designing for price and security to dealing with maintenance

Spark: The Definitive Guide - Bill Chambers 2018-02-08

Learn how to use, deploy, and maintain Apache Spark with this comprehensive guide, written by the creators of the open-source cluster-computing framework. With an emphasis on improvements and new features in Spark 2.0, authors Bill Chambers and Matei Zaharia break down Spark topics into distinct sections, each with unique goals. You'll explore the basic operations and common functions of Spark's structured APIs, as well as Structured Streaming, a new high-level API for building end-to-end streaming applications. Developers and system administrators will learn the fundamentals of monitoring, tuning, and debugging Spark, and explore machine learning techniques and scenarios for employing MLlib, Spark's scalable machine-learning library. Get a gentle overview of big data and Spark Learn about DataFrames, SQL, and Datasets—Spark's core APIs—through worked examples Dive into Spark's low-level APIs, RDDs, and execution of SQL and DataFrames Understand how Spark runs on a cluster Debug, monitor, and tune Spark clusters and applications Learn the power of Structured Streaming, Spark's stream-processing engine Learn how you can apply MLlib to a variety of problems, including classification or recommendation

Top 50 SQL Tricky Interview Questions - Knowledge Powerhouse 2016-12-11

This book contains tricky and nasty SQL interview questions that an interviewer asks. It is a compilation of advanced SQL interview questions after attending dozens of technical interviews in top-notch companies like- Oracle, Google, Ebay, Amazon etc.Each question is accompanied with an answer because you want to save your time while preparing for an interview.The difficulty rating on these Questions varies from a Junior level programmer to Architect level. Sample Questions are:How can we retrieve alternate records from a table in Oracle?Given a list of student names and grade. Write a query to print a comma separated list of student names in a grade.Write SQL Query to get Student Name and number of Students in same grade.Write SQL query to delete duplicate rows in a table?Write SQL query to get the second highest salary among all Employees?Write SQL Query to get Employee Name, Manager ID and number of employees in the department?Write SQL query to get the nth highest salary among all Employees.Given an Employee table with Manager_ID as column, print First name, Manager ID and Level of employees in Organization Structure?Why is the difference between NVL and NVL2 functions in SQL?What is the difference between UNION and UNION ALL?What are the reasons for de-normalizing the data?What is a Pseudocolumn?How can you find 10 employees with Odd number as Employee ID?What is the difference between DELETE and TRUNCATE in SQL?Which SQL feature can be used to view data in a table sequentially?What are the differences between CASE and DECODE in SQL?Write a SQL Query to get the Quarter from date.http://www.knowledgepowerhouse.com

Learning Spark - Holden Karau 2015-01-28

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

SAP® MM Questions and Answers - Kogent Inc 2010-10-25

Designed for SAP users as a quick reference or for computer science and business students, SAP MM Questions and Answers includes all the major concepts related to SAP MM functionality, technical configuration, and implementation in an easy-to-understand question and answer format. It discusses the new aspects related to SAP ERP 6.0 and all the important MM codes and concepts for materials and vendors, including clients, company codes, plants, storage locations, purchase organizations, etc. The organized and accessible format allows the reader to quickly find the questions on specific subjects and provides all of the details to pass certification exams in a step-by-step, easy-to-read method of instruction.

Data Algorithms - Mahmoud Parsian 2015-07-13

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

Functional Programming in Scala - Paul Chiusano 2014-09-01

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

Spark GraphX in Action - Michael Malak 2016-06-12

Summary Spark GraphX in Action starts out with an overview of Apache Spark and the GraphX graph processing API. This example-based tutorial then teaches you how to configure GraphX and how to use it interactively. Along the way, you'll collect practical techniques for enhancing applications and applying machine learning algorithms to graph data. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology GraphX is a powerful graph processing API for the Apache Spark analytics engine that lets you draw insights from large datasets. GraphX gives you unprecedented speed and capacity for running massively parallel and machine learning algorithms. About the Book Spark GraphX in Action begins with the big picture of what graphs can be used for. This example-based tutorial teaches you how to use GraphX interactively. You'll start with a crystal-clear introduction to building big data graphs from regular data, and then explore the problems and possibilities of implementing graph algorithms and architecting graph processing pipelines. Along the way, you'll collect practical techniques for enhancing applications and applying machine learning algorithms to graph data. What's Inside Understanding graph technology Using the GraphX API Developing algorithms for big graphs Machine learning with graphs Graph visualization About the Reader Readers should be comfortable writing code. Experience with Apache Spark and Scala is not required. About the Authors Michael Malak has worked on Spark applications for Fortune 500 companies since early 2013. Robin East has worked as a consultant to large organizations for over 15 years and is a data scientist at Worldpay. Table of Contents PART 1 SPARK AND GRAPHS Two important technologies: Spark and graphs GraphX quick start Some fundamentals PART 2 CONNECTING VERTICES GraphX Basics Built-in algorithms Other useful graph algorithms Machine learning PART 3 OVER THE ARC The missing algorithms Performance and monitoring Other languages and tools

A Guide to the Project Management Body of Knowledge (PMBOK® Guide) - Seventh Edition and The Standard for Project Management (BRAZILIAN PORTUGUESE) - Project Management Institute Project Management Institute 2021-08-01

PMBOK&® Guide is the go-to resource for project management practitioners. The project management profession has significantly evolved due to emerging technology, new approaches and rapid market changes. Reflecting this

evolution, The Standard for Project Management enumerates 12 principles of project management and the PMBOK&® Guide &- Seventh Edition is structured around eight project performance domains.This edition is designed to address practitioners' current and future needs and to help them be more proactive, innovative and nimble in enabling desired project outcomes.This edition of the PMBOK&® Guide:•Reflects the full range of development approaches (predictive, adaptive, hybrid, etc.);•Provides an entire section devoted to tailoring the development approach and processes;•Includes an expanded list of models, methods, and artifacts;•Focuses on not just delivering project outputs but also enabling outcomes; and• Integrates with PMIstandards+™ for information and standards application content based on project type, development approach, and industry sector.

Stream Processing with Apache Spark - Gerard Maas 2019-06-05

Before you can build analytics tools to gain quick insights, you first need to know how to process data in real time. With this practical guide, developers familiar with Apache Spark will learn how to put this in-memory framework to use for streaming data. You'll discover how Spark enables you to write streaming jobs in almost the same way you write batch jobs. Authors Gerard Maas and François Garillot help you explore the theoretical underpinnings of Apache Spark. This comprehensive guide features two sections that compare and contrast the streaming APIs Spark now supports: the original Spark Streaming library and the newer Structured Streaming API. Learn fundamental stream processing concepts and examine different streaming architectures Explore Structured Streaming through practical examples; learn different aspects of stream processing in detail Create and operate streaming jobs and applications with Spark Streaming; integrate Spark Streaming with other Spark APIs Learn advanced Spark Streaming techniques, including approximation algorithms and machine learning algorithms Compare Apache Spark to other stream processing projects, including Apache Storm, Apache Flink, and Apache Kafka Streams

Big Data Hadoop Interview Guide - Vishwanathan Narayanan 2021

A power-packed guide with solutions to crack a Big data Hadoop interview, this book covers many interview questions and the best possible ways to answer them, and provides real-world examples that will help you understand the concepts of Big Data. --

SAS Certified Specialist Prep Guide - SAS Institute 2019-02-11

The SAS® Certified Specialist Prep Guide: Base Programming Using SAS® 9.4 prepares you to take the new SAS 9.4 Base Programming -- Performance-Based Exam. This is the official guide by the SAS Global Certification Program. This prep guide is for both new and experienced SAS users, and it covers all the objectives that are tested on the exam. New in this edition is a workbook whose sample scenarios require you to write code to solve problems and answer questions. Answers for the chapter quizzes and solutions for the sample scenarios in the workbook are included. You will also find links to exam objectives, practice exams, and other resources such as the Base SAS® glossary and a list of practice data sets. Major topics include importing data, creating and modifying SAS data sets, and identifying and correcting both data syntax and programming logic errors. All exam topics are covered in these chapters: Setting Up Practice Data Basic Concepts Accessing Your Data Creating SAS Data Sets Identifying and Correcting SAS Language Errors Creating Reports Understanding DATA Step Processing BY-Group Processing Creating and Managing Variables Combining SAS Data Sets Processing Data with DO Loops SAS Formats and Informats SAS Date, Time, and Datetime Values Using Functions to Manipulate Data Producing Descriptive Statistics Creating Output Practice Programming Scenarios (Workbook)

Programming Hive - Edward Capriolo 2012-09-26

Describes the features and functions

of Apache Hive, the data infrastructure for Hadoop.

Top 1000 Java Interview Questions and Answers: Includes Spring, Hibernate, Microservices, GIT, Maven, JSP, AWS, Cloud Computing - Knowledge Powerhouse 2018-05-06

This is the ultimate book for interview preparation for Java jobs. It has questions on Java, Stream, Collections, Multi-threading, Spring, Hibernate, JSP, Design patterns, GIT, Maven, AWS and Cloud computing. It is a digest of questions from multiple sources. It covers almost all the technical areas of an interview for Java engineer position. The difficulty level of questions in this book vary from beginner to expert level.Once you go through this book, you will be very well prepared for facing Java interview for an experienced Software Developer.This books also contains Java tricky Interview questions, Java 8, Microserivces and AWS questions. Technical job applicants save pervious time in interview preparation by reading this book. You do not have to waste time in searching for questions and answers online. This books is your main book for Java based jobs.

Apache Spark Graph Processing - Rindra Ramamonjison 2015-09-10

Build, process and analyze large-scale graph data effectively with Spark About This Book Find solutions for every stage of data processing from loading and transforming graph data to Improve the scalability of your graphs with a variety of real-world applications with complete Scala code. A concise guide to processing large-scale networks with Apache Spark. Who This Book Is For This book is for data scientists and big data developers who want to learn the processing and analyzing graph datasets at scale. Basic programming experience with Scala is assumed. Basic knowledge of Spark is assumed. What You Will Learn Write, build and deploy Spark applications with the Scala Build Tool. Build and analyze large-scale network datasets Analyze and transform graphs using RDD and graph-specific operations Implement new custom graph operations tailored to specific needs. Develop iterative and efficient graph algorithms using message aggregation and Pregel abstraction Extract subgraphs and use it to discover common clusters Analyze graph data and solve various data science problems using real-world datasets. In Detail Apache Spark is the next standard of open-source cluster-computing engine for processing big data. Many practical computing problems concern large graphs, like the Web graph and various social networks. The scale of these graphs - in some cases billions of vertices, trillions of edges - poses challenges to their efficient processing. Apache Spark GraphX API combines the advantages of both data-parallel and graph-parallel systems by efficiently expressing graph computation within the Spark data-parallel framework. This book will teach the user to do graphical programming in Apache Spark, apart from an explanation of the entire process of graphical data analysis. You will journey through the creation of graphs, its uses, its exploration and analysis and finally will also cover the conversion of graph elements into graph structures. This book begins with an introduction of the Spark system, its libraries and the Scala Build Tool. Using a hands-on approach, this book will quickly teach you how to install and leverage Spark interactively on the command line and in a standalone Scala program. Then, it presents all the methods for building Spark graphs using illustrative network datasets. Next, it will walk you through the process of exploring, visualizing and analyzing different network characteristics. This book will also teach you how to transform raw datasets into a usable form. In addition, you will learn powerful operations that can be used to transform graph elements and graph structures. Furthermore, this book also teaches how to create custom graph operations that are tailored for specific needs with efficiency in mind. The later chapters of this book cover more advanced topics such as clustering graphs, implementing graph-parallel iterative algorithms

and learning methods from graph data. Style and approach A step-by-step guide that will walk you through the key ideas and techniques for processing big graph data at scale, with practical examples that will ensure an overall understanding of the concepts of Spark.

Android Development Interview Questions You'll Most Likely Be Asked - Vibrant Publishers 2012-01-22

Android Development Interview Questions You'll Most Likely Be Asked is a perfect companion to stand ahead above the rest in today's competitive job market.

Scala and Spark for Big Data Analytics - Md. Rezaul Karim 2017-07-25

Harness the power of Scala to program Spark and analyze tonnes of data in the blink of an eye! About This Book Learn Scala's sophisticated type system that combines Functional Programming and object-oriented concepts Work on a wide array of applications, from simple batch jobs to stream processing and machine learning Explore the most common as well as some complex use-cases to perform large-scale data analysis with Spark Who This Book Is For Anyone who wishes to learn how to perform data analysis by harnessing the power of Spark will find this book extremely useful. No knowledge of Spark or Scala is assumed, although prior programming experience (especially with other JVM languages) will be useful to pick up concepts quicker. What You Will Learn Understand object-oriented & functional programming concepts of Scala In-depth understanding of Scala collection APIs Work with RDD and DataFrame to learn Spark's core abstractions Analysing structured and unstructured data using SparkSQL and GraphX Scalable and fault-tolerant streaming application development using Spark structured streaming Learn machine-learning best practices for classification, regression, dimensionality reduction, and recommendation system to build predictive models with widely used algorithms in Spark MLlib & ML Build clustering models to cluster a vast amount of data Understand tuning, debugging, and monitoring Spark applications Deploy Spark applications on real clusters in Standalone, Mesos, and YARN In Detail Scala has been observing wide adoption over the past few years, especially in the field of data science and analytics. Spark, built on Scala, has gained a lot of recognition and is being used widely in productions. Thus, if you want to leverage the power of Scala and Spark to make sense of big data, this book is for you. The first part introduces you to Scala, helping you understand the object-oriented and functional programming concepts needed for Spark application development. It then moves on to Spark to cover the basic abstractions using RDD and DataFrame. This will help you develop scalable and fault-tolerant streaming applications by analyzing structured and unstructured data using SparkSQL, GraphX, and Spark structured streaming. Finally, the book moves on to some advanced topics, such as monitoring, configuration, debugging, testing, and deployment. You will also learn how to develop Spark applications using SparkR and PySpark APIs, interactive data analytics using Zeppelin, and in-memory data processing with Alluxio. By the end of this book, you will have a thorough understanding of Spark, and you will be able to perform full-stack data analytics with a feel that no amount of data is too big. Style and approach Filled with practical examples and use cases, this book will hot only help you get up and running with Spark, but will also take you farther down the road to becoming a data scientist.

Artificial Intelligence with Python - Prateek Joshi 2017-01-27

Build real-world Artificial Intelligence applications with Python to intelligently interact with the world around you About This Book Step into the amazing world of intelligent apps using this comprehensive guide Enter the world of Artificial Intelligence, explore it, and create your own applications Work through simple yet insightful examples that will get you up and running with Artificial Intelligence in no time

Who This Book Is For This book is for Python developers who want to build real-world Artificial Intelligence applications. This book is friendly to Python beginners, but being familiar with Python would be useful to play around with the code. It will also be useful for experienced Python programmers who are looking to use Artificial Intelligence techniques in their existing technology stacks. What You Will Learn Realize different classification and regression techniques Understand the concept of clustering and how to use it to automatically segment data See how to build an intelligent recommender system Understand logic programming and how to use it Build automatic speech recognition systems Understand the basics of heuristic search and genetic programming Develop games using Artificial Intelligence Learn how reinforcement learning works Discover how to build intelligent applications centered on images, text, and time series data See how to use deep learning algorithms and build applications based on it In Detail Artificial Intelligence is becoming increasingly relevant in the modern world where everything is driven by technology and data. It is used extensively across many fields such as search engines, image recognition, robotics, finance, and so on. We will explore various real-world scenarios in this book and you'll learn about various algorithms that can be used to build Artificial Intelligence applications. During the course of this book, you will find out how to make informed decisions about what algorithms to use in a given context. Starting from the basics of Artificial Intelligence, you will learn how to develop various building blocks using different data mining techniques. You will see how to implement different algorithms to get the best possible results, and will understand how to apply them to real-world scenarios. If you want to add an intelligence layer to any application that's based on images, text, stock market, or some other form of data, this exciting book on Artificial Intelligence will definitely be your guide! Style and approach This highly practical book will show you how to implement Artificial Intelligence. The book provides multiple examples enabling you to create smart applications to meet the needs of your organization. In every chapter, we explain an algorithm, implement it, and then build a smart application.

Hands-On Data Science and Python Machine Learning - Frank Kane 2017-07-31

This book covers the fundamentals of machine learning with Python in a concise and dynamic manner. It covers data mining and large-scale machine learning using Apache Spark. About This Book Take your first steps in the world of data science by understanding the tools and techniques of data analysis Train efficient Machine Learning models in Python using the supervised and unsupervised learning methods Learn how to use Apache Spark for processing Big Data efficiently Who This Book Is For If you are a budding data scientist or a data analyst who wants to analyze and gain actionable insights from data using Python, this book is for you. Programmers with some experience in Python who want to enter the lucrative world of Data Science will also find this book to be very useful, but you don't need to be an expert Python coder or mathematician to get the most from this book. What You Will Learn Learn how to clean your data and ready it for analysis Implement the popular clustering and regression methods in Python Train efficient machine learning models using decision trees and random forests Visualize the results of your analysis using Python's Matplotlib library Use Apache Spark's MLlib package to perform machine learning on large datasets In Detail Join Frank Kane, who worked on Amazon and IMDb's machine learning algorithms, as he guides you on your first steps into the world of data science. Hands-On Data Science and Python Machine Learning gives you the tools that you need to understand and explore the core topics in the field, and the confidence and practice to build and analyze your own machine learning

models. With the help of interesting and easy-to-follow practical examples, Frank Kane explains potentially complex topics such as Bayesian methods and K-means clustering in a way that anybody can understand them. Based on Frank's successful data science course, Hands-On Data Science and Python Machine Learning empowers you to conduct data analysis and perform efficient machine learning using Python. Let Frank help you unearth the value in your data using the various data mining and data analysis techniques available in Python, and to develop efficient predictive models to predict future results. You will also learn how to perform large-scale machine learning on Big Data using Apache Spark. The book covers preparing your data for analysis, training machine learning models, and visualizing the final data analysis. Style and approach This comprehensive book is a perfect blend of theory and hands-on code examples in Python which can be used for your reference at any time.

OCP Oracle Certified Professional Java SE 11 Programmer I Study Guide - Jeanne Boyarsky 2019-11-19

The comprehensive study aide for those preparing for the new Oracle Certified Professional Java SE Programmer I Exam 1Z0-815 Used primarily in mobile and desktop application development, Java is a platform-independent, object-oriented programming language. It is the principal language used in Android application development as well as a popular language for client-side cloud applications. Oracle has updated its Java Programmer certification tracks for Oracle Certified Professional. OCP Oracle Certified Professional Java SE 11 Programmer I Study Guide covers 100% of the exam objectives, ensuring that you are thoroughly prepared for this challenging certification exam. This comprehensive, in-depth study guide helps you develop the functional-programming knowledge required to pass the exam and earn certification. All vital topics are covered, including Java building blocks, operators and loops, String and StringBuilder, Array and ArrayList, and more. Included is access to Sybex's superior online interactive learning environment and test bank—containing self-assessment tests, chapter tests, bonus practice exam questions, electronic flashcards, and a searchable glossary of important terms. This indispensable guide: Clarifies complex material and strengthens your comprehension and retention of key topics Covers all exam objectives such as methods and encapsulation, exceptions, inheriting abstract classes and interfaces, and Java 8 Dates and Lambda Expressions Explains object-oriented design principles and patterns Helps you master the fundamentals of functional programming Enables you to create Java solutions applicable to real-world scenarios There are over 9 millions developers using Java around the world, yet hiring managers face challenges filling open positions with qualified candidates. The OCP Oracle Certified Professional Java SE 11 Programmer I Study Guide will help you take the next step in your career.

Spark - Ilya Ganelin 2016-03-21

Production-targeted Spark guidance with real-world use cases Spark: Big Data Cluster Computing in Production goes beyond general Spark overviews to provide targeted guidance toward using lightning-fast big-data clustering in production. Written by an expert team well-known in the big data community, this book walks you through the challenges in moving from proof-of-concept or demo Spark applications to live Spark in production. Real use cases provide deep insight into common problems, limitations, challenges, and opportunities, while expert tips and tricks help you get the most out of Spark performance. Coverage includes Spark SQL, Tachyon, Kerberos, ML Lib, YARN, and Mesos, with clear, actionable guidance on resource scheduling, db connectors, streaming, security, and much more. Spark has become the tool of choice for many Big Data problems, with more active contributors than any other Apache Software project. General

introductory books abound, but this book is the first to provide deep insight and real-world advice on using Spark in production. Specific guidance, expert tips, and invaluable foresight make this guide an incredibly useful resource for real production settings. Review Spark hardware requirements and estimate cluster size Gain insight from real-world production use cases Tighten security, schedule resources, and fine-tune performance Overcome common problems encountered using Spark in production Spark works with other big data tools including MapReduce and Hadoop, and uses languages you already know like Java, Scala, Python, and R. Lightning speed makes Spark too good to pass up, but understanding limitations and challenges in advance goes a long way toward easing actual production implementation. Spark: Big Data Cluster Computing in Production tells you everything you need to know, with real-world production insight and expert guidance, tips, and tricks.

Real-World Hadoop - Ted Dunning 2015-03-24

If you're a business team leader, CIO, business analyst, or developer interested in how Apache Hadoop and Apache HBase-related technologies can address problems involving large-scale data in cost-effective ways, this book is for you. Using real-world stories and situations, authors Ted Dunning and Ellen Friedman show Hadoop newcomers and seasoned users alike how NoSQL databases and Hadoop can solve a variety of business and research issues. You'll learn about early decisions and pre-planning that can make the process easier and more productive. If you're already using these technologies, you'll discover ways to gain the full range of benefits possible with Hadoop. While you don't need a deep technical background to get started, this book does provide expert guidance to help managers, architects, and practitioners succeed with their Hadoop projects. Examine a day in the life of big data: India's ambitious Aadhaar project Review tools in the Hadoop ecosystem such as Apache's Spark, Storm, and Drill to learn how they can help you Pick up a collection of technical and strategic tips that have helped others succeed with Hadoop Learn from several prototypical Hadoop use cases, based on how organizations have actually applied the technology Explore real-world stories that reveal how MapR customers combine use cases when putting Hadoop and NoSQL to work, including in production

PySpark Recipes - Raju Kumar Mishra 2017-12-09

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

Java/J2EE Job Interview Companion - Arulkumaran Kumaraswamipillai 2007

400+ Java/J2EE Interview questions with clear and concise answers for: job seekers (junior/senior developers, architects, team/technical leads), promotion seekers, pro-active learners and interviewers. Lulu top 100 best seller. Increase your earning potential by learning, applying and succeeding. Learn the fundamentals relating to Java/J2EE in an easy to understand questions and answers approach. Covers 400+ popular interview Q&A with lots of diagrams, examples, code snippets, cross referencing and comparisons. This is not only an interview guide but also a quick reference guide, a refresher material and a roadmap covering a

wide range of Java/J2EE related topics. More Java J2EE interview questions and answers & resume resources at http: //www.lulu.com/java-succes

Data Analytics Basics - Simplilearn 2020-12-14

Data analytics is increasingly becoming a key element in shaping a company's business strategy. Today, data influences every decision made by an organization, and this is driving the wide-scale adoption of data analytics, including machine learning technologies and artificial intelligence solutions. The heightened focus is propelling a surge in data analytics spending, reflected in various studies conducted by leading market research firms. The field of data analytics offers some amazing salaries and is not only the hottest IT job, but it is also one of the best-paying jobs in the world. This guide aims at providing the readers with everything they need to know about the data analytics field, basic terminologies, key concepts, real-life use cases, skills you must master in order to scale up your career, and training and certifications you might need to reach your dream job.

Top 200 Data Engineer Interview Questions and Answers - Knowledge Powerhouse 2017-03-19

Top 200 Data Engineer Interview Questions Big Data and Data Science are the most popular technology trends. There is a growing demand for Data Engineer job in technology companies. This book contains technical interview questions that an interviewer asks for Data Engineer position. Each question is accompanied with an answer so that you can prepare for job interview in short time. The book contains questions on Apache Hadoop, Hive, Spark, SQL and MySQL. It is a combination of our five other books. We have compiled this list after attending dozens of technical interviews in top-notch companies like- Airbnb, Netflix, Amazon etc.Often, these questions and concepts are used in our daily work. But these are most helpful when an Interviewer is trying to test your deep knowledge of Big Data topics like- Hadoop, Hive, Spark, SQL, MySQL etc. What are the Big Data topics covered in this book? We cover a wide variety of Big Data and Data Science topics in this book. Some of the topics are Apache Hadoop, Hive, Spark, SQL, MySql etc. How will this book help me? By reading this book, you do not have to spend time searching the Internet for Data Engineer interview questions. We have already compiled the list of the most popular and the latest Data Engineer Interview questions. Are there answers in this book? Yes, in this book each question is followed by an answer. So you can save time in interview preparation. What is the best way of reading this book? You have to first do a slow reading of all the questions in this book. Once you go through them in the first pass, mark the questions that you could not answer by yourself. Then, in second pass go through only the difficult questions. After going through this book 2-3 times, you will be well prepared to face a technical interview for a Data Engineer position. What is the level of questions in this book? This book contains questions that are good for a beginner Data engineer to a senior Data engineer. The difficulty level of question varies in the book from Fresher to a Seasoned professional. What are the sample questions in this book? What is the difference between ROLLBACK TO SAVEPOINT and RELEASE SAVEPOINT? How will you see the current user logged into MySQL connection? Can we create multiple tables in Hive for a data file? Can we use Hive for Online Transaction Processing (OLTP) systems? Can we use same name for a TABLE and VIEW in Hive? How can we get a random number between 1 and 100 in MySQL? How can you copy the structure of a table into another table without copying the data? How can you find 10 employees with Odd number as Employee ID? How does CONCAT function work in Hive? How will you change the data type of a column in Hive? How will you check if a file exists in HDFS? How will you check if a table exists in MySQL? How will you run Unix

commands from Hive? How will you search for a String in MySQL column? How will you see the structure of a table in MySQL? How will you select the storage level in Apache Spark? How will you synchronize the changes made to a file in Distributed Cache in Hadoop? If we set Replication factor 3 for a file, does it mean any computation will also take place 3 times? Is it safe to use ROWID to locate a record in Oracle SQL queries? What are different Persistence levels in Apache Spark? What are the common Transformations in Apache Spark?
http://www.knowledgepowerhouse.com

Build a Career in Data Science - Emily Robinson 2020-03-06

Summary You are going to need more than technical knowledge to succeed as a data scientist. Build a Career in Data Science teaches you what school leaves out, from how to land your first job to the lifecycle of a data science project, and even how to become a manager. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology What are the keys to a data scientist's long-term success? Blending your technical know-how with the right "soft skills" turns out to be a central ingredient of a rewarding career. About the book Build a Career in Data Science is your guide to landing your first data science job and developing into a valued senior employee. By following clear and simple instructions, you'll learn to craft an amazing resume and ace your interviews. In this demanding, rapidly changing field, it can be challenging to keep projects on track, adapt to company needs, and manage tricky stakeholders. You'll love the insights on how to handle expectations, deal with failures, and plan your career path in the stories from seasoned data scientists included in the book. What's inside Creating a portfolio of data science projects Assessing and negotiating an offer Leaving gracefully and moving up the ladder Interviews with professional data scientists About the reader For readers who want to begin or advance a data science career. About the author Emily Robinson is a data scientist at Warby Parker. Jacqueline Nolis is a data science consultant and mentor. Table of Contents: PART 1 - GETTING STARTED WITH DATA SCIENCE 1. What is data science? 2. Data science companies 3. Getting the skills 4. Building a portfolio PART 2 - FINDING YOUR DATA SCIENCE JOB 5. The search: Identifying the right job for you 6. The application: Résumés and cover letters 7. The interview: What to expect and how to handle it 8. The offer: Knowing what to accept PART 3 - SETTLING INTO DATA SCIENCE 9. The first months on the job 10. Making an effective analysis 11. Deploying a model into production 12. Working with stakeholders PART 4 - GROWING IN YOUR DATA SCIENCE ROLE 13. When your data science project fails 14. Joining the data science community 15. Leaving your job gracefully 16. Moving up the ladder

Probability for Statistics and Machine Learning - Anirban DasGupta 2011-05-17

This book provides a versatile and lucid treatment of classic as well as modern probability theory, while integrating them with core topics in statistical theory and also some key tools in machine learning. It is written in an extremely accessible style, with elaborate motivating discussions and numerous worked out examples and exercises. The book has 20 chapters on a wide range of topics, 423 worked out examples, and 808 exercises. It is unique in its unification of probability and statistics, its coverage and its superb exercise sets, detailed bibliography, and in its substantive treatment of many topics of current importance. This book can be used as a text for a year long graduate course in statistics, computer science, or mathematics, for self-study, and as an invaluable research reference on probabiliity and its applications. Particularly worth mentioning are the treatments of distribution theory, asymptotics, simulation and Markov Chain Monte Carlo, Markov chains and martingales, Gaussian processes, VC theory, probability metrics, large

deviations, bootstrap, the EM algorithm, confidence intervals, maximum likelihood and Bayes estimates, exponential families, kernels, and Hilbert spaces, and a self contained complete review of univariate probability.

High Performance Spark - Holden Karau 2017-05-25

Apache Spark is amazing when everything clicks. But if you haven't seen the performance improvements you expected, or still don't feel confident enough to use Spark in production, this practical book is for you. Authors Holden Karau and Rachel Warren demonstrate performance optimizations to help your Spark queries run faster and handle larger data sizes, while using fewer resources. Ideal for software engineers, data engineers, developers, and system administrators working with large-scale data applications, this book describes techniques that can reduce data infrastructure costs and developer hours. Not only will you gain a more comprehensive understanding of Spark, you'll also learn how to make it sing. With this book, you'll explore: How Spark SQL's new interfaces improve performance over SQL's RDD data structure The choice between data joins in Core Spark and Spark SQL Techniques for getting the most out of standard RDD transformations How to work around performance issues in Spark's key/value pair paradigm Writing high-performance Spark code without Scala or the JVM How to test for functionality and performance when applying suggested improvements Using Spark MLlib and Spark ML machine learning libraries Spark's Streaming components and external community packages

Apache Spark Implementation on IBM z/OS - Lydia Parziale 2016-08-13

The term big data refers to extremely large sets of data that are analyzed to reveal insights, such as patterns, trends, and associations. The algorithms that analyze this data to provide these insights must extract value from a wide range of data sources, including business data and live, streaming, social media data. However, the real value of these insights comes from their timeliness. Rapid delivery of insights enables anyone (not only data scientists) to make effective decisions, applying deep intelligence to every enterprise application. Apache Spark is an integrated analytics framework and runtime to accelerate and simplify algorithm development, depoyment, and realization of business insight from analytics. Apache Spark on IBM® z/OS® puts the open source engine, augmented with unique differentiated features, built specifically for data science, where big data resides. This IBM Redbooks® publication describes the installation and configuration of IBM z/OS Platform for Apache Spark for field teams and clients. Additionally, it includes examples of business analytics scenarios.

Fountainhead of Jihad - Vahid Brown 2013

Drawing upon a wealth of previously unresearched primary sources in many languages, the authors shed much new light on a group frequently described as the most lethal actor in the current Afghan insurgency, and shown here to have been for decades at the centre of a nexus of transnational Islamist militancy, fostering the development of jihadi organisations from Southeast Asia to East Africa. Addressing the abundant new evidence documenting the Haqqani network's pivotal role in the birth and evolution of the global jihadi movement, the book also represents a significant advance in our knowledge of the history of al-Qaeda, fundamentally altering the picture painted by the existing literature on the subject.

Angular 2 Interview Questions and Answers - Anil Singh 2017-08-16

This book has the collection of Angular 2 Interview Questions and Answers with TypeScript and basic of Angular 4. Angular 2 is a most popular framework for developing mobile as well as web applications. Angular 2 is so simpler, faster, modular and instrumented design and it is developed by Google and Misko Hevery is the father of Angular. You can learn complete knowledge of Angular 2, TypeScript and Angular 4 simpler and faster with examples.

This book has important questions and answers for beginner and expert level of developers and it's containing 115 questions and answers with TypeScript, Angular 4 and examples.

Advanced Analytics with Spark - Sandy Ryza 2015-04-02

In this practical book, four Cloudera data scientists present a set of self-contained patterns for performing large-scale data analysis with Spark. The authors bring Spark, statistical methods, and real-world data sets together to teach you how to approach analytics problems by example. You'll start with an introduction to Spark and its ecosystem, and then dive into patterns that apply common techniques—classification, collaborative filtering, and anomaly detection among others—to fields such as genomics, security, and finance. If you have an entry-level understanding of machine learning and statistics, and you program in Java, Python, or Scala, you'll find these patterns useful for working on your own data applications. Patterns include: Recommending music and the Audioscrobbler data set Predicting forest cover with decision trees Anomaly detection in network traffic with K-means clustering Understanding Wikipedia with Latent Semantic Analysis Analyzing co-occurrence networks with GraphX Geospatial and temporal data analysis on the New York City Taxi Trips data Estimating financial risk through Monte Carlo simulation Analyzing genomics data and the BDG project Analyzing neuroimaging data with PySpark and Thunder

Spark Cookbook - Rishi Yadav 2015-07-27

By introducing in-memory persistent storage, Apache Spark eliminates the need to store intermediate data in filesystems, thereby increasing processing speed by up to 100 times. This book will focus on how to analyze large and complex sets of data. Starting with installing and configuring Apache Spark with various cluster managers, you will cover setting up development environments. You will then cover various recipes to perform interactive queries using Spark SQL and real-time streaming with various sources such as Twitter Stream and Apache Kafka. You will then focus on machine learning, including supervised learning, unsupervised learning, and recommendation engine algorithms. After mastering graph processing using GraphX, you will cover various recipes for cluster optimization and troubleshooting.

Data Pipelines Pocket Reference - James Densmore 2021-02-10

Data pipelines are the foundation for success in data analytics. Moving data from numerous diverse sources and transforming it to provide context is the difference between having data and actually gaining value from it. This pocket reference defines data pipelines and explains how they work in today's modern data stack. You'll learn common considerations and key decision points when implementing pipelines, such as batch versus streaming data ingestion and build versus buy. This book addresses the most common decisions made by data professionals and discusses foundational concepts that apply to open source frameworks, commercial products, and homegrown solutions. You'll learn: What a data pipeline is and how it works How data is moved and processed on modern data infrastructure, including cloud platforms Common tools and products used by data engineers to build pipelines How pipelines support analytics and reporting needs Considerations for pipeline maintenance, testing, and alerting

Frank Kane's Taming Big Data with Apache Spark and Python - Frank Kane 2017-06-30

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.