Pyspark Coding Interview Questions

Data Analysis with Python and PySpark
Spark for Python Developers
Cassandra: The Definitive Guide
Scala for Machine Learning
PySpark Recipes
R for Marketing Research and Analytics
Introducing Data Science
A Guide to the Project Management Body of Knowledge (PMBOK® Guide) – Seventh Edition and The Standard for Project Management (ENGLISH)
Building Machine Learning Pipelines
Advanced Analytics with Spark
Powerful Python
Spark: The Definitive Guide
Cracking the Data Science Interview
Parallel and Concurrent Programming in Haskell
Learning PySpark
Learning Spark
Data Engineering with Google Cloud Platform
Data Science in Production
High Performance Spark
Deep Learning for Coders with fastai and PyTorch
Java/J2EE Job Interview Companion
Koalas / Koalas
Learning Spark
Top 50 Apache Spark Interview Questions and Answers
Kafka Streams - Real-time Stream Processing
Beyond the Basic Stuff with Python
PySpark Cookbook
Frank Kane's Taming Big Data with Apache Spark and Python
Hadoop: The Definitive Guide
Computer Graphics from Scratch
Mastering Social Media Mining with Python
Build a Career in Data Science
Mastering Spark with R
Artificial Intelligence with Python
Python for Geeks
Rapid GUI Programming with Python and Qt
Scala and Spark for Big Data Analytics
Database Reliability Engineering

SAVANAH JADON

Data Analysis with Python and PySpark Packt Publishing Ltd
A concise guide to implementing Spark Big Data analytics for Python developers, and building a real-time and insightful trend tracker data intensive appAbout This Book• Set up real-time streaming and batch data intensive infrastructure using Spark and Python• Deliver insightful visualizations in a web app using Spark (PySpark)• Inject live data using Spark Streaming with real-time eventsWho This Book Is ForThis book is for data scientists and software developers with a focus on Python who want to work with the Spark engine, and it will also benefit Enterprise Architects. All you need to have is a good background of Python and an inclination to work with Spark.What You Will Learn• Create a Python development environment powered by Spark (PySpark), Blaze, and Bookeh• Build a real-time trend tracker data intensive app• Visualize the trends and insights gained from data using Bookeh• Generate insights from data using machine learning through Spark MLLIB• Juggle with data using Blaze• Create training data sets and train the Machine Learning models• Test the machine learning models on test datasets• Deploy the machine learning algorithms and models and scale it for real-time eventsIn DetailLooking for a cluster computing system that provides high-level APIs? Apache Spark is your answer—an open source, fast, and general purpose cluster computing system. Spark's multi-stage memory primitives provide performance up to 100 times faster than Hadoop, and it is also well-suited for machine learning algorithms.Are you a Python developer inclined to work with Spark engine? If so, this book will be your companion as you create data-intensive app using Spark as a processing engine, Python visualization libraries, and web frameworks such as Flask.To begin with, you will learn the most effective way to install the Python development environment powered by Spark, Blaze, and Bookeh. You will then find out how to connect with data stores such as MySQL, MongoDB, Cassandra, and Hadoop.You'll expand your skills throughout, getting familiarized with the various data sources (Github, Twitter, Meetup, and Blogs), their data structures, and solutions to effectively tackle complexities. You'll explore datasets using iPython Notebook and will discover how to optimize the data models and pipeline. Finally, you'll get to know how to create training datasets and train the machine learning models.By the end of the book, you will have created a real-time and insightful trend tracker data-intensive app with Spark.Style and approach This is a comprehensive guide packed with easy-to-follow examples that will take your skills to the next level and will get you up and running with Spark.
Spark for Python Developers "O'Reilly Media, Inc."
Imagine what you could do if scalability wasn't a problem. With this hands-on guide, you'll learn how the Cassandra database management system handles hundreds of terabytes of data while remaining highly available across multiple data centers. This expanded second edition—updated for Cassandra 3.0—provides the technical details and practical examples you need to put this database to work in a production environment. Authors Jeff Carpenter and Eben Hewitt demonstrate the advantages of Cassandra's non-relational design, with special attention to data modeling. If you're a developer, DBA, or application architect looking to solve a database scaling issue or future-proof your application, this guide helps you harness Cassandra's speed and flexibility. Understand Cassandra's distributed and decentralized structure Use the Cassandra Query Language (CQL) and cqlsh—the CQL shell Create a working data model and compare it with an equivalent relational model Develop sample applications using client drivers for languages including Java, Python, and Node.js Explore cluster topology and learn how nodes exchange data Maintain a high level of performance in your cluster Deploy Cassandra on site, in the Cloud, or with Docker Integrate Cassandra with Spark, Hadoop, Elasticsearch, Solr, and Lucene
Cassandra: The Definitive Guide Simon and Schuster
Harness the power of Scala to program Spark and analyze tonnes of data in the blink of an eye! About This Book Learn Scala's sophisticated type system that combines Functional Programming and object-oriented concepts Work on a wide array of applications, from simple batch jobs to stream processing and machine learning Explore the most common as well as some complex use-cases to perform large-scale data analysis with Spark Who This Book Is For Anyone who wishes to learn how to perform data analysis by harnessing the power of Spark will find this book extremely useful. No knowledge of Spark or Scala is assumed, although prior programming experience (especially with other JVM languages) will be useful to pick up concepts quicker. What You Will Learn Understand object-oriented & functional programming concepts of Scala In-depth understanding of Scala collection APIs Work with RDD and DataFrame to learn Spark's core abstractions Analysing structured and unstructured data using SparkSQL and GraphX Scalable and fault-tolerant streaming application development using Spark structured streaming Learn machine-learning best practices for classification, regression, dimensionality reduction, and recommendation system to build predictive models with widely used algorithms in Spark MLlib & ML Build clustering models to cluster a vast amount of data Understand tuning, debugging, and monitoring Spark applications Deploy Spark applications on real clusters in Standalone, Mesos, and YARN In Detail Scala has been observing wide adoption over the past few years, especially in the field of data science and analytics. Spark, built on Scala, has gained a lot of recognition and is being used widely in productions. Thus, if you want to leverage the power of Scala and Spark to make sense of big data, this book is for you. The first part introduces you to Scala, helping you understand the object-oriented and functional programming concepts needed for Spark application development. It then moves on to Spark to cover the basic abstractions using RDD and DataFrame. This will help you develop scalable and fault-tolerant streaming applications by analyzing structured and unstructured data using SparkSQL, GraphX, and Spark structured streaming. Finally, the book moves on to some advanced topics, such as monitoring, configuration, debugging, testing, and deployment. You will also learn how to develop Spark applications using SparkR and PySpark APIs, interactive data analytics using Zeppelin, and in-memory data processing with Alluxio. By the end of this book, you will have a thorough understanding of Spark, and you will be able to perform full-stack data analytics with a feel that no amount of data is too big. Style and approach Filled with practical examples and use cases, this book will hot only help you get up and running with Spark, but will also take you farther down the

road to becoming a data scientist.
Scala for Machine Learning No Starch Press
Summary You are going to need more than technical knowledge to succeed as a data scientist. Build a Career in Data Science teaches you what school leaves out, from how to land your first job to the lifecycle of a data science project, and even how to become a manager. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology What are the keys to a data scientist’s long-term success? Blending your technical know-how with the right “soft skills” turns out to be a central ingredient of a rewarding career. About the book Build a Career in Data Science is your guide to landing your first data science job and developing into a valued senior employee. By following clear and simple instructions, you’ll learn to craft an amazing resume and ace your interviews. In this demanding, rapidly changing field, it can be challenging to keep projects on track, adapt to company needs, and manage tricky stakeholders. You’ll love the insights on how to handle expectations, deal with failures, and plan your career path in the stories from seasoned data scientists included in the book. What's inside Creating a portfolio of data science projects Assessing and negotiating an offer Leaving gracefully and moving up the ladder Interviews with professional data scientists About the reader For readers who want to begin or advance a data science career. About the author Emily Robinson is a data scientist at Warby Parker. Jacqueline Nolis is a data science consultant and mentor. Table of Contents: PART 1 - GETTING STARTED WITH DATA SCIENCE 1. What is data science? 2. Data science companies 3. Getting the skills 4. Building a portfolio PART 2 - FINDING YOUR DATA SCIENCE JOB 5. The search: Identifying the right job for you 6. The application: Résumés and cover letters 7. The interview: What to expect and how to handle it 8. The offer: Knowing what to accept PART 3 - SETTLING INTO DATA SCIENCE 9. The first months on the job 10. Making an effective analysis 11. Deploying a model into production 12. Working with stakeholders PART 4 - GROWING IN YOUR DATA SCIENCE ROLE 13. When your data science project fails 14. Joining the data science community 15. Leaving your job gracefully 16. Moving up the ladder
PySpark Recipes Packt Publishing Ltd
Learn about Koalas, where they live and why it is almost always in a tree.
R for Marketing Research and Analytics Packt Publishing Ltd
BRIDGE THE GAP BETWEEN NOVICE AND PROFESSIONAL You've completed a basic Python programming tutorial or finished Al Sweigart's bestseller, Automate the Boring Stuff with Python. What's the next step toward becoming a capable, confident software developer? Welcome to Beyond the Basic Stuff with Python. More than a mere collection of advanced syntax and masterful tips for writing clean code, you'll learn how to advance your Python programming skills by using the command line and other professional tools like code formatters, type checkers, linters, and version control. Sweigart takes you through best practices for setting up your development environment, naming variables, and improving readability, then tackles documentation, organization and performance measurement, as well as object-oriented design and the Big-O algorithm analysis commonly used in coding interviews. The skills you learn will boost your ability to program--not just in Python but in any language. You'll learn: Coding style, and how to use Python's Black auto-formatting tool for cleaner code Common sources of bugs, and how to detect them with static analyzers How to structure the files in your code projects with the Cookiecutter template tool Functional programming techniques like lambda and higher-order functions How to profile the speed of your code with Python's built-in timeit and cProfile modules The computer science behind Big-O algorithm analysis How to make your comments and docstrings informative, and how often to write them How to create classes in object-oriented programming, and why they're used to organize code Toward the end of the book you'll read a detailed source-code breakdown of two classic command-line games, the Tower of Hanoi (a logic puzzle) and Four-in-a-Row (a two-player tile-dropping game), and a breakdown of how their code follows the book's best practices. You'll test your skills by implementing the program yourself. Of course, no single book can make you a professional software developer. But Beyond the Basic Stuff with Python will get you further down that path and make you a better programmer, as you learn to write readable code that's easy to debug and perfectly Pythonic Requirements: Covers Python 3.6 and higher
Introducing Data Science Packt Publishing Ltd
Quickly find solutions to common programming problems encountered while processing big data. Content is presented in the popular problem-solution format. Look up the programming problem that you want to solve. Read the solution. Apply the solution directly in your own code. Problem solved! PySpark Recipes covers Hadoop and its shortcomings. The architecture of Spark, PySpark, and RDD are presented. You will learn to apply RDD to solve day-to-day big data problems. Python and NumPy are included and make it easy for new learners of PySpark to understand and adopt the model. What You Will Learn Understand the advanced features of PySpark2 and SparkSQL Optimize your code Program SparkSQL with Python Use Spark Streaming and Spark MLlib with Python Perform graph analysis with GraphFrames Who This Book Is For Data analysts, Python programmers, big data enthusiasts
A Guide to the Project Management Body of Knowledge (PMBOK® Guide) – Seventh Edition and The Standard for Project Management (ENGLISH) "O'Reilly Media, Inc."
This book is a complete introduction to the power of R for marketing research practitioners. The text describes statistical models from a conceptual point of view with a minimal amount of mathematics, presuming only an introductory knowledge of statistics. Hands-on chapters accelerate the learning curve by asking readers to interact with R from the beginning. Core topics include the R language, basic statistics, linear modeling, and data visualization, which is presented throughout as an integral part of analysis. Later chapters cover more advanced topics yet are intended to be approachable for all analysts. These sections examine logistic regression, customer segmentation, hierarchical linear modeling, market basket analysis, structural equation modeling, and conjoint analysis in R. The text uniquely presents Bayesian models with a minimally complex approach, demonstrating and explaining Bayesian methods alongside traditional analyses for analysis of variance, linear models, and metric and choice-based conjoint analysis. With its emphasis on data visualization, model assessment, and development of statistical intuition, this book provides guidance for any analyst looking to develop or improve skills in R for marketing applications.
Building Machine Learning Pipelines Apress
Learn how to use, deploy, and maintain Apache Spark with this comprehensive guide, written by the creators of the open-source cluster-computing framework. With an emphasis on improvements and

new features in Spark 2.0, authors Bill Chambers and Matei Zaharia break down Spark topics into distinct sections, each with unique goals. Youâ??ll explore the basic operations and common functions of Sparkâ??s structured APIs, as well as Structured Streaming, a new high-level API for building end-to-end streaming applications. Developers and system administrators will learn the fundamentals of monitoring, tuning, and debugging Spark, and explore machine learning techniques and scenarios for employing MLlib, Sparkâ??s scalable machine-learning library. Get a gentle overview of big data and Spark Learn about DataFrames, SQL, and Datasetsâ??Sparkâ??s core APIsâ??through worked examples Dive into Sparkâ??s low-level APIs, RDDs, and execution of SQL and DataFrames Understand how Spark runs on a cluster Debug, monitor, and tune Spark clusters and applications Learn the power of Structured Streaming, Sparkâ??s stream-processing engine Learn how you can apply MLlib to a variety of problems, including classification or recommendation

Advanced Analytics with Spark Packt Publishing Ltd

Think big about your data! PySpark brings the powerful Spark big data processing engine to the Python ecosystem, letting you seamlessly scale up your data tasks and create lightning-fast pipelines.In Data Analysis with Python and PySpark you will learn how to:Manage your data as it scales across multiple machines, Scale up your data programs with full confidence, Read and write data to and from a variety of sources and formats, Deal with messy data with PySpark's data manipulation functionality, Discover new data sets and perform exploratory data analysis, Build automated data pipelines that transform, summarize, and get insights from data, Troubleshoot common PySpark errors, Creating reliable long-running jobs. Data Analysis with Python and PySpark is your guide to delivering successful Python-driven data projects. Packed with relevant examples and essential techniques, this practical book teaches you to build pipelines for reporting, machine learning, and other data-centric tasks. Quick exercises in every chapter help you practice what you've learned, and rapidly start implementing PySpark into your data systems. No previous knowledge of Spark is required.Data Analysis with Python and PySpark helps you solve the daily challenges of data science with PySpark. You'll learn how to scale your processing capabilities across multiple machines while ingesting data from any source--whether that's Hadoop clusters, cloud data storage, or local data files. Once you've covered the fundamentals, you'll explore the full versatility of PySpark by building machine learning pipelines, and blending Python, pandas, and PySpark code.

Powerful Python Data Analysis with Python and PySpark

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

Spark: The Definitive Guide "O'Reilly Media, Inc."

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

Cracking the Data Science Interview "O'Reilly Media, Inc."

Build real-world Artificial Intelligence applications with Python to intelligently interact with the world around you About This Book Step into the amazing world of intelligent apps using this comprehensive guide Enter the world of Artificial Intelligence, explore it, and create your own applications Work through simple yet insightful examples that will get you up and running with Artificial Intelligence in no time Who This Book Is For This book is for Python developers who want to build real-world Artificial Intelligence applications. This book is friendly to Python beginners, but being familiar with Python would be useful to play around with the code. It will also be useful for experienced Python programmers who are looking to use Artificial Intelligence techniques in their existing technology stacks. What You Will Learn Realize different classification and regression techniques Understand the concept of clustering and how to use it to automatically segment data See how to build an intelligent recommender system Understand logic programming and how to use it Build automatic speech recognition systems Understand the basics of heuristic search and genetic programming Develop games using Artificial Intelligence Learn how reinforcement learning works Discover how to build intelligent applications centered on images, text, and time series data See how to use deep learning algorithms and build applications based on it In Detail Artificial Intelligence is becoming increasingly relevant in the modern world where everything is driven by technology and data. It is used extensively across many fields such as search engines, image recognition, robotics, finance, and so on. We will explore various real-world scenarios in this book and you'll learn about various algorithms that can be used to build Artificial Intelligence applications. During the course of this book, you will find out how to make informed decisions about what algorithms to use in a given context. Starting from the basics of Artificial Intelligence, you will learn how to develop various building blocks using different data mining techniques. You will see how to implement different algorithms to get the best possible results, and will understand how to apply them to real-world scenarios. If you want to add an intelligence layer to any application that's based on images, text, stock market, or some other form of data, this exciting book on Artificial Intelligence will definitely be your guide! Style and approach This highly practical book will show you how to implement

Artificial Intelligence. The book provides multiple examples enabling you to create smart applications to meet the needs of your organization. In every chapter, we explain an algorithm, implement it, and then build a smart application.

Parallel and Concurrent Programming in Haskell Packt Publishing

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

Learning PySpark "O'Reilly Media, Inc."

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

Learning Spark "O'Reilly Media, Inc."

The "Writing Idiomatic Python" book is finally here! Chock full of code samples, you'll learn the "Pythonic" way to accomplish common tasks. Each idiom comes with a detailed description, example code showing the "wrong" way to do it, and code for the idiomatic, "Pythonic" alternative. *This version of the book is for Python 3.3+. There is also a Python 2.7.3+ version available.* "Writing Idiomatic Python" contains the most common and important Python idioms in a format that maximizes identification and understanding. Each idiom is presented as a recommendation to write some commonly used piece of code. It is followed by an explanation of why the idiom is important. It also contains two code samples: the "Harmful" way to write it and the "Idiomatic" way. * The "Harmful" way helps you identify the idiom in your own code. * The "Idiomatic" way shows you how to easily translate that code into idiomatic Python. This book is perfect for you: * If you're coming to Python from another programming language * If you're learning Python as a first programming language * If you're looking to increase the readability, maintainability, and correctness of your Python code What is "Idiomatic" Python? Every programming language has its own idioms. Programming language idioms are nothing more than the generally accepted way of writing a certain piece of code. Consistently writing idiomatic code has a number of important benefits: * Others can read and understand your code easily * Others can maintain and enhance your code with minimal effort * Your code will contain fewer bugs * Your code will teach others to write correct code without any effort on your part

"O'Reilly Media, Inc."

Take your Python skills to the next level to develop scalable, real-world applications for local as well as cloud deployment Key FeaturesAll code examples have been tested with Python 3.7 and Python 3.8 and are expected to work with any future 3.x releaseLearn how to build modular and object-oriented applications in PythonDiscover how to use advanced Python techniques for the cloud and clustersBook Description Python is a multipurpose language that can be used for multiple use cases. Python for Geeks will teach you how to advance in your career with the help of expert tips and tricks. You'll start by exploring the different ways of using Python optimally, both from the design and implementation point of view. Next, you'll understand the life cycle of a large-scale Python project. As you advance, you'll focus on different ways of creating an elegant design by modularizing a Python project and learn best practices and design patterns for using Python. You'll also discover how to scale out Python beyond a single thread and how to implement multiprocessing and multithreading in Python. In addition to this, you'll understand how you can not only use Python to deploy on a single machine but also use clusters in private as well as in public cloud computing environments. You'll then explore data processing techniques, focus on reusable, scalable data pipelines, and learn how to use these advanced techniques for network automation, serverless functions, and machine learning. Finally, you'll focus on strategizing web development design using the techniques and best practices covered in the book. By the end of this Python book, you'll be able to do some serious Python programming for large-scale complex projects. What you will learnUnderstand how to design and manage complex Python projectsStrategize test-driven development (TDD) in PythonExplore multithreading and multiprogramming in PythonUse Python for data processing with Apache Spark and Google Cloud Platform (GCP)Deploy serverless programs on public clouds such as GCPUse Python to build web applications and application programming interfacesApply Python for network automation and serverless functionsGet to grips with Python for data analysis and machine learningWho this book is for This book is for intermediate-level Python developers in any field who are looking to build their skills to develop and manage large-scale complex projects. Developers who want to create reusable modules and Python libraries and cloud developers building applications for cloud deployment will also find this book useful. Prior experience with Python will help you get the most out of this book.

Data Engineering with Google Cloud Platform Gareth Stevens Publishing LLLP

Build and deploy your own data pipelines on GCP, make key architectural decisions, and gain the

confidence to boost your career as a data engineer Key FeaturesUnderstand data engineering concepts, the role of a data engineer, and the benefits of using GCP for building your solutionLearn how to use the various GCP products to ingest, consume, and transform data and orchestrate pipelinesDiscover tips to prepare for and pass the Professional Data Engineer examBook Description With this book, you'll understand how the highly scalable Google Cloud Platform (GCP) enables data engineers to create end-to-end data pipelines right from storing and processing data and workflow orchestration to presenting data through visualization dashboards. Starting with a quick overview of the fundamental concepts of data engineering, you'll learn the various responsibilities of a data engineer and how GCP plays a vital role in fulfilling those responsibilities. As you progress through the chapters, you'll be able to leverage GCP products to build a sample data warehouse using Cloud Storage and BigQuery and a data lake using Dataproc. The book gradually takes you through operations such as data ingestion, data cleansing, transformation, and integrating data with other sources. You'll learn how to design IAM for data governance, deploy ML pipelines with the Vertex AI, leverage pre-built GCP models as a service, and visualize data with Google Data Studio to build compelling reports. Finally, you'll find tips on how to boost your career as a data engineer, take the Professional Data Engineer certification exam, and get ready to become an expert in data engineering with GCP. By the end of this data engineering book, you'll have developed the skills to perform core data engineering tasks and build efficient ETL data pipelines with GCP. What you will learnLoad data into BigQuery and materialize its output for downstream consumptionBuild data pipeline orchestration using Cloud ComposerDevelop Airflow jobs to orchestrate and automate a data warehouseBuild a Hadoop data lake, create ephemeral clusters, and run jobs on the Dataproc clusterLeverage Pub/Sub for messaging and ingestion for event-driven systemsUse Dataflow to perform ETL on streaming dataUnlock the power of your data with Data StudioCalculate the GCP cost estimation for your end-to-end data solutionsWho this book is for This book is for data engineers, data analysts, and anyone looking to design and manage data processing pipelines using GCP. You'll find this book useful if you are preparing to take Google's Professional Data Engineer exam. Beginner-level understanding of data science, the Python programming language, and Linux commands is necessary. A basic understanding of data processing and cloud computing, in general, will help you make the most out of this book.

Data Science in Production "O'Reilly Media, Inc."

Summary Introducing Data Science teaches you how to accomplish the fundamental tasks that occupy data scientists. Using the Python language and common Python libraries, you'll experience firsthand the challenges of dealing with data at scale and gain a solid foundation in data science. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Many companies need developers with data science skills to work on projects ranging from social media marketing to machine learning. Discovering what you need to learn to begin a career as a data scientist can seem bewildering. This book is designed to help you get started. About the Book Introducing Data ScienceIntroducing Data Science explains vital data science concepts and teaches you how to accomplish the fundamental tasks that occupy data scientists. You'll explore data visualization, graph databases, the use of NoSQL, and the data science process. You'll use the Python language and common Python libraries as you experience firsthand the challenges of dealing with data at scale. Discover how Python allows you to gain insights from data sets so big that they need to be stored on multiple machines, or from data moving so quickly that no single machine can handle it. This book gives you hands-on experience with the most popular Python data science libraries, Scikit-learn and StatsModels. After reading this book, you'll have the solid foundation you need to start a career in data science. What's Inside Handling large data Introduction to machine learning Using Python to work with data Writing data science algorithms About the Reader This book assumes you're comfortable reading code in Python or a similar language, such as C, Ruby, or JavaScript. No prior experience with data science is required. About the Authors Davy Cielen, Arno D. B. Meysman, and Mohamed Ali are the founders and managing partners of Optimately and Maiton, where they focus on developing data science projects and solutions in various sectors. Table of Contents Data science in a big data world The data science process Machine learning Handling large data on a single computer First steps in big data Join the NoSQL movement The rise of graph databases Text mining and text analytics Data visualization to the end user

High Performance Spark Packt Publishing Ltd

Companies are spending billions on machine learning projects, but it's money wasted if the models can't be deployed effectively. In this practical guide, Hannes Hapke and Catherine Nelson walk you through the steps of automating a machine learning pipeline using the TensorFlow ecosystem. You'll learn the techniques and tools that will cut deployment time from days to minutes, so that you can focus on developing new models rather than maintaining legacy systems. Data scientists, machine learning engineers, and DevOps engineers will discover how to go beyond model development to successfully productize their data science projects, while managers will better understand the role they play in helping to accelerate these projects. Understand the steps to build a machine learning pipeline Build your pipeline using components from TensorFlow Extended Orchestrate your machine learning pipeline with Apache Beam, Apache Airflow, and Kubeflow Pipelines Work with data using TensorFlow Data Validation and TensorFlow Transform Analyze a model in detail using TensorFlow Model Analysis Examine fairness and bias in your model performance Deploy models with TensorFlow Serving or TensorFlow Lite for mobile devices Learn privacy-preserving machine learning techniques

Related with Pyspark Coding Interview Questions:
© Pyspark Coding Interview Questions Cells The Basic Unit Of Life Answer Key
© Pyspark Coding Interview Questions Certified Dialysis Nurse Exam Practice Questions
© Pyspark Coding Interview Questions Celtics Number 9 History