99 Apache Spark Interview Questions For Professionals A Guide To Prepare For Apache Spark Interview Questions

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.

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

Hadoop Administration : Apache Ambari Interview QuestionsTotal 118 Questions: Quickly Become Hadoop AdministratorHadoopExam Learning Resources

Discover the capabilities of PySpark and its application in the realm of data science. This comprehensive guide with hand-picked examples of daily use cases will walk you through the end-to-end predictive model-building cycle with the latest techniques and tricks of the trade. Applied Data Science Using PySpark is divided unto six sections which walk you through the book. In section 1, you start with the basics of PySpark focusing on data manipulation. We make you comfortable with the language and then build upon it to introduce you to the mathematical functions available off the shelf. In section 2, you will dive into the art of variable selection where we demonstrate various selection techniques available in PySpark. In section 3, we take you on a journey through machine learning algorithms, implementations, and fine-tuning techniques. We will also talk about different validation metrics and how to use them for picking the best models. Sections 4 and 5 go through machine learning pipelines and various methods available to operationalize the model and serve it through Docker/an API. In the final section, you will cover reusable objects for easy experimentation and learn some tricks that can help you optimize your programs and machine learning pipelines. By the end of this book, you will have seen the flexibility and advantages of PySpark in data science applications. This book is recommended to those who want to unleash the power of parallel computing by simultaneously working with big datasets. What You Will Learn Build an end-to-end predictive model Implement multiple variable selection techniques Operationalize models Master multiple algorithms and implementations Who This Book is For Data scientists and machine learning and deep learning engineers who want to learn and use PySpark for real-time analysis of streaming data.

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.

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.

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.

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

Upgrade your machine learning models with graph-based algorithms, the perfect structure for complex and interlinked data. Summary In Graph-Powered Machine Learning, you will learn: The lifecycle of a machine learning project Graphs in big data platforms Data source modeling using graphs Graph-based natural language processing, recommendations, and fraud detection techniques Graph algorithms Working with Neo4J Graph-Powered Machine Learning teaches to use graph-based algorithms and data organization strategies to develop superior machine learning applications. You’ll dive into the role of graphs in machine learning and big data platforms, and take an in-depth look at data source modeling, algorithm design, recommendations, and fraud detection. Explore end-to-end projects that illustrate architectures and help you optimize with best design practices. Author Alessandro Negro’s extensive experience shines through in every chapter, as you learn from examples and concrete scenarios based on his work with real clients! Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Identifying relationships is the foundation of machine learning. By recognizing and analyzing the connections in your data, graph-centric algorithms like K-nearest neighbor or PageRank radically improve the effectiveness of ML applications. Graph-based machine learning techniques offer a powerful new perspective for machine learning in social networking, fraud detection, natural language processing, and recommendation systems. About the book Graph-Powered Machine Learning teaches you how to exploit the natural relationships in structured and unstructured datasets using

graph-oriented machine learning algorithms and tools. In this authoritative book, you'll master the architectures and design practices of graphs, and avoid common pitfalls. Author Alessandro Negro explores examples from real-world applications that connect GraphML concepts to real world tasks. What's inside Graphs in big data platforms Recommendations, natural language processing, fraud detection Graph algorithms Working with the Neo4J graph database About the reader For readers comfortable with machine learning basics. About the author Alessandro Negro is Chief Scientist at GraphAware. He has been a speaker at many conferences, and holds a PhD in Computer Science. Table of Contents PART 1 INTRODUCTION 1 Machine learning and graphs: An introduction 2 Graph data engineering 3 Graphs in machine learning applications PART 2 RECOMMENDATIONS 4 Content-based recommendations 5 Collaborative filtering 6 Session-based recommendations 7 Context-aware and hybrid recommendations PART 3 FIGHTING FRAUD 8 Basic approaches to graph-powered fraud detection 9 Proximity-based algorithms 10 Social network analysis against fraud PART 4 TAMING TEXT WITH GRAPHS 11 Graph-based natural language processing 12 Knowledge graphs

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

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

If you 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 book, 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 migration. 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

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

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

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.

Dig deep into the data with a hands-on guide to machine learning with updated examples and more! Machine Learning: Hands-On for Developers and Technical Professionals provides hands-on instruction and fully-coded working examples for the most common machine learning techniques used by developers and technical professionals. The book contains a breakdown of each ML variant, explaining how it works and how it is used within certain industries, allowing readers to incorporate the presented techniques into their own work as they follow along. A core tenant of machine learning is a strong focus on data preparation, and a full exploration of the various types of learning algorithms illustrates how the proper tools can help any developer extract information and insights from existing data. The book includes a full complement of Instructor's Materials to facilitate use in the classroom, making this resource useful for students and as a professional reference. At its core, machine learning is a mathematical, algorithm-based technology that forms the basis of historical data mining and modern big data science. Scientific analysis of big data requires a working knowledge of machine learning, which forms predictions based on known properties learned from training data. Machine Learning is an accessible, comprehensive guide for the non-mathematician, providing clear guidance that allows readers to: Learn the languages of machine learning including Hadoop, Mahout, and Weka Understand decision trees, Bayesian networks, and artificial neural networks Implement Association Rule, Real Time, and Batch learning Develop a strategic plan for safe, effective, and efficient machine learning By learning to construct a system that can learn from data, readers can increase their utility across industries. Machine learning sits at the core of deep dive data analysis and visualization, which is increasingly in demand as companies discover the goldmine hiding in their existing data. For the tech professional involved in data science, Machine Learning: Hands-On for Developers and Technical Professionals provides the skills and techniques required to dig deeper.

Sharpen your coding skills by exploring established computer science problems! Classic Computer Science Problems in Java challenges you with time-tested scenarios and algorithms. Summary Sharpen your coding skills by exploring established computer science problems! Classic Computer Science Problems in Java challenges you with time-tested scenarios and algorithms. You'll work through a series of exercises based in computer science fundamentals that are designed to improve your software development abilities, improve your understanding of artificial intelligence, and even prepare you to ace an interview. As you work through examples in search, clustering, graphs, and more, you'll remember important things you've forgotten and discover classic solutions to your "new" problems! Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Whatever software development problem you're facing, odds are someone has already uncovered a solution. This book collects the most useful solutions devised, guiding you through a variety of challenges and tried-and-true problem-solving techniques. The principles and algorithms presented here are guaranteed to save you countless hours in project after project. About the book Classic Computer Science Problems in Java is a master class in computer programming designed around 55 exercises that have been used in computer science classrooms for years. You'll work through hands-on examples as you explore core algorithms, constraint problems, AI applications, and much more. What's inside Recursion, memoization, and bit manipulation Search, graph, and genetic algorithms Constraint-satisfaction problems K-means clustering, neural networks, and adversarial search About the reader For intermediate Java programmers. About the author David Kopec is an assistant professor of Computer Science and Innovation at Champlain College in Burlington, Vermont. Table of Contents 1 Small problems 2 Search problems 3 Constraint-satisfaction problems 4 Graph problems 5 Genetic algorithms 6 K-means clustering 7 Fairly simple neural networks 8 Adversarial search 9 Miscellaneous problems 10 Interview with Brian Goetz

Get up and running, and productive in no time with MLflow using the most effective machine learning engineering approach Key Features Explore machine learning workflows for stating ML problems in a concise and clear manner using

MLflow Use MLflow to iteratively develop a ML model and manage it Discover and work with the features available in MLflow to seamlessly take a model from the development phase to a production environment Book Description MLflow is a platform for the machine learning life cycle that enables structured development and iteration of machine learning models and a seamless transition into scalable production environments. This book will take you through the different features of MLflow and how you can implement them in your ML project. You will begin by framing an ML problem and then transform your solution with MLflow, adding a workbench environment, training infrastructure, data management, model management, experimentation, and state-of-the-art ML deployment techniques on the cloud and premises. The book also explores techniques to scale up your workflow as well as performance monitoring techniques. As you progress, you'll discover how to create an operational dashboard to manage machine learning systems. Later, you will learn how you can use MLflow in the AutoML, anomaly detection, and deep learning context with the help of use cases. In addition to this, you will understand how to use machine learning platforms for local development as well as for cloud and managed environments. This book will also show you how to use MLflow in non-Python-based languages such as R and Java, along with covering approaches to extend MLflow with Plugins. By the end of this machine learning book, you will be able to produce and deploy reliable machine learning algorithms using MLflow in multiple environments. What you will learn Develop your machine learning project locally with MLflow's different features Set up a centralized MLflow tracking server to manage multiple MLflow experiments Create a model life cycle with MLflow by creating custom models Use feature streams to log model results with MLflow Develop the complete training pipeline infrastructure using MLflow features Set up an inference-based API pipeline and batch pipeline in MLflow Scale large volumes of data by integrating MLflow with high-performance big data libraries Who this book is for This book is for data scientists, machine learning engineers, and data engineers who want to gain hands-on machine learning engineering experience and learn how they can manage an end-to-end machine learning life cycle with the help of MLflow. Intermediate-level knowledge of the Python programming language is expected.

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

Get started with Apache Flink, the open source framework that powers some of the world’s largest stream processing applications. With this practical book, you’ll explore the fundamental concepts of parallel stream processing and discover how this technology differs from traditional batch data processing. Longtime Apache Flink committers Fabian Hueske and Vasia Kalavri show you how to implement scalable streaming applications with Flink’s DataStream API and continuously run and maintain these applications in operational environments. Stream processing is ideal for many use cases, including low-latency ETL, streaming analytics, and real-time dashboards as well as fraud detection, anomaly detection, and alerting. You can process continuous data of any kind, including user interactions, financial transactions, and IoT data, as soon as you generate them. Learn concepts and challenges of distributed stateful stream processing Explore Flink’s system architecture, including its event-time processing mode and fault-tolerance model Understand the fundamentals and building blocks of the DataStream API, including its time-based and statefuloperators Read data from and write data to external systems with exactly-once consistency Deploy and configure Flink clusters Operate continuously running streaming applications

This Book is published by www.HadoopExam.com (HadoopExam Learning Resources). Where you can find material and training's for preparing for BigData, Cloud Computing, Analytics, Data Science and popular Programming Language. This Book will contain how to setup 4 node cluster using VMWare workstation on your windows machine (similar you can try on MacBook) as well. There are in total 15 chapters and we have also give 6 problem scenarios for practice. However, you can get more than 50 practice scenarios from www.HadoopExam.com for preparing CCA131 certification exam. www.HadoopExam.com currently have in total 44 (Few more will be added soon) solved problem scenarios which you can get directly from website. This book not only provides how to prepare for CCA131 exam, but also gives you the platform detail to practice the material as well as how to setup the same. Currently we are providing or in process of Developing following material for Hadoop Big Data Certification. Please visit website for more detail.

Have you ever... - Wanted to work at an exciting futuristic company? - Struggled with an interview problem that could have been solved in 15 minutes? - Wished you could study real-world computing problems? If so, you need to read Elements of Programming Interviews (EPI). EPI is your comprehensive guide to interviewing for software development roles. The core of EPI is a collection of over 250 problems with detailed solutions. The problems are representative of interview questions asked at leading software companies. The problems are illustrated with 200 figures, 300 tested programs, and 150 additional variants. The book begins with a summary of the nontechnical aspects of interviewing, such as strategies for a great interview, common mistakes, perspectives

from the other side of the table, tips on negotiating the best offer, and a guide to the best ways to use EPI. We also provide a summary of data structures, algorithms, and problem solving patterns. Coding problems are presented through a series of chapters on basic and advanced data structures, searching, sorting, algorithm design principles, and concurrency. Each chapter stars with a brief introduction, a case study, top tips, and a review of the most important library methods. This is followed by a broad and thought-provoking set of problems. A practical, fun approach to computer science fundamentals, as seen through the lens of common programming interview questions. Jeff Atwood/Co-founder, Stack Overflow and Discourse

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)

A step-by-step solution-based guide to preparing building, training, and deploying high-quality machine learning models with Amazon SageMaker Key Features Perform ML experiments with built-in and custom algorithms in SageMaker Explore proven solutions when working with TensorFlow, PyTorch, Hugging Face Transformers, and scikit-learn Use the different features and capabilities of SageMaker to automate relevant ML processes Book Description Amazon SageMaker is a fully managed machine learning (ML) service that helps data scientists and ML practitioners manage ML experiments. In this book, you'll use the different capabilities and features of Amazon SageMaker to solve relevant data science and ML problems. This step-by-step guide features 80 proven recipes designed to give you the hands-on machine learning experience needed to contribute to real-world experiments and projects. You'll cover the algorithms and techniques that are commonly used when training and deploying NLP, time series forecasting, and computer vision models to solve ML problems. You'll explore various solutions for working with deep learning libraries and frameworks such as TensorFlow, PyTorch, and Hugging Face Transformers in Amazon SageMaker. You'll also learn how to use SageMaker Clarify, SageMaker Model Monitor, SageMaker Debugger, and SageMaker Experiments to debug, manage, and monitor multiple ML experiments and deployments. Moreover, you'll have a better understanding of how SageMaker Feature Store, Autopilot, and Pipelines can meet the specific needs of data science teams. By the end of this book, you'll be able to combine the different solutions you've learned as building blocks to solve real-world ML problems. What you will learn Train and deploy NLP, time series forecasting, and computer vision models to solve different business problems Push the limits of customization in SageMaker using custom container images Use AutoML capabilities with SageMaker Autopilot to create high-quality models Work with effective data analysis and preparation techniques Explore solutions for debugging and managing ML experiments and deployments Deal with bias detection and ML explainability requirements using SageMaker Clarify Automate intermediate and complex deployments and workflows using a variety of solutions Who this book is for This book is for developers, data scientists, and machine learning practitioners interested in using Amazon SageMaker to build, analyze, and deploy machine learning models with 80 step-by-step recipes. All you need is an AWS account to get things running. Prior knowledge of AWS, machine learning, and the Python programming language will help you to grasp the concepts covered in this book more effectively.

Data Science and Big Data Analytics is about harnessing the power of data for new insights. The book covers the breadth of activities and methods and tools that Data Scientists use. The content focuses on concepts, principles and practical applications that are applicable to any industry and technology environment, and the learning is supported and explained with examples that you can replicate using open-source software. This book will help you: Become a contributor on a data science team Deploy a structured lifecycle approach to data analytics problems Apply appropriate analytic techniques and tools to analyzing big data Learn how to tell a compelling story with data to drive business action Prepare for EMC Proven Professional Data Science Certification Corresponding data sets are available from the book’s page at Wiley which you can find on the Wiley site by searching for the ISBN 9781118876138. Get started discovering, analyzing, visualizing, and presenting data in a meaningful way today!

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

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.

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.

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.

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.

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.”

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

Get Started Fast with Apache Hadoop® 2, YARN, and Today’s Hadoop Ecosystem With Hadoop 2.x and YARN, Hadoop moves beyond MapReduce to become practical for virtually any type of data processing. Hadoop 2.x and the Data Lake concept represent a radical shift away from conventional approaches to data usage and storage. Hadoop 2.x installations offer unmatched scalability and breakthrough extensibility that supports new and existing Big Data analytics processing methods and models. Hadoop® 2 Quick-Start Guide is the first easy, accessible guide to Apache Hadoop 2.x, YARN, and the modern Hadoop ecosystem. Building on his unsurpassed experience teaching Hadoop and Big Data, author Douglas Eadline covers all the basics you need to know to install and use Hadoop 2 on personal computers or servers, and to navigate the powerful technologies that complement it. Eadline concisely introduces and explains every key Hadoop 2 concept, tool, and service, illustrating each with a simple “beginning-to-end” example and identifying trustworthy, up-to-date resources for learning more. This guide is ideal if you want to learn about Hadoop 2 without getting mired in technical details. Douglas Eadline will bring you up to speed quickly, whether you’re a user, admin, devops specialist, programmer, architect, analyst, or data scientist. Coverage Includes Understanding what Hadoop 2 and YARN do, and how they improve on Hadoop 1 with MapReduce Understanding Hadoop-based Data Lakes versus RDBMS Data Warehouses Installing Hadoop 2 and core services on Linux machines, virtualized sandboxes, or clusters Exploring the Hadoop Distributed File System (HDFS) Understanding the essentials of MapReduce and YARN application programming Simplifying programming and data movement with Apache Pig, Hive, Sqoop, Flume, Oozie, and HBase Observing application progress, controlling jobs, and managing workflows Managing Hadoop efficiently with Apache Ambari–including recipes for HDFS to NFSv3 gateway, HDFS snapshots, and YARN configuration Learning basic Hadoop 2 troubleshooting, and installing Apache Hue and Apache Spark

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.

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

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.

Real-world practice problems to bring your SQL skills to the next level It's easy to find basic SQL syntax and keyword information online. What's hard to find is challenging, well-designed, real-world problems--the type of problems that come up all the time when you're dealing with data. Learning how to solve these problems will give you the skill and confidence to step up in your career. With SQL Practice Problems, you can get that level of experience by solving sets of targeted problems. These aren't just problems designed to give an example of specific syntax, or keyword. These are the common problems you run into all the time when you deal with data. You will get real world practice, with real world data. I'll teach you how to "think" in SQL, how to analyze data problems, figure out the fundamentals, and work towards a solution that you can be proud of. It contains challenging problems, that hone your ability to write high quality SQL code. What do you get when you buy SQL Practice Problems? You get instructions on how set up MS SQL Server Express Edition 2016 and SQL Server Management Studio 2016, both free downloads. Almost all the SQL presented here works for previous versions of MS SQLServer, and any exceptions are highlighted. You'll also get a customized sample database, with video walk-through instructions on how to set it up on your computer. And of course, you get the actual practice problems - 57 problems that you work through step-by-step. There are targeted hints if you need them that help guide you through the question. For the more complex questions there are multiple levels of hints. Each answer comes with a short, targeted discussion section with alternative answers and tips on usage and good programming practice. What kind of problems are there in SQL Practice Problems? SQL Practice Problems has data analysis and reporting oriented challenges that are designed to step you through introductory, intermediate and advanced SQL Select statements, with a learn-by-doing technique. Most textbooks and courses have some practice problems. But most often, they're used just to illustrate a particular piece of syntax, with no filtering on what's most useful. What you'll get with SQL Practice Problems is the problems that illustrate some the most common challenges you'll run into with data, and the best, most useful techniques to solve them. These practice problems involve only Select statements, used for data analysis and reporting, and not statements to modify data (insert, delete, update), or to create stored procedures. About the author: Hi, my name is Sylvia Moestl Vasilik. I've been a database programmer and engineer for more than 15 years, working at top organizations like Expedia, Microsoft, T-Mobile, and the Gates Foundation. In 2015, I was teaching a SQL Server Certificate course at the University of Washington Continuing Education. It was a 10 week course, and my students paid more than $1000 for it. My students learned the basics of SQL, most of the keywords, and worked through practice problems every week of the course. But because of the emphasis on getting a broad overview of all features of SQL, we didn't spend enough time on the types of SQL that's used 95% of the time--intermediate and advanced Select statements. After the course was over, some of my students emailed me to ask where they could get more practice. That's when I was inspired to start work on this book.

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

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.

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

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.