99 Apache Spark Interview Questions For Professionals A Guide To Prepa

Leverage the power of Scala with different tools to build scalable, robust data science applications About This Book A complete guide for scalable data science solutions, from data ingestion to data visualization Deploy horizontally scalable data processing pipelines and take advantage of web frameworks to build engaging visualizations Build functional, type-safe routines to interact with relational and NoSQL databases with the help of tutorials and examples provided Who This Book Is For If you are a Scala developer or data scientist, or if you want to enter the field of data science, then this book will give you all the tools you need to implement data science solutions. What You Will Learn Transform and filter tabular data to extract features for machine learning Implement your own algorithms or take advantage of MLLib's extensive suite of models to build distributed machine learning pipelines Read, transform, and write data to both SQL and NoSQL databases in a functional manner Write robust routines to query web APIs Read data from web APIs such as the GitHub or Twitter API Use Scala to interact with MongoDB, which offers high performance and helps to store large data sets with uncertain query requirements Create Scala web applications that couple with JavaScript libraries such as D3 to create compelling interactive visualizations Deploy scalable parallel applications using Apache Spark, loading data from HDFS or Hive In Detail Scala is a multi-paradigm programming language (it supports both object-oriented and functional programming) and scripting language used to build applications for the JVM. Languages such as R, Python, Java, and so on are mostly used for data science. It is particularly good at analyzing large sets of data without any significant impact on performance and thus Scala is being adopted by many developers and data scientists. Data scientists might be aware that building applications that are truly scalable is hard. Scala, with its powerful functional libraries for interacting with databases and building scalable frameworks will give you the tools to construct robust data pipelines. This book will introduce you to the libraries for ingesting, storing, manipulating, processing, and visualizing data in Scala. Packed with real-world examples and interesting data sets, this book will teach you to ingest data from flat files and web APIs and store it in a SQL or NoSQL database. It will show you how to design scalable architectures to process and modelling your data, starting from simple concurrency constructs such as parallel collections and futures, through to actor systems and Apache Spark. As well as Scala's emphasis on functional structures and immutability, you will learn how to use the right parallel construct for the job at hand, minimizing development time without compromising scalability. Finally, you will learn how to build beautiful interactive visualizations using web frameworks. This book gives tutorials on some of the most common Scala libraries for data science, allowing you to quickly get up to speed with building data science and data engineering solutions. Style and approach A tutorial with complete examples, this book will give you the tools to start building useful data engineering and data science solutions straightaway

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

Design, build, and deploy your own machine learning applications by leveraging key Java machine learning librariesAbout This Book- Develop a sound strategy to solve predictive modelling problems using the most popular machine learning Java libraries- Explore a broad variety of data processing, machine learning, and natural language processing through diagrams, source code, and real-world applications- Packed with practical advice and tips to help you get to grips with applied machine learningWho This Book Is ForIf you want to learn how to use Java's machine learning libraries to gain insight from your data, this book is for you. It will get you up and running quickly and provide you with the skills you need to successfully create, customize, and deploy machine learning applications in real life. You should be familiar with Java programming and data mining concepts to make the most of this book, but no prior experience with data mining packages is necessary.What You Will Learn- Understand the basic steps of applied machine learning and how to differentiate among various machine learning approaches- Discover key Java machine learning libraries, what each library brings to the table, and what kind of problems each are able to solve- Learn how to implement classification, regression, and clustering- Develop a sustainable strategy for customer retention by predicting likely churn candidates- Build a scalable recommendation engine with Apache Mahout- Apply machine learning to fraud, anomaly, and outlier detection- Experiment with deep learning concepts, algorithms, and the toolbox for deep learning- Write your own activity recognition model for eHealth applications using mobile sensorsIn DetailAs the amount of data continues to grow at an almost incomprehensible rate, being able to understand and process data is becoming a key differentiator for competitive organizations. Machine learning applications are everywhere, from self-driving cars, spam detection, document search, and trading strategies, to speech recognition. This makes machine learning well-suited to the present-day era of Big Data and Data Science. The main challenge is how to transform data into actionable knowledge.Machine Learning in Java will provide you with the techniques and tools you need to quickly gain insight from complex data. You will start by learning how to apply machine learning methods to a variety of common tasks including classification, prediction, forecasting, market basket analysis, and clustering.Moving on, you will discover how to detect anomalies and fraud, and ways to perform activity recognition, image recognition, and text analysis. By the end of the book, you will explore related web resources and technologies that will help you take your learning to the next level.By applying the most effective machine learning methods to real-world problems, you will gain hands-on experience that will transform the way you think about data.Style and approachThis is a practical tutorial that uses hands-on examples to step through some real-world applications of machine learning. Without shying away from the technical details, you will explore machine learning with Java libraries using clear and practical examples. You will explore how to prepare data for analysis, choose a machine learning method, and measure the success of the process.

Describes the features and functions of Apache Hive, the data infrastructure for Hadoop.

The Startup Owner's Manual

Principles of Database Management

A Weekly Journal

Functional and Reactive Domain Modeling

SAS Certified Specialist Prep Guide

The Future of the Internet--And How to Stop It

Learn what it takes to succeed in the the most in-demand tech job Harvard Business Review calls it the sexiest tech job of the 21st century. Data scientists are in demand, and this unique book shows you exactly what employers want and the skill set that separates the quality data scientist from other talented IT professionals. Data science involves extracting, creating, and processing data to turn it into business value. With over 15 years of big data, predictive modeling, and business analytics experience, author Vincent Granville is no stranger to data science. In this one-of-a-kind guide, he provides insight into the essential data science skills, such as statistics and visualization

techniques, and covers everything from analytical recipes and data science tricks to common job interview questions, sample resumes, and source code. The applications are endless and varied: automatically detecting spam and plagiarism, optimizing bid prices in keyword advertising, identifying new molecules to fight cancer, assessing the risk of meteorite impact. Complete with case studies, this book is a must, whether you're looking to become a data scientist or to hire one. Explains the finer points of data science, the required skills, and how to acquire them, including analytical recipes, standard rules, source code, and a dictionary of terms Shows what companies are looking for and how the growing importance of big data has increased the demand for data scientists Features job interview questions, sample resumes, salary surveys, and examples of job ads Case studies explore how data science is used on Wall Street, in botnet detection, for online advertising, and in many other business-critical situations Developing Analytic Talent: Becoming a Data Scientist is essential reading for those aspiring to this hot career choice and for employers seeking the best candidates.

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.

Journalist Walls grew up with parents whose ideals and stubborn nonconformity were their curse and their salvation. Rex and Rose Mary and their four children lived like nomads, moving among Southwest desert towns, camping in the mountains. Rex was a charismatic, brilliant man who, when sober, captured his children's imagination, teaching them how to embrace life fearlessly. Rose Mary painted and wrote and couldn't stand the responsibility of providing for her family. When the money ran out, the Walls retreated to the dismal West Virginia mining town Rex had tried to escape. As the dysfunction escalated, the children had to fend for themselves, supporting one another as they found the resources and will to leave home. Yet Walls describes her parents with deep affection in this tale of unconditional love in a family that, despite its profound flaws, gave her the fiery determination to carve out a successful life. -- From publisher description.

Hadoop Admin: Apache Ambari interview Questions which include the 118 questions in total and it will prepare you for the Hadoop Administration. It is not necessary this all questions would be asked during the interview process. But HadoopExam tries to cover all possible concepts which needs to learn for knowing the Apache Ambari Hadoop Cluster management tool. These questions and answer would be helpful to understand the various components, operations, monitoring and administering the Hadoop cluster for sure. The benefit of Question and answer format is that, it would allow you to understand the thing in depth and you can get the better insight on the subject. This book was created by the Engineering team of HadoopExam which has in depth knowledge about the Hadoop Cluster Administration and Created HandsOn Hadoop Administration training. The team target is to make you learn the subject as in depth as possible with the minimum effort hence we have material in Question, Answers format, On-demand video trainings, E-Books, Projects and POC etc. We are delighted when learners come and give the feedback about our material and become repeat subscriber because they regularly get new material as well as updated material. Again all the best and please provide the feedback on the admin@hadoopexam.com or hadoopexam@gmail.com . Wherever possible we are trying to help you in your career.

Designing Data-Intensive Applications

Kafka Streams - Real-time Stream Processing
Spark in Action
Patterns for Learning from Data at Scale
Learn Amazon SageMaker
High Performance Spark

Today, software engineers need to know not only how to program effectively but also how to develop proper engineering practices to make their codebase sustainable and healthy. This book emphasizes this difference between programming and software engineering. How can software engineers manage a living codebase that evolves and responds to changing requirements and demands over the length of its life? Based on their experience at Google, software engineers Titus Winters and Hyrum Wright, along with technical writer Tom Manshreck, present a candid and insightful look at how some of the world's leading practitioners construct and maintain software. This book covers Google's unique engineering culture, processes, and tools and how these aspects contribute to the effectiveness of an engineering organization. You'll explore three fundamental principles that software organizations should keep in mind when designing, architecting, writing, and maintaining code: How time affects the sustainability of software and how to make your code resilient over time How scale affects the viability of software practices within an engineering organization What trade-offs a typical engineer needs to make when evaluating design and development decisions

Introductory, theory-practice balanced text teaching the fundamentals of databases to advanced undergraduates or graduate students in information systems or computer science.

Summary Functional and Reactive Domain Modeling teaches you how to think of the domain model in terms of pure functions and how to compose them to build larger abstractions. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Traditional distributed applications won't cut it in the reactive world of microservices, fast data, and sensor networks. To capture their dynamic relationships and dependencies, these systems require a different approach to domain modeling. A domain model composed of pure functions is a more natural way of representing a process in a reactive system, and it maps directly onto technologies and patterns like Akka, CQRS, and event sourcing. About the Book Functional and Reactive Domain Modeling teaches you consistent, repeatable techniques for building domain models in reactive systems. This book reviews the relevant concepts of FP and reactive architectures and then methodically introduces this new approach to domain modeling. As you read, you'll learn where and how to apply it, even if your systems aren't purely reactive or functional. An expert blend of theory and practice, this book presents strong examples you'll return to again and again as you apply these principles to your own projects. What's Inside Real-world libraries and frameworks Establish meaningful reliability guarantees Isolate domain logic from side effects Introduction to reactive design patterns About the Reader Readers should be comfortable with functional programming and traditional domain modeling. Examples use the Scala language. About the Author Software architect Debasish Ghosh was an early adopter of reactive design using Scala and Akka. He's the author of DSLs in Action, published by Manning in 2010. Table of Contents Functional domain modeling: an introduction Scala for functional domain models Designing functional domain models Functional patterns for domain models Modularization of domain models Being reactive Modeling with reactive streams Reactive persistence and event sourcing Testing your domain model Summary - core thoughts and principles

This Book is published by www.HadoopExam.com (HadoopExam Learning Resources). Where you can find material and training's for preparing for Big-data, Cloud Computing, Analytics, Data Science and popular Programming Language. This Book will contain 130+ frequent interview questions for Spark 2.0 framework, which also covers the YARN framework, Spark streaming, Core Spark and SparkSQL, PySpark, these questions will not only help you in clearing interview process, but also you can understand various underline concepts, which Spark engine uses internally. Also, it is recommended that you go through the Spark Hands On Training provided by HadoopExam. In training we have created concepts as well as practicals by creating simple and complex problems with the use of Spark framework API. While publishing this book there are 32 modules available, which are in-line with Spark technology to be used on Hadoop Framework.As you know, Spark is one the most popular computing framework used and very well integrate with the Hadoop framework. You can see previously professionals were using MapReduce framework as a computing engine, but since Spark developed it is almost replaced by Spark engine, because Spark can give you rich API as well as it do most of the time data processing by having data in memory. Having data in-memory can save lot of disk I/O and drastically improve the performanced of submitted application. If you see now a days IOT and Machine learning are catching up and most of the professional started using higher level API created using Spark framework like MLib, Graphx etc. Spark technology is now a days an exclusive skill, which most of developer want to learn. So to fulfill this need HadoopExam.com has many learning resources for learning Spark and doing certifications. Currently we have following products available to make you master in Apache Framework, visit HadoopExam.com for more detail. 1. Apache Spark Professional Training with Hands On Lab Sessions 2. Oreilly Databricks Apache Spark Developer Certification Simulator3. Hortonworks Spark Developer Certification 4. Cloudera CCA175 Hadoop and Spark Developer Certification 5. MapR Spark Certification preparation materialThis book has collection of questions, which are usually asked by the interviewer while filtering the candidates who had really worked on Spark framework which is well integrated with the Hadoop Framework.

Mastering Machine Learning on AWS

Mastering Apache Spark
Covers Apache Spark 3 with Examples in Java, Python, and Scala
Household Words
Unleashing Large Cluster Analytics in the Cloud
Learn English

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

Gain expertise in processing and storing data by using advanced techniques with Apache SparkAbout This Book- Explore the integration of Apache Spark with third party applications such as H20, Databricks and Titan- Evaluate how Cassandra and Hbase can be used for storage- An advanced guide with a combination of instructions and practical examples to extend the most up-to date Spark functionalitiesWho This Book Is ForIf you are a developer with some experience with Spark and want to strengthen your knowledge of how to get around in the world of Spark, then this book is ideal for you. Basic knowledge of Linux, Hadoop and Spark is assumed. Reasonable knowledge of Scala is expected.What You Will Learn- Extend the tools available for processing and storage- Examine clustering and classification using MLlib- Discover Spark stream processing via Flume, HDFS- Create a schema in Spark SQL, and learn how a Spark schema can be populated with data- Study Spark based graph processing using Spark GraphX- Combine Spark with H20 and deep learning and learn why it is useful- Evaluate how graph storage works with Apache Spark, Titan, HBase and Cassandra- Use Apache Spark in the cloud with Databricks and AWSIn DetailApache Spark is an in-memory cluster based parallel processing system that provides a wide range of functionality like graph processing, machine learning, stream processing and SQL. It operates at unprecedented speeds, is easy to use and offers a rich set of data transformations.This book aims to take your limited knowledge of Spark to the next level by teaching you how to expand Spark functionality. The book commences with an overview of the Spark eco-system. You will learn how to use MLlib to create a fully working neural net for handwriting recognition. You will then discover how stream processing can be tuned for optimal performance and to ensure parallel processing. The book extends to show how to incorporate H20 for machine learning, Titan for graph based storage, Databricks for cloud-based Spark. Intermediate Scala based code examples are provided for Apache Spark module processing in a CentOS Linux and Databricks cloud environment.Style and approachThis book is an extensive guide to Apache Spark modules and tools and shows how Spark's functionality can be extended for real-time processing and storage with worked examples.

This book aims to provide a broad overview of various topics of Internet of Things (IoT), ranging from research, innovation and development priorities to enabling technologies, nanoelectronics, cyber-physical systems, architecture, interoperability and industrial applications. All this is happening in a global context, building towards intelligent, interconnected decision making as an essential driver for new growth and co-competition across a wider set of markets. It is intended to be a standalone book in a series that covers the Internet of Things activities of the IERC – Internet of Things European Research Cluster from research to technological innovation, validation and deployment. The book builds on the ideas put forward by the European Research Cluster on the Internet of Things Strategic Research and Innovation Agenda, and presents global views and state of the art results on the challenges facing the research, innovation, development and deployment of IoT in future years. The concept of IoT could disrupt consumer and industrial product markets generating new revenues and serving as a growth driver for semiconductor, networking equipment, and service provider end-markets globally. This will create new application and product end-markets, change the value chain of companies that creates the IoT technology and deploy it in various end sectors, while impacting the business models of semiconductor, software, device, communication and service provider stakeholders. The proliferation of intelligent devices at the edge of the network with the introduction of embedded software and app-driven hardware into manufactured devices, and the ability, through

embedded software/hardware developments, to monetize those device functions and features by offering novel solutions, could generate completely new types of revenue streams. Intelligent and IoT devices leverage software, software licensing, entitlement management, and Internet connectivity in ways that address many of the societal challenges that we will face in the next decade.

Review "I really appreciate the quality and depth this book has. This book covers wide range of day to day conversation. It is very useful for everyone who has fear of speaking in English language. I have never seen such a great value at a highly reasonable price. Strongly recommend for people those are interested in speaking English in a short span of time." CA Shiv Singhal, Director, Devalya Education Private LimitedProduct Description You want to speak English. This is the book you need. You may find it difficult to learn and speak English in the absence of right guidance. This book provides a blueprint to practice every day conversation. It offers practical approach to learn English, and it also helps you to overcome grammatical blunders by covering following aspects:Part 1- How to describe your present Part 2 - How to describe a person or a place Part 3 - How to describe your past Part 4 - How to describe a conversation Part 5 - How to describe something step by step Part 6 - How to describe a social topic Part 7 - How to describe your imagination In my first book Speak English Like a Star, my purpose was to help you to understand conceptual English. This book is meant to take you to next level and it will help you to speak English confidently and comfortably in office or at home or with strangers.Answer following questions to know whether this book is right choice for you 1. Do you find it difficult to express your views? 2. Do you commit grammatical errors while writing English or speaking English? 3. Do you need a script for everyday English conversation? If answer to any of above questions is "yes", this book is one of the best resources to overcome your challenges. Buy this book if English speaking has become a barrier in your career and you want to speak English to take your career to next level. About Author Yogesh has been helping people to develop command of English, communication and career skills. He caters to job seekers, working professionals and corporate primarily. He is author of Speak English Like a Star and Unlock Your Confidence Overnight. He has also developed video training program on spoken English; namely, Learn English at Home. His articles on communication and career development have been read by more than 200000 people on different sites. He makes you speak English from day one and his passion is to help you to become better communicator and achieve your goals by applying best strategies.

Your Guide to Everyday Conversation

The Glass Castle

Developing Analytic Talent

Internet of Things Research and Innovation Value Chains, Ecosystems and Markets

Learning Spark

97 Things Every Cloud Engineer Should Know

With this practical book, AI and machine learning practitioners will learn how to successfully build and deploy data science projects on Amazon Web Services. The Amazon AI and machine learning stack unifies data science, data engineering, and application development to help level upyour skills. This guide shows you how to build and run pipelines in the cloud, then integrate the results into applications in minutes instead of days. Throughout the book, authors Chris Fregly and Antje Barth demonstrate how to reduce cost and improve performance. Apply the Amazon AI and ML stack to real-world use cases for natural language processing, computer vision, fraud detection, conversational devices, and more Use automated machine learning to implement a specific subset of use cases with SageMaker Autopilot Dive deep into the complete model development lifecycle for a BERT-based NLP use case including data ingestion, analysis, model training, and deployment Tie everything together into a repeatable machine learning operations pipeline Explore real-time ML, anomaly detection, and streaming analytics on data streams with Amazon Kinesis and Managed Streaming for Apache Kafka Learn security best practices for data science projects and workflows including identity and access management, authentication, authorization, and more

This extraordinary book explains the engine that has catapulted the Internet from backwater to ubiquity—and reveals that it is sputtering precisely because of its runaway success. With the unwitting help of its users, the generative Internet is on a path to a lockdown, ending its cycle of innovation—and facilitating unsettling new kinds of control. IPods, iPhones, Xboxes, and TiVos represent the first wave of Internet-centered products that can't be easily modified by anyone except their vendors or selected partners. These “tethered appliances” have already been used in remarkable but little-known ways: car GPS systems have been reconfigured at the demand of law enforcement to eavesdrop on the occupants at all times, and digital video recorders have been ordered to self-destruct thanks to a lawsuit against the manufacturer thousands of miles away. New Web 2.0 platforms like Google mash-ups and Facebook are rightly touted—but their applications can be similarly monitored and eliminated from a central source. As tethered appliances and applications eclipse the PC, the very nature of the Internet—its “generativity,” or innovative character—is at risk. The Internet's current trajectory is one of lost opportunity. Its salvation, Zittrain argues, lies in the hands of its millions of users. Drawing on generative technologies like Wikipedia that have so far survived their own successes, this book shows how to develop new technologies and social structures that allow users to work creatively and collaboratively, participate in solutions, and become true “netizens.”

The book contains the latest trend in IT industry 'BigData and Hadoop'. It explains how big is 'Big Data' and why everybody is trying to implement this into their IT project.It includes research work on various topics, theoretical and practical approach, each component of the architecture is described along with current industry trends.Big Data and Hadoop have taken together are a new skill as per the industry standards. Readers will get a compact book along with the industry experience and would be a reference to help readers.KEY FEATURES Overview Of Big Data, Basics of Hadoop, Hadoop Distributed File System, HBase, MapReduce, HIVE: The Dataware House Of Hadoop, PIG: The Higher Level Programming Environment, SQOOP: Importing Data From Heterogeneous Sources, Flume, Ozzie, Zookeeper & Big Data Stream Mining, Chapter-wise Questions & Previous Years Questions

Updated with new code, new projects, and new chapters, Machine Learning with TensorFlow, Second Edition gives readers a solid foundation in machine-learning concepts and the TensorFlow library. Summary Updated with new code, new projects, and new chapters, Machine Learning with TensorFlow, Second Edition gives readers a solid foundation in machine-learning concepts and the TensorFlow library. Written by NASA JPL Deputy CTO and Principal Data Scientist Chris Mattmann, all examples are accompanied by downloadable Jupyter Notebooks for a hands-on experience coding TensorFlow with Python. New and revised content expands coverage of core machine learning algorithms, and advancements in neural networks such as VGG-Face facial identification classifiers and deep speech classifiers. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Supercharge your data analysis with machine learning! ML algorithms automatically improve as they process data, so results get better over time. You don't have to be a mathematician to use ML: Tools like Google's TensorFlow library help with complex calculations so you can focus on getting the answers you need. About the book Machine Learning with TensorFlow, Second Edition is a fully revised guide to building machine learning models using Python and TensorFlow. You'll apply core ML concepts to real-world challenges, such as sentiment analysis, text classification, and image recognition. Hands-on examples illustrate neural network techniques for deep speech processing, facial identification, and auto-encoding with CIFAR-10. What's inside Machine Learning with TensorFlow Choosing the best ML approaches Visualizing algorithms with TensorBoard Sharing results with collaborators Running models in Docker About the reader Requires intermediate Python skills and knowledge of general algebraic concepts like vectors and matrices. Examples use the super-stable 1.15.x branch of TensorFlow and TensorFlow 2.x. About the author Chris Mattmann is the Division Manager of the Artificial Intelligence, Analytics, and Innovation Organization at NASA Jet Propulsion Lab. The first edition of this book was written by Nishant Shukla with Kenneth Fricklas. Table of Contents PART 1 - YOUR MACHINE-LEARNING RIG 1 A machine-learning odyssey 2 TensorFlow essentials PART 2 - CORE LEARNING ALGORITHMS 3 Linear regression and beyond 4 Using regression for call-center volume prediction 5 A gentle introduction to classification 6 Sentiment classification: Large movie-review dataset 7 Automatically clustering data 8 Inferring user activity from Android accelerometer data 9 Hidden Markov models 10 Part-of-speech tagging and word-sense disambiguation PART 3 - THE NEURAL NETWORK PARADIGM 11 A peek into autoencoders 12 Applying autoencoders: The CIFAR-10 image dataset 13 Reinforcement learning 14 Convolutional neural networks 15 Building a real-world CNN: VGG-Face ad VGG-Face Lite 16 Recurrent neural networks 17 LSTMs and automatic speech recognition 18 Sequence-to-sequence models for chatbots 19 Utility landscape

A guide to building, training, and deploying machine learning models for developers and data scientists

Best Practices for Scaling and Optimizing Apache Spark

Introduction to Information Retrieval

Beginning Apache Spark Using Azure Databricks

Total 130+ Questions

Scala for Data Science

If you create, manage, operate, or configure systems running in the cloud, you're a cloud engineer--even if you work as a system administrator, software developer, data scientist, or site reliability engineer. With this bo professionals from around the world provide valuable insight into today's cloud engineering role. These concise articles explore the entire cloud computing experience, including fundamentals, architecture, and migrati You'll delve into security and compliance, operations and reliability, and software development. And examine networking, organizational culture, and more. You're sure to find 1, 2, or 97 things that inspire you to dig deeper and expand your own career. "Three Keys to Making the Right Multicloud Decisions," Brendan O'Leary "Serverless Bad Practices," Manases Jesus Galindo Bello "Failing a Cloud Migration," Lee Atchison "Treat Your Cloud Environment as If It Were On Premises," Iyana Garry "What Is Toil, and Why Are SREs Obsessed with It?", Zachary Nickens "Lean QA: The QA Evolving in the DevOps World," Theresa Neate "How Economies of Scale Work in the Cloud," Jon Moore "The Cloud Is Not About the Cloud," Ken Corless "Data Gravity: The Importance of Data Management in the Cloud," Geoff Hughes "Even in the Cloud, the Network Is the Foundation," David Murray "Cloud Engineering Is About Culture, Not Containers," Holly Cummins

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 tog 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 filterin 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 fi 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 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

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 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 oth 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 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)

The AWS Certified Machine Learning Specialty 2020 Certification Guide covers everything you need to pass the MLS-C01 certification exam and serves as a handy, on-the-job reference guide. You'll also find the book useful if you're looking to get up to speed with AWS services for machine learning.

Building the Hyperconnected Society

Data Science on AWS
Disruptive Analytics
A Beginner's Guide to Storytelling with Data
Practical SQL
AWS Certified Machine Learning Specialty: MLS-C01 Certification Guide

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.

More than 100,000 entrepreneurs rely on this book for detailed, step-by-step instructions on building successful, scalable, profitable startups. The National Science Foundation pays hundreds of startup teams each year to follow the process outlined in the book, and it's taught at Stanford, Berkeley, Columbia and more than 100 other leading universities worldwide. Why? The Startup Owner's Manual guides you, step-by-step, as you put the Customer Development process to work. This method was created by renowned Silicon Valley startup expert Steve Blank, co-creator with Eric Ries of the "Lean Startup" movement and tested and refined by him for more than a decade. This 608-page how-to guide includes over 100 charts, graphs, and diagrams, plus 77 valuable checklists that guide you as you drive your company toward profitability. It will help you: • Avoid the 9 deadly sins that destroy startups' chances for success • Use the Customer Development method to bring your business idea to life • Incorporate the Business Model Canvas as the organizing principle for startup hypotheses • Identify your customers and determine how to "get, keep and grow" customers profitably • Compute how you'll drive your startup to repeatable, scalable profits. The Startup Owner's Manual was originally published by K&S Ranch Publishing Inc. and is now available from Wiley. The cover, design, and content are the same as the prior release and should not be considered a new or updated product.

Summary Spark in Action teaches you the theory and skills you need to effectively handle batch and streaming data using Spark. Fully updated for Spark 2.0. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Big data systems distribute datasets across clusters of machines, making it a challenge to efficiently query, stream, and interpret them. Spark can help. It is a processing system designed specifically for distributed data. It provides easy-to-use interfaces, along with the performance you need for production-quality analytics and machine learning. Spark 2 also adds improved programming APIs, better performance, and countless other upgrades. About the Book Spark in Action teaches you the theory and skills you need to effectively handle batch and streaming data using Spark. You'll get comfortable with the Spark CLI as you work through a few introductory examples. Then, you'll start programming Spark using its core APIs. Along the way, you'll work with structured data using Spark SQL, process near-real-time streaming data, apply machine learning algorithms, and munge graph data using Spark GraphX. For a zero-effort startup, you can download the preconfigured virtual machine ready for you to try the book's code. What's Inside Updated for Spark 2.0 Real-life case studies Spark DevOps with Docker Examples in Scala, and online in Java and Python About the Reader Written for experienced programmers with some background in big data or machine learning. About the Authors Petar Ze?evi? and Marko Bona?i are seasoned developers heavily involved in the Spark community. Table of Contents PART 1 - FIRST STEPS Introduction to Apache

Spark Spark fundamentals Writing Spark applications The Spark API in depth PART 2 - MEET THE SPARK FAMILY Sparkling queries with Spark SQL Ingesting data with Spark Streaming Getting smart with MLlib ML: classification and clustering Connecting the dots with GraphX PART 3 - SPARK OPS Running Spark Running on a Spark standalone cluster Running on YARN and Mesos PART 4 - BRINGING IT TOGETHER Case study: real-time dashboard Deep learning on Spark with H2O
Practical SQL is an approachable and fast-paced guide to SQL (Structured Query Language), the standard programming language for defining, organizing, and exploring data in relational databases. The book focuses on using SQL to find the story your data tells, with the popular open-source database PostgreSQL and the pgAdmin interface as its primary tools. You'll first cover the fundamentals of databases and the SQL language, then build skills by analyzing data from the U.S. Census and other federal and state government agencies. With exercises and real-world examples in each chapter, this book will teach even those who have never programmed before all the tools necessary to build powerful databases and access information quickly and efficiently. You'll learn how to: - Create databases and related tables using your own data - Define the right data types for your information - Aggregate, sort, and filter data to find patterns - Use basic math and advanced statistical functions - Identify errors in data and clean them up - Import and export data using delimited text files - Write queries for geographic information systems (GIS) - Create advanced queries and automate tasks Learning SQL doesn't have to be dry and complicated. Practical SQL delivers clear examples with an easy-to-follow approach to teach you the tools you need to build and manage your own databases. This book uses PostgreSQL, but the SQL syntax is applicable to many database applications, including Microsoft SQL Server and MySQL.
Learning PySpark
Base Programming Using SAS 9.4
Software Engineering at Google
Becoming a Data Scientist
Lessons Learned from Programming Over Time
Machine Learning with TensorFlow, Second Edition
Learn all you need to know about seven key innovations disrupting business analytics today. These innovations—the open source business model, cloud analytics, the Hadoop ecosystem, Spark and in-memory analytics, streaming analytics, Deep Learning, and self-service analytics—are radically changing how businesses use data for competitive advantage. Taken together, they are disrupting the business analytics value chain, creating new opportunities. Enterprises who seize the opportunity will thrive and prosper, while others struggle and decline: disrupt or be disrupted. Disruptive Business Analytics provides strategies to profit from disruption. It shows you how to organize for insight, build and provision an open source stack, how to practice lean data warehousing, and how to assimilate disruptive innovations into an organization. Through a short history of business analytics and a detailed survey of products and services, analytics authority Thomas W. Dinsmore provides a practical explanation of the most compelling innovations available today. What You'll Learn Discover how the open source business model works and how to make it work for you See how cloud computing completely changes the economics of analytics Harness the power of Hadoop and its ecosystem Find out why Apache Spark is everywhere Discover the potential of streaming and real-time analytics Learn what Deep Learning can do and why it matters See how self-service analytics can change the way organizations do business Who This Book Is For Corporate actors at all levels of responsibility for analytics: analysts, CIOs, CTOs, strategic decision makers, managers, systems architects, technical marketers, product developers, IT personnel, and consultants.
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
Class-tested and coherent, this textbook teaches classical and web information retrieval, including web search and the related areas of text classification and text clustering from basic concepts. It gives an up-to-date treatment of all aspects of the design and implementation of systems for gathering, indexing, and searching documents; methods for evaluating systems; and an introduction to the use of machine learning methods on text collections. All the important ideas are explained using examples and figures, making it perfect for introductory courses in information retrieval for advanced undergraduates and graduate students in computer science. Based on feedback from extensive classroom experience, the book has been carefully structured in order to make teaching more natural and effective. Slides and additional exercises (with solutions for lecturers) are also available through the book's supporting website to help course instructors prepare their lectures.
Summary The Spark distributed data processing platform provides an easy-to-implement tool for ingesting, streaming, and processing data from any source. In Spark in Action, Second Edition, you'll learn to take advantage of Spark's core features and incredible processing speed, with applications including real-time computation, delayed evaluation, and machine learning. Spark skills are a hot commodity in enterprises worldwide, and with Spark's powerful and flexible Java APIs, you can reap all the benefits without first learning Scala or Hadoop. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Analyzing enterprise data starts by reading, filtering, and merging files and streams from many sources. The Spark data processing engine handles this varied volume like a champ, delivering speeds 100 times faster than Hadoop systems. Thanks to SQL support, an intuitive interface, and a straightforward multilanguage API, you can use Spark without learning a complex new ecosystem. About the book Spark in Action, Second Edition, teaches you to create end-to-end analytics applications. In this entirely new book, you'll learn from interesting Java-

based examples, including a complete data pipeline for processing NASA satellite data. And you'll discover Java, Python, and Scala code samples hosted on GitHub that you can explore and adapt, plus appendixes that give you a cheat sheet for installing tools and understanding Spark-specific terms. What's inside Writing Spark applications in Java Spark application architecture Ingestion through files, databases, streaming, and Elasticsearch Querying distributed datasets with Spark SQL About the reader This book does not assume previous experience with Spark, Scala, or Hadoop. About the author Jean-Georges Perrin is an experienced data and software architect. He is France's first IBM Champion and has been honored for 12 consecutive years. Table of Contents PART 1 - THE THEORY CRIPPLED BY AWESOME EXAMPLES 1 So, what is Spark, anyway? 2 Architecture and flow 3 The majestic role of the dataframe 4 Fundamentally lazy 5 Building a simple app for deployment 6 Deploying your simple app PART 2 - INGESTION 7 Ingestion from files 8 Ingestion from databases 9 Advanced ingestion: finding data sources and building your own 10 Ingestion through structured streaming PART 3 - TRANSFORMING YOUR DATA 11 Working with SQL 12 Transforming your data 13 Transforming entire documents 14 Extending transformations with user-defined functions 15 Aggregating your data PART 4 - GOING FURTHER 16 Cache and checkpoint: Enhancing Spark's performances 17 Exporting data and building full data pipelines 18 Exploring deployment

A Memoir

Advanced Analytics with Spark

Practical Recommender Systems

Lightning-Fast Big Data Analysis

Hands-On Data Science and Python Machine Learning

Top 200 Data Engineer Interview Questions and Answers

The book Kafka Streams - Real-time Stream Processing helps you understand the stream processing in general and apply that skill to Kafka streams programming. This book is focusing mainly on the new generation of the Kafka Streams library available in the Apache Kafka 2.x. The primary focus of this book is on Kafka Streams. However, the book also touches on the other Apache Kafka capabilities and concepts that are necessary to grasp the Kafka Streams programming. Who should read this book? Kafka Streams: Real-time Stream Processing is written for software engineers willing to develop a stream processing application using Kafka Streams library. I am also writing this book for data architects and data engineers who are responsible for designing and building the organization's data-centric infrastructure. Another group of people is the managers and architects who do not directly work with Kafka implementation, but they work with the people who implement Kafka Streams at the ground level. What should you already know? This book assumes that the reader is familiar with the basics of Java programming language. The source code and examples in this book are using Java 8, and I will be using Java 8 lambda syntax, so experience with lambda will be helpful. Kafka Streams is a library that runs on Kafka. Having a good fundamental knowledge of Kafka is essential to get the most out of Kafka Streams. I will touch base on the mandatory Kafka concepts for those who are new to Kafka. The book also assumes that you have some familiarity and experience in running and working on the Linux operating system.

Summary Online recommender systems help users find movies, jobs, restaurants-even romance! There's an art in combining statistics, demographics, and query terms to achieve results that will delight them. Learn to build a recommender system the right way: it can make or break your application! Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Recommender systems are everywhere, helping you find everything from movies to jobs, restaurants to hospitals, even romance. Using behavioral and demographic data, these systems make predictions about what users will be most interested in at a particular time, resulting in high-quality, ordered, personalized suggestions. Recommender systems are practically a necessity for keeping your site content current, useful, and interesting to your visitors. About the Book Practical Recommender Systems explains how recommender systems work and shows how to create and apply them for your site. After covering the basics, you'll see how to collect user data and produce personalized recommendations. You'll learn how to use the most popular recommendation algorithms and see examples of them in action on sites like Amazon and Netflix. Finally, the book covers scaling problems and other issues you'll encounter as your site grows. What's inside How to collect and understand user behavior Collaborative and content-based filtering Machine learning algorithms Real-world examples in Python About the Reader Readers need intermediate programming and database skills. About the Author Kim Falk is an experienced data scientist who works daily with machine learning and recommender systems. Table of Contents PART 1 - GETTING READY FOR RECOMMENDER SYSTEMS What is a recommender? User behavior and how to collect it Monitoring the system Ratings and how to calculate them Non-personalized recommendations The user (and content) who came in from the cold PART 2 - RECOMMENDER ALGORITHMS Finding similarities among users and among content Collaborative filtering in the neighborhood Evaluating and testing your recommender Content-based filtering Finding hidden genres with matrix factorization Taking the best of all algorithms: implementing hybrid recommenders Ranking and learning to rank Future of recommender systems

Data is at the center of many challenges in system design today. Difficult issues need to be figured out, such as scalability, consistency, reliability, efficiency, and maintainability. In addition, we have an overwhelming variety of tools, including relational databases, NoSQL datastores, stream or batch processors, and message brokers. What are the right choices for your application? How do you make sense of all these buzzwords? In this

practical and comprehensive guide, author Martin Kleppmann helps you navigate this diverse landscape by examining the pros and cons of various technologies for processing and storing data. Software keeps changing, but the fundamental principles remain the same. With this book, software engineers and architects will learn how to apply those ideas in practice, and how to make full use of data in modern applications. Peer under the hood of the systems you already use, and learn how to use and operate them more effectively Make informed decisions by identifying the strengths and weaknesses of different tools Navigate the trade-offs around consistency, scalability, fault tolerance, and complexity Understand the distributed systems research upon which modern databases are built Peek behind the scenes of major online services, and learn from their architectures

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.

The Step-By-Step Guide for Building a Great Company

Machine Learning in Java

Programming Hive

Charting Your Strategy for Next-Generation Business Analytics

Learn by example

Hadoop Administration : Apache Ambari Interview Questions

Analyze vast amounts of data in record time using Apache Spark with Databricks in the Cloud. Learn the fundamentals, and more, of running analytics on large clusters in Azure and AWS, using Apache Spark with Databricks on top. Discover how to squeeze the most value out of your data at a mere fraction of what classical analytics solutions cost, while at the same time getting the results you need, incrementally faster. This book explains how the confluence of these pivotal technologies gives you enormous power, and cheaply, when it comes to huge datasets. You will begin by learning how cloud infrastructure makes it possible to scale your code to large amounts of processing units, without having to pay for the machinery in advance. From there you will learn how Apache Spark, an open source framework, can enable all those CPUs for data analytics use. Finally, you will see how services such as Databricks provide the power of Apache Spark, without you having to know anything about configuring hardware or software. By removing the need for expensive experts and hardware, your resources can instead be allocated to actually finding business value in the data. This book guides you through some advanced topics such as analytics in the cloud, data lakes, data ingestion, architecture, machine learning, and tools, including Apache Spark, Apache Hadoop, Apache Hive, Python, and SQL. Valuable exercises help reinforce what you have learned. What You Will Learn Discover the value of big data analytics that leverage the power of the cloud Get started with Databricks using SQL and Python in either Microsoft Azure or AWS Understand the underlying technology, and how the cloud and Apache Spark fit into the bigger picture See how these tools are used in the real world Run basic analytics, including machine learning, on billions of rows at a fraction of a cost or free Who This Book Is For Data engineers, data scientists, and cloud architects who want or need to run advanced analytics in the cloud. It is assumed that the reader has data experience, but perhaps minimal exposure to Apache Spark and Azure Databricks. The book is also recommended for people who want to get started in the analytics field, as it provides a strong foundation.

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

Spark 2. 0 Interview Questions

Big Data and Hadoop

The Big Ideas Behind Reliable, Scalable, and Maintainable Systems

Total 118 Questions: Quickly Become Hadoop Administrator

The definitive guide to passing the MLS-C01 exam on the very first attempt

The Practical Guide to Storing, Managing and Analyzing Big and Small Data