Get Free A Data Pipeline For Phm Data Driven Analytics In Large Pdf For Free

Data Engineering with Apache Spark, Delta Lake, and Lakehouse Mar 10 2022 Understand the complexities of modern-day data engineering platforms and explore strategies to deal with them with the help of use case scenarios led by an industry expert in big data Key Features Become well-versed with the core concepts of Apache Spark and Delta Lake for building data platforms Learn how to ingest, process, and analyze data that can be later used for training machine learning models Understand how to operationalize data models in production using curated data Book Description In the world of ever-changing data and schemas, it is important to build data pipelines that can auto-adjust to changes. This book will help you build scalable data platforms that managers, data scientists, and data analysts can rely on. Starting with an introduction to data engineering, along with its key concepts and architectures, this book will show you how to use Microsoft Azure Cloud services effectively for data engineering. You'll cover data lake

design patterns and the different stages through which the data needs to flow in a typical data lake. Once you've explored the main features of Delta Lake to build data lakes with fast performance and governance in mind, you'll advance to implementing the lambda architecture using Delta Lake. Packed with practical examples and code snippets, this book takes you through real-world examples based on production scenarios faced by the author in his 10 years of experience working with big data. Finally, you'll cover data lake deployment strategies that play an important role in provisioning the cloud resources and deploying the data pipelines in a repeatable and continuous way. By the end of this data engineering book, you'll know how to effectively deal with ever-changing data and create scalable data pipelines to streamline data science, ML, and artificial intelligence (AI) tasks. What you will learn Discover the challenges you may face in the data engineering world Add ACID transactions to Apache Spark using Delta Lake Understand effective design strategies to build enterprise-grade data lakes Explore architectural and design patterns for building efficient data ingestion pipelines Orchestrate a data pipeline for preprocessing data using Apache Spark and Delta Lake APIs Automate deployment and monitoring of data pipelines in production Get to grips with securing, monitoring, and managing data pipelines models efficiently Who this book is for This book is for aspiring data engineers and data analysts who are new to the world of data engineering and are looking for a practical guide to building scalable data platforms. If you already work with PySpark and want to use Delta Lake for data engineering, you'll find

this book useful. Basic knowledge of Python, Spark, and SQL is expected.

Data Science on the Google Cloud Platform Aug 03 2021 Learn how easy it is to apply sophisticated statistical and machine learning methods to real-world problems when you build on top of the Google Cloud Platform (GCP). This hands-on guide shows developers entering the data science field how to implement an end-to-end data pipeline, using statistical and machine learning methods and tools on GCP. Through the course of the book, you'll work through a sample business decision by employing a variety of data science approaches. Follow along by implementing these statistical and machine learning solutions in your own project on GCP, and discover how this platform provides a transformative and more collaborative way of doing data science. You'll learn how to: Automate and schedule data ingest, using an App Engine application Create and populate a dashboard in Google Data Studio Build a real-time analysis pipeline to carry out streaming analytics Conduct interactive data exploration with Google BigQuery Create a Bayesian model on a Cloud Dataproc cluster Build a logistic regression machine-learning model with Spark Compute time-aggregate features with a Cloud Dataflow pipeline Create a high-performing prediction model with TensorFlow Use your deployed model as a microservice you can access from both batch and real-time pipelines

Data Engineering with Apache Spark, Delta Lake, and Lakehouse Mar 30 2021 Understand the complexities of modern-day data engineering platforms and explore strategies to deal with them with the help of use case

scenarios led by an industry expert in big data Key Features: Become well-versed with the core concepts of Apache Spark and Delta Lake for building data platforms Learn how to ingest, process, and analyze data that can be later used for training machine learning models Understand how to operationalize data models in production using curated data Book Description: In the world of ever-changing data and ever-evolving schemas, it is important to build data pipelines that can auto-adjust to changes. This book will help you build scalable data platforms that managers, data scientists, and data analysts can rely on. Starting with an introduction to data engineering, along with its key concepts and architectures, this book will show you how to use Microsoft Azure Cloud services effectively for data engineering. You'll cover data lake design patterns and the different stages through which the data needs to flow in a typical data lake. Once you've explored the main features of Delta Lake to build data lakes with fast performance and governance in mind, you'll advance to implementing the lambda architecture using Delta Lake. Packed with practical examples and code snippets, this book takes you through real-world examples based on production scenarios faced by the author in his 10 years of experience working with big data. Finally, you'll cover data lake deployment strategies that play an important role in provisioning the cloud resources and deploying the data pipelines in a repeatable and continuous way. By the end of this data engineering book, you'll have learned how to effectively deal with ever-changing data and create scalable data pipelines to streamline data science, ML, and artificial intelligence (AI) tasks. What

You Will Learn: Discover the challenges you may face in the data engineering world Add ACID transactions to Apache Spark using Delta Lake Understand effective design strategies to build enterprise-grade data lakes Explore architectural and design patterns for building efficient data ingestion pipelines Orchestrate a data pipeline for preprocessing data using Apache Spark and Delta Lake APIs Automate deployment and monitoring of data pipelines in production Get to grips with securing, monitoring, and managing data pipelines models efficiently Who this book is for: This book is for aspiring data engineers and data analysts who are new to the world of data engineering and are looking for a practical guide to building scalable data platforms. If you already work with PySpark and want to use Delta Lake for data engineering, you'll find this book useful. Basic knowledge of Python, Spark, and SQL is expected.

Data Pipelines with Apache Airflow Jan 20 2023 "An Airflow bible. Useful for all kinds of users, from novice to expert." - Rambabu Posa, Sai Aashika Consultancy Data Pipelines with Apache Airflow teaches you how to build and maintain effective data pipelines. A successful pipeline moves data efficiently, minimizing pauses and blockages between tasks, keeping every process along the way operational. Apache Airflow provides a single customizable environment for building and managing data pipelines, eliminating the need for a hodgepodge collection of tools, snowflake code, and homegrown processes. Using real-world scenarios and examples, Data Pipelines with Apache Airflow teaches you how to simplify and automate data pipelines, reduce operational overhead, and smoothly integrate all the

technologies in your stack. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Data pipelines manage the flow of data from initial collection through consolidation, cleaning, analysis, visualization, and more. Apache Airflow provides a single platform you can use to design, implement, monitor, and maintain your pipelines. Its easy-to-use UI, plug-and-play options, and flexible Python scripting make Airflow perfect for any data management task. About the book Data Pipelines with Apache Airflow teaches you how to build and maintain effective data pipelines. You'll explore the most common usage patterns, including aggregating multiple data sources, connecting to and from data lakes, and cloud deployment. Part reference and part tutorial, this practical guide covers every aspect of the directed acyclic graphs (DAGs) that power Airflow, and how to customize them for your pipeline's needs. What's inside Build, test, and deploy Airflow pipelines as DAGs Automate moving and transforming data Analyze historical datasets using backfilling Develop custom components Set up Airflow in production environments About the reader For DevOps, data engineers, machine learning engineers, and sysadmins with intermediate Python skills. About the author Bas Harenslak and Julian de Ruiter are data engineers with extensive experience using Airflow to develop pipelines for major companies. Bas is also an Airflow committer. Table of Contents PART 1 - GETTING STARTED 1 Meet Apache Airflow 2 Anatomy of an Airflow DAG 3 Scheduling in Airflow 4 Templating tasks using the Airflow context 5

Defining dependencies between tasks PART 2 - BEYOND THE BASICS 6 Triggering workflows 7 Communicating with external systems 8 Building custom components 9 Testing 10 Running tasks in containers PART 3 - AIRFLOW IN PRACTICE 11 Best practices 12 Operating Airflow in production 13 Securing Airflow 14 Project: Finding the fastest way to get around NYC PART 4 - IN THE CLOUDS 15 Airflow in the clouds 16 Airflow on AWS 17 Airflow on Azure 18 Airflow in GCP

Practical Real-time Data Processing and Analytics Oct 25 2020 A practical guide to help you tackle different real-time data processing and analytics problems using the best tools for each scenario About This Book Learn about the various challenges in real-time data processing and use the right tools to overcome them This book covers popular tools and frameworks such as Spark, Flink, and Apache Storm to solve all your distributed processing problems A practical guide filled with examples, tips, and tricks to help you perform efficient Big Data processing in real-time Who This Book Is For If you are a Java developer who would like to be equipped with all the tools required to devise an end-to-end practical solution on real-time data streaming, then this book is for you. Basic knowledge of real-time processing would be helpful, and knowing the fundamentals of Maven, Shell, and Eclipse would be great. What You Will Learn Get an introduction to the established real-time stack Understand the key integration of all the components Get a thorough understanding of the basic building blocks for real-time solution designing Garnish the search and visualization aspects for your real-time solution Get conceptually and

practically acquainted with real-time analytics Be well equipped to apply the knowledge and create your own solutions In Detail With the rise of Big Data, there is an increasing need to process large amounts of data continuously, with a shorter turnaround time. Real-time data processing involves continuous input, processing and output of data, with the condition that the time required for processing is as short as possible. This book covers the majority of the existing and evolving open source technology stack for real-time processing and analytics. You will get to know about all the real-time solution aspects, from the source to the presentation to persistence. Through this practical book, you'll be equipped with a clear understanding of how to solve challenges on your own. We'll cover topics such as how to set up components, basic executions, integrations, advanced use cases, alerts, and monitoring. You'll be exposed to the popular tools used in real-time processing today such as Apache Spark, Apache Flink, and Storm. Finally, you will put your knowledge to practical use by implementing all of the techniques in the form of a practical, real-world use case. By the end of this book, you will have a solid understanding of all the aspects of real-time data processing and analytics, and will know how to deploy the solutions in production environments in the best possible manner. Style and Approach In this practical guide to real-time analytics, each chapter begins with a basic high-level concept of the topic, followed by a practical, hands-on implementation of each concept, where you can see the working and execution of it. The book is written in a DIY style, with plenty of practical use cases, well-explained code

examples, and relevant screenshots and diagrams.
Data Science in Education Using R Jul 02 2021 Data Science in Education Using R is the go-to reference for learning data science in the education field. The book answers questions like: What does a data scientist in education do? How do I get started learning R, the popular open-source statistical programming language? And what does a data analysis project in education look like? If you're just getting started with R in an education job, this is the book you'll want with you. This book gets you started with R by teaching the building blocks of programming that you'll use many times in your career. The book takes a "learn by doing" approach and offers eight analysis walkthroughs that show you a data analysis from start to finish, complete with code for you to practice with. The book finishes with how to get involved in the data science community and how to integrate data science in your education job. This book will be an essential resource for education professionals and researchers looking to increase their data analysis skills as part of their professional and academic development.
Performance Dashboards Jun 01 2021 Tips, techniques, and trends on how to use dashboard technology tooptimize business performance Business performance management is a hot new management disciplinethat delivers tremendous value when supported by informationtechnology. Through case studies and industry research, this bookshows how leading companies are using performance dashboards toexecute strategy, optimize business processes, and improveperformance. Wayne W. Eckerson (Hingham, MA) is the Director of Research for TheData Warehousing

Institute (TDWI), the leading association ofbusiness intelligence and data warehousing professionals worldwidethat provide high-quality, in-depth education, training, andresearch. He is a columnist for SearchCIO.com, DM Review,Application Development Trends, the Business Intelligence Journal,and TDWI Case Studies & Solution.

Spotlight on Learning from Failure Nov 25 2020 When considering the complexities of implementing a next-generation data pipeline, the risk of over-promising and under-delivering is incredibly high due to the overwhelming expectations placed on predictive analytics today. It's critical to look at a number of factors-like your company's culture and the historic promise gap between ETL (extract, transform, load) in the 1970s and AI (artificial intelligence) in the 2010s-to overcome the challenges and succeed. Natalino Busa discusses how to bust the myths, deconstruct the false assumptions, and avoid common pitfalls in order to build and deliver successful data pipelines that deliver real value to your organization. Recorded on February 26, 2019. See the original event page for resources for further learning. Find future live events to attend or watch recordings of other past events . O'Reilly Spotlight explores emerging business and technology topics and ideas through a series of one-hour interactive events. In live conversations, participants share their questions and ideas while hearing the experts' unique perspectives, insights, fears, and predictions for the future. In every edition of Spotlight on Learning from Failure , you'll discover the lessons learned from failures both large and small. You'll discover how successful companies have addressed setbacks, missteps, and challenges and how you

can grow from their examples.

Building Data Pipelines with Python Sep 04 2021 "This course shows you how to build data pipelines and automate workflows using Python 3. From simple task-based messaging queues to complex frameworks like Luigi and Airflow, the course delivers the essential knowledge you need to develop your own automation solutions. You'll learn the architecture basics, and receive an introduction to a wide variety of the most popular frameworks and tools. Designed for the working data professional who is new to the world of data pipelines and distributed solutions, the course requires intermediate level Python experience and the ability to manage your own system set-ups."--Resource description page.

Cost-Effective Data Pipelines Nov 06 2021 The low cost of getting started with cloud services can easily evolve into a significant expense down the road. That's challenging for teams developing data pipelines, particularly when rapid changes in technology and workload require a constant cycle of redesign. How do you deliver scalable, highly available products while keeping costs in check? With this practical guide, author Sev Leonard provides a holistic approach to designing scalable data pipelines in the cloud. Intermediate data engineers, software developers, and architects will learn how to navigate cost/performance trade-offs and how to choose and configure compute and storage. You'll also pick up best practices for code development, testing, and monitoring. By focusing on the entire design process, you'll be able to deliver cost-effective, high-quality products. This book helps you: Reduce cloud spend with lower cost cloud

service offerings and smart design strategies Minimize waste without sacrificing performance by rightsizing compute resources Drive pipeline evolution, head off performance issues, and quickly debug with effective monitoring Set up development and test environments that minimize cloud service dependencies Create data pipeline code bases that are testable and extensible, fostering rapid development and evolution Improve data quality and pipeline operation through validation and testing

R for Data Science Dec 27 2020 Learn how to use R to turn raw data into insight, knowledge, and understanding. This book introduces you to R, RStudio, and the tidyverse, a collection of R packages designed to work together to make data science fast, fluent, and fun. Suitable for readers with no previous programming experience, R for Data Science is designed to get you doing data science as quickly as possible. Authors Hadley Wickham and Garrett Grolemund guide you through the steps of importing, wrangling, exploring, and modeling your data and communicating the results. You'll get a complete, big-picture understanding of the data science cycle, along with basic tools you need to manage the details. Each section of the book is paired with exercises to help you practice what you've learned along the way. You'll learn how to: Wrangle—transform your datasets into a form convenient for analysis Program—learn powerful R tools for solving data problems with greater clarity and ease Explore—examine your data, generate hypotheses, and quickly test them Model—provide a low-dimensional summary that captures true "signals" in your dataset Communicate—learn R Markdown for integrating prose,

code, and results

Reproducible Data Science with Pachyderm Sep 23 2020 Create scalable and reliable data pipelines easily with Pachyderm Key FeaturesLearn how to build an enterprise-level reproducible data science platform with PachydermDeploy Pachyderm on cloud platforms such as AWS EKS, Google Kubernetes Engine, and Microsoft Azure Kubernetes ServiceIntegrate Pachyderm with other data science tools, such as Pachyderm NotebooksBook Description Pachyderm is an open source project that enables data scientists to run reproducible data pipelines and scale them to an enterprise level. This book will teach you how to implement Pachyderm to create collaborative data science workflows and reproduce your ML experiments at scale. You'll begin your journey by exploring the importance of data reproducibility and comparing different data science platforms. Next, you'll explore how Pachyderm fits into the picture and its significance, followed by learning how to install Pachyderm locally on your computer or a cloud platform of your choice. You'll then discover the architectural components and Pachyderm's main pipeline principles and concepts. The book demonstrates how to use Pachyderm components to create your first data pipeline and advances to cover common operations involving data, such as uploading data to and from Pachyderm to create more complex pipelines. Based on what you've learned, you'll develop an end-to-end ML workflow, before trying out the hyperparameter tuning technique and the different supported Pachyderm language clients. Finally, you'll learn how to use a SaaS version of Pachyderm with Pachyderm Notebooks.

By the end of this book, you will learn all aspects of running your data pipelines in Pachyderm and manage them on a day-to-day basis. What you will learnUnderstand the importance of reproducible data science for enterpriseExplore the basics of Pachyderm, such as commits and branchesUpload data to and from PachydermImplement common pipeline operations in PachydermCreate a real-life example of hyperparameter tuning in PachydermCombine Pachyderm with Pachyderm language clients in Python and GoWho this book is for This book is for new as well as experienced data scientists and machine learning engineers who want to build scalable infrastructures for their data science projects. Basic knowledge of Python programming and Kubernetes will be beneficial. Familiarity with Golang will be helpful.

Cloud Native Data Pipelines with Apache Kafka Jun 13 2022 As microservices, data services, and serverless APIs proliferate in a cloud native world, analysts still need to report on the business as a whole. Data engineers need to collect and standardize data in an increasingly complex and diverse system. Luckily, the problem is also the solution. The way to manage data in a cloud native environment is to build cloud native data pipelines. Gwen Shapira (Confluent) discusses how data engineering requirements have changed in a cloud native world and how the solutions have changed with them. She then shares architectural patterns that are commonly used to build cloud native data infrastructure and explains how they help you build flexible, scalable, and reliable pipelines to give your business visibility on all your data. This session was recorded at the 2019 O'Reilly Strata Data Conference in San Francisco.

Streaming Data May 12 2022 Summary Streaming Data introduces the concepts and requirements of streaming and real-time data systems. The book is an idea-rich tutorial that teaches you to think about how to efficiently interact with fast-flowing data. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology As humans, we're constantly filtering and deciphering the information streaming toward us. In the same way, streaming data applications can accomplish amazing tasks like reading live location data to recommend nearby services, tracking faults with machinery in real time, and sending digital receipts before your customers leave the shop. Recent advances in streaming data technology and techniques make it possible for any developer to build these applications if they have the right mindset. This book will let you join them. About the Book Streaming Data is an idea-rich tutorial that teaches you to think about efficiently interacting with fast-flowing data. Through relevant examples and illustrated use cases, you'll explore designs for applications that read, analyze, share, and store streaming data. Along the way, you'll discover the roles of key technologies like Spark, Storm, Kafka, Flink, RabbitMQ, and more. This book offers the perfect balance between big-picture thinking and implementation details. What's Inside The right way to collect real-time data Architecting a streaming pipeline Analyzing the data Which technologies to use and when About the Reader Written for developers familiar with relational database concepts. No experience with streaming or real-time applications required. About the Author Andrew Psaltis is a software engineer

focused on massively scalable real-time analytics. Table of Contents PART 1 - A NEW HOLISTIC APPROACH Introducing streaming data Getting data from clients: data ingestion Transporting the data from collection tier: decoupling the data pipeline Analyzing streaming data Algorithms for data analysis Storing the analyzed or collected data Making the data available Consumer device capabilities and limitations accessing the data PART 2 - TAKING IT REAL WORLD Analyzing Meetup RSVPs in real time

AWS Data Pipeline Developer Guide Oct 05 2021 AWS Data Pipeline is a web service that you can use to automate the movement and transformation of data. With AWS Data Pipeline, you can define data-driven workflows, so that tasks can be dependent on the successful completion of previous tasks. You define the parameters of your data transformations and AWS Data Pipeline enforces the logic that you've set up.

Data Engineering with AWS Jan 28 2021 Start your AWS data engineering journey with this easy-to-follow, hands-on guide and get to grips with foundational concepts through to building data engineering pipelines using AWS Key Features: Learn about common data architectures and modern approaches to generating value from big data Explore AWS tools for ingesting, transforming, and consuming data, and for orchestrating pipelines Learn how to architect and implement data lakes and data lakehouses for big data analytics Book Description: Knowing how to architect and implement complex data pipelines is a highly sought-after skill. Data engineers are responsible for building

these pipelines that ingest, transform, and join raw datasets - creating new value from the data in the process. Amazon Web Services (AWS) offers a range of tools to simplify a data engineer's job, making it the preferred platform for performing data engineering tasks. This book will take you through the services and the skills you need to architect and implement data pipelines on AWS. You'll begin by reviewing important data engineering concepts and some of the core AWS services that form a part of the data engineer's toolkit. You'll then architect a data pipeline, review raw data sources, transform the data, and learn how the transformed data is used by various data consumers. The book also teaches you about populating data marts and data warehouses along with how a data lakehouse fits into the picture. Later, you'll be introduced to AWS tools for analyzing data, including those for ad-hoc SQL queries and creating visualizations. In the final chapters, you'll understand how the power of machine learning and artificial intelligence can be used to draw new insights from data. By the end of this AWS book, you'll be able to carry out data engineering tasks and implement a data pipeline on AWS independently. What You Will Learn: Understand data engineering concepts and emerging technologies Ingest streaming data with Amazon Kinesis Data Firehose Optimize, denormalize, and join datasets with AWS Glue Studio Use Amazon S3 events to trigger a Lambda process to transform a file Run complex SQL queries on data lake data using Amazon Athena Load data into a Redshift data warehouse and run queries Create a visualization of your data using Amazon QuickSight Extract sentiment data from a dataset using Amazon Comprehend

Who this book is for: This book is for data engineers, data analysts, and data architects who are new to AWS and looking to extend their skills to the AWS cloud. Anyone who is new to data engineering and wants to learn about the foundational concepts while gaining practical experience with common data engineering services on AWS will also find this book useful. A basic understanding of big data-related topics and Python coding will help you get the most out of this book but is not needed. Familiarity with the AWS console and core services is also useful but not necessary.
Data Science on the Google Cloud Platform Aug 15 2022 Learn how easy it is to apply sophisticated statistical and machine learning methods to real-world problems when you build using Google Cloud Platform (GCP). This hands-on guide shows data engineers and data scientists how to implement an end-to-end data pipeline, using statistical and machine learning methods and tools on GCP. Through the course of this updated second edition, you'll work through a sample business decision by employing a variety of data science approaches. Follow along by implementing these statistical and machine learning solutions in your own project on GCP, and discover how this platform provides a transformative and more collaborative way of doing data science. You'll learn how to: Employ best practices in building highly scalable data and ML pipelines on Google Cloud Automate and schedule data ingest using Cloud Run Create and populate a dashboard in Data Studio Build a real-time analytics pipeline using Pub/Sub, Dataflow, and BigQuery Conduct interactive data exploration with BigQuery Create a Bayesian model with Spark on Cloud

Dataproc Forecast time series and do anomaly detection with BigQuery ML Aggregate within time windows with Dataflow Train explainable machine learning models with Vertex AI Operationalize ML with Vertex AI Pipelines

Data Pipelines Pocket Reference Feb 21 2023 Data pipelines are the foundation for success in data analytics. Moving data from numerous diverse sources and transforming it to provide context is the difference between having data and actually gaining value from it. This pocket reference defines data pipelines and explains how they work in today's modern data stack. You'll learn common considerations and key decision points when implementing pipelines, such as batch versus streaming data ingestion and build versus buy. This book addresses the most common decisions made by data professionals and discusses foundational concepts that apply to open source frameworks, commercial products, and homegrown solutions. You'll learn: What a data pipeline is and how it works How data is moved and processed on modern data infrastructure, including cloud platforms Common tools and products used by data engineers to build pipelines How pipelines support analytics and reporting needs Considerations for pipeline maintenance, testing, and alerting

Architecting Data Intensive Applications Sep 16 2022 Architect and design data-intensive applications and, in the process, learn how to collect, process, store, govern, and expose data for a variety of use cases Key Features Integrate the data-intensive approach into your application architecture Create a robust application layout with effective messaging and data querying architecture Enable smooth data flow and

make the data of your application intensive and fast Book Description Are you an architect or a developer who looks at your own applications gingerly while browsing through Facebook and applauding it silently for its data-intensive, yet ?uent and efficient, behaviour? This book is your gateway to build smart data-intensive systems by incorporating the core data-intensive architectural principles, patterns, and techniques directly into your application architecture. This book starts by taking you through the primary design challenges involved with architecting data-intensive applications. You will learn how to implement data curation and data dissemination, depending on the volume of your data. You will then implement your application architecture one step at a time. You will get to grips with implementing the correct message delivery protocols and creating a data layer that doesn't fail when running high traffic. This book will show you how you can divide your application into layers, each of which adheres to the single responsibility principle. By the end of this book, you will learn to streamline your thoughts and make the right choice in terms of technologies and architectural principles based on the problem at hand. What you will learn Understand how to envision a data-intensive system Identify and compare the non-functional requirements of a data collection component Understand patterns involving data processing, as well as technologies that help to speed up the development of data processing systems Understand how to implement Data Governance policies at design time using various Open Source Tools Recognize the anti-patterns to avoid while designing a data store for applications Understand the

different data dissemination technologies available to query the data in an efficient manner Implement a simple data governance policy that can be extended using Apache Falcon Who this book is for This book is for developers and data architects who have to code, test, deploy, and/or maintain large-scale, high data volume applications. It is also useful for system architects who need to understand various non-functional aspects revolving around Data Intensive Systems.

Big Data Processing with Apache Spark Nov 13 2019 Apache Spark is a popular open-source big-data processing framework thatÕs built around speed, ease of use, and unified distributed computing architecture. Not only it supports developing applications in different languages like Java, Scala, Python, and R, itÕs also hundred times faster in memory and ten times faster even when running on disk compared to traditional data processing frameworks. Whether you are currently working on a big data project or interested in learning more about topics like machine learning, streaming data processing, and graph data analytics, this book is for you. You can learn about Apache Spark and develop Spark programs for various use cases in big data analytics using the code examples provided. This book covers all the libraries in Spark ecosystem: Spark Core, Spark SQL, Spark Streaming, Spark ML, and Spark GraphX.

Fast Data Processing with Spark 2 Mar 18 2020 Learn how to use Spark to process big data at speed and scale for sharper analytics. Put the principles into practice for faster, slicker big data projects. About This Book A quick way to get started with Spark – and reap the rewards From analytics to engineering your big data architecture, we've got it

covered Bring your Scala and Java knowledge – and put it to work on new and exciting problems Who This Book Is For This book is for developers with little to no knowledge of Spark, but with a background in Scala/Java programming. It's recommended that you have experience in dealing and working with big data and a strong interest in data science. What You Will Learn Install and set up Spark in your cluster Prototype distributed applications with Spark's interactive shell Perform data wrangling using the new DataFrame APIs Get to know the different ways to interact with Spark's distributed representation of data (RDDs) Query Spark with a SQL-like query syntax See how Spark works with big data Implement machine learning systems with highly scalable algorithms Use R, the popular statistical language, to work with Spark Apply interesting graph algorithms and graph processing with GraphX In Detail When people want a way to process big data at speed, Spark is invariably the solution. With its ease of development (in comparison to the relative complexity of Hadoop), it's unsurprising that it's becoming popular with data analysts and engineers everywhere. Beginning with the fundamentals, we'll show you how to get set up with Spark with minimum fuss. You'll then get to grips with some simple APIs before investigating machine learning and graph processing – throughout we'll make sure you know exactly how to apply your knowledge. You will also learn how to use the Spark shell, how to load data before finding out how to build and run your own Spark applications. Discover how to manipulate your RDD and get stuck into a range of DataFrame APIs. As if that's not enough, you'll also learn some useful Machine Learning

algorithms with the help of Spark MLlib and integrating Spark with R. We'll also make sure you're confident and prepared for graph processing, as you learn more about the GraphX API. Style and approach This book is a basic, step-by-step tutorial that will help you take advantage of all that Spark has to offer.

Data Pipelines with Apache Airflow Jul 14 2022 This book teaches you how to build and maintain effective data pipelines. Youll explore the most common usage patterns, including aggregating multiple data sources, connecting to and from data lakes, and cloud deployment. --

Mastering Social Media Mining with Python Dec 19 2022 Acquire and analyze data from all corners of the social web with Python About This Book Make sense of highly unstructured social media data with the help of the insightful use cases provided in this guide Use this easy-to-follow, step-by-step guide to apply analytics to complicated and messy social data This is your one-stop solution to fetching, storing, analyzing, and visualizing social media data Who This Book Is For This book is for intermediate Python developers who want to engage with the use of public APIs to collect data from social media platforms and perform statistical analysis in order to produce useful insights from data. The book assumes a basic understanding of the Python Standard Library and provides practical examples to guide you toward the creation of your data analysis project based on social data. What You Will Learn Interact with a social media platform via their public API with Python Store social data in a convenient format for data analysis Slice and dice social data using Python tools for data science Apply text

analytics techniques to understand what people are talking about on social media Apply advanced statistical and analytical techniques to produce useful insights from data Build beautiful visualizations with web technologies to explore data and present data products In Detail Your social media is filled with a wealth of hidden data – unlock it with the power of Python. Transform your understanding of your clients and customers when you use Python to solve the problems of understanding consumer behavior and turning raw data into actionable customer insights. This book will help you acquire and analyze data from leading social media sites. It will show you how to employ scientific Python tools to mine popular social websites such as Facebook, Twitter, Quora, and more. Explore the Python libraries used for social media mining, and get the tips, tricks, and insider insight you need to make the most of them. Discover how to develop data mining tools that use a social media API, and how to create your own data analysis projects using Python for clear insight from your social data. Style and approach This practical, hands-on guide will help you learn everything you need to perform data mining for social media. Throughout the book, we take an example-oriented approach to use Python for data analysis and provide useful tips and tricks that you can use in day-to-day tasks.

LSST Data Pipeline Prototyping Plans and Strategy Feb 15 2020 In this document we describe our approach and strategy for building the prototype for the image-stream analysis data pipeline. We start by describing the main research areas upon which we will be focusing; we then describe our plans on how to carry these research ideas to implement the data

pipeline.

Building Machine Learning Pipelines Oct 17 2022 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

Data Engineering with Python Nov 18 2022 Build, monitor, and manage real-time data pipelines to create data engineering infrastructure efficiently using open-source Apache projects Key FeaturesBecome well-versed in data architectures, data preparation, and data optimization skills

with the help of practical examplesDesign data models and learn how to extract, transform, and load (ETL) data using PythonSchedule, automate, and monitor complex data pipelines in productionBook Description Data engineering provides the foundation for data science and analytics, and forms an important part of all businesses. This book will help you to explore various tools and methods that are used for understanding the data engineering process using Python. The book will show you how to tackle challenges commonly faced in different aspects of data engineering. You'll start with an introduction to the basics of data engineering, along with the technologies and frameworks required to build data pipelines to work with large datasets. You'll learn how to transform and clean data and perform analytics to get the most out of your data. As you advance, you'll discover how to work with big data of varying complexity and production databases, and build data pipelines. Using real-world examples, you'll build architectures on which you'll learn how to deploy data pipelines. By the end of this Python book, you'll have gained a clear understanding of data modeling techniques, and will be able to confidently build data engineering pipelines for tracking data, running quality checks, and making necessary changes in production. What you will learnUnderstand how data engineering supports data science workflowsDiscover how to extract data from files and databases and then clean, transform, and enrich itConfigure processors for handling different file formats as well as both relational and NoSQL databasesFind out how to implement a data pipeline and dashboard to visualize resultsUse staging and validation to check data before

landing in the warehouseBuild real-time pipelines with staging areas that perform validation and handle failuresGet to grips with deploying pipelines in the production environmentWho this book is for This book is for data analysts, ETL developers, and anyone looking to get started with or transition to the field of data engineering or refresh their knowledge of data engineering using Python. This book will also be useful for students planning to build a career in data engineering or IT professionals preparing for a transition. No previous knowledge of data engineering is required.

Building Big Data Pipelines with Apache Beam Apr 11 2022 Implement, run, operate, and test data processing pipelines using Apache Beam Key FeaturesUnderstand how to improve usability and productivity when implementing Beam pipelinesLearn how to use stateful processing to implement complex use cases using Apache BeamImplement, test, and run Apache Beam pipelines with the help of expert tips and techniquesBook Description Apache Beam is an open source unified programming model for implementing and executing data processing pipelines, including Extract, Transform, and Load (ETL), batch, and stream processing. This book will help you to confidently build data processing pipelines with Apache Beam. You'll start with an overview of Apache Beam and understand how to use it to implement basic pipelines. You'll also learn how to test and run the pipelines efficiently. As you progress, you'll explore how to structure your code for reusability and also use various Domain Specific Languages (DSLs). Later chapters will show you how to use schemas and query your

data using (streaming) SQL. Finally, you'll understand advanced Apache Beam concepts, such as implementing your own I/O connectors. By the end of this book, you'll have gained a deep understanding of the Apache Beam model and be able to apply it to solve problems. What you will learnUnderstand the core concepts and architecture of Apache BeamImplement stateless and stateful data processing pipelinesUse state and timers for processing real-time event processingStructure your code for reusabilityUse streaming SQL to process real-time data for increasing productivity and data accessibilityRun a pipeline using a portable runner and implement data processing using the Apache Beam Python SDKImplement Apache Beam I/O connectors using the Splittable DoFn APIWho this book is for This book is for data engineers, data scientists, and data analysts who want to learn how Apache Beam works. Intermediate-level knowledge of the Java programming language is assumed.

Open Source Data Pipelines for Intelligent Applications Apr 30 2021 For decades, businesses have used information about their customers to make critical decisions on what to stock in inventory, which items to recommend to customers, and when to run promotions. But the advent of big data early in this century changed the game considerably. The key to achieving a competitive advantage today is the ability to process and store ever-increasing amounts of information that affect those decisions. In this report, solutions specialists from Red Hat provide an architectural guide to help you navigate the modern data analytics ecosystem. You'll learn how the industry has evolved and examine current

approaches to storage. That includes a deep dive into the anatomy of a portable data platform architecture, along with several aspects of running data pipelines and intelligent applications with Kubernetes. Explore the history of open source data processing and the evolution of container scheduling Get a concise overview of intelligent applications Learn how to use storage with Kubernetes to produce effective intelligent applications Understand how to structure applications on Kubernetes in your platform architecture Delve into example pipeline architectures for deploying intelligent applications on Kubernetes.

GIT-archive Data Pipeline Feb 26 2021

The Essential Guide to Data Integration Jul 22 2020

Modern Enterprise Data Pipelines Feb 09 2022 A Dell Technologies perspective on today's data landscape and the key ingredients for planning a modern, distributed data pipeline for your multicloud data-driven enterprise

Azure Data Factory Cookbook Aug 23 2020 With the help of well-structured and practical recipes, this book will teach you how to integrate data from the cloud and on-premise. You'll learn how to transform, clean, and consolidate data into a single data platform and get to grips with using ADF as the main ETL and orchestration tool for your data warehouse or data platform project.

Understanding Challenges in the Data Pipeline for Development Data Dec 07 2021 The developing world is relying more and more on data driven policies. Numerous development agencies have pushed for on-ground data collection to support the development work they pursue. Many governments have launched efforts for more frequent

information gathering. Overall, the amount of data collected is tremendous, yet we face significant issues in doing useful analysis. Most of these barriers are around data cleaning and merging, and they require a data engineer to support some parts of the analysis. This thesis aims to understand the pain points of cleaning development data. It also proposes solutions that harness the thought process of a data engineer to reduce the manual workload of the tedious process of cleaning such data. To achieve these goals, two research areas are critical: (1) to discern current data usage patterns and to build a taxonomy of data cleaning in the developing world; and (2) to create algorithms to support automated data cleaning, which target selected problems including matching transliterated names. With these goals, this thesis will empower regular data users to easily do the necessary data cleaning and scrubbing for analysis.

Building an Anonymization Pipeline Jan 08 2022 How can you use data in a way that protects individual privacy but still provides useful and meaningful analytics? With this practical book, data architects and engineers will learn how to establish and integrate secure, repeatable anonymization processes into their data flows and analytics in a sustainable manner. Luk Arbuckle and Khaled El Emam from Privacy Analytics explore end-to-end solutions for anonymizing device and IoT data, based on collection models and use cases that address real business needs. These examples come from some of the most demanding data environments, such as healthcare, using approaches that have withstood the test of time. Create anonymization solutions diverse enough to cover a spectrum of use cases Match your solutions to the

data you use, the people you share it with, and your analysis goals Build anonymization pipelines around various data collection models to cover different business needs Generate an anonymized version of original data or use an analytics platform to generate anonymized outputs Examine the ethical issues around the use of anonymized data

Concurrent Data Processing in Elixir Dec 15 2019 Learn different ways of writing concurrent code in Elixir and increase your application's performance, without sacrificing scalability or fault-tolerance. Most projects benefit from running background tasks and processing data concurrently, but the world of OTP and various libraries can be challenging. Which Supervisor and what strategy to use? What about GenServer? Maybe you need back-pressure, but is GenStage, Flow, or Broadway a better choice? You will learn everything you need to know to answer these questions, start building highly concurrent applications in no time, and write code that's not only fast, but also resilient to errors and easy to scale. Whether you are building a high-frequency stock trading application or a consumer web app, you need to know how to leverage concurrency to build applications that are fast and efficient. Elixir and the OTP offer a range of powerful tools, and this guide will show you how to choose the best tool for each job, and use it effectively to quickly start building highly concurrent applications. Learn about Tasks, supervision trees, and the different types of Supervisors available to you. Understand why processes and process linking are the building blocks of concurrency in Elixir. Get comfortable with the OTP and use the GenServer behaviour to maintain process state for long-running jobs.

Easily scale the number of running processes using the Registry. Handle large volumes of data and traffic spikes with GenStage, using back-pressure to your advantage. Create your first multi-stage data processing pipeline using producer, consumer, and producer-consumer stages. Process large collections with Flow, using MapReduce and more in parallel. Thanks to Broadway, you will see how easy it is to integrate with popular message broker systems, or even existing GenStage producers. Start building the high-performance and fault-tolerant applications Elixir is famous for today. What You Need: You'll need Elixir 1.9+ and Erlang/OTP 22+ installed on a Mac OS X, Linux, or Windows machine.

MLOps Engineering at Scale Oct 13 2019 Dodge costly and time-consuming infrastructure tasks, and rapidly bring your machine learning models to production with MLOps and pre-built serverless tools! In MLOps Engineering at Scale you will learn: Extracting, transforming, and loading datasets Querying datasets with SQL Understanding automatic differentiation in PyTorch Deploying model training pipelines as a service endpoint Monitoring and managing your pipeline’s life cycle Measuring performance improvements MLOps Engineering at Scale shows you how to put machine learning into production efficiently by using pre-built services from AWS and other cloud vendors. You’ll learn how to rapidly create flexible and scalable machine learning systems without laboring over time-consuming operational tasks or taking on the costly overhead of physical hardware. Following a real-world use case for calculating taxi fares, you will engineer an MLOps pipeline for a

PyTorch model using AWS server-less capabilities. About the technology A production-ready machine learning system includes efficient data pipelines, integrated monitoring, and means to scale up and down based on demand. Using cloud-based services to implement ML infrastructure reduces development time and lowers hosting costs. Serverless MLOps eliminates the need to build and maintain custom infrastructure, so you can concentrate on your data, models, and algorithms. About the book MLOps Engineering at Scale teaches you how to implement efficient machine learning systems using pre-built services from AWS and other cloud vendors. This easy-to-follow book guides you step-by-step as you set up your serverless ML infrastructure, even if you've never used a cloud platform before. You'll also explore tools like PyTorch Lightning, Optuna, and MLFlow that make it easy to build pipelines and scale your deep learning models in production. What's inside Reduce or eliminate ML infrastructure management Learn state-of-the-art MLOps tools like PyTorch Lightning and MLFlow Deploy training pipelines as a service endpoint Monitor and manage your pipeline's life cycle Measure performance improvements About the reader Readers need to know Python, SQL, and the basics of machine learning. No cloud experience required. About the author Carl Osipov implemented his first neural net in 2000 and has worked on deep learning and machine learning at Google and IBM. Table of Contents PART 1 - MASTERING THE DATA SET 1 Introduction to serverless machine learning 2 Getting started with the data set 3 Exploring and preparing the data set 4 More exploratory data analysis and data preparation PART 2 - PYTORCH

FOR SERVERLESS MACHINE LEARNING 5 Introducing PyTorch: Tensor basics 6 Core PyTorch: Autograd, optimizers, and utilities 7 Serverless machine learning at scale 8 Scaling out with distributed training PART 3 - SERVERLESS MACHINE LEARNING PIPELINE 9 Feature selection 10 Adopting PyTorch Lightning 11 Hyperparameter optimization 12 Machine learning pipeline

Data Science on AWS Jun 20 2020 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
Data Engineering with AWS May 20 2020 The missing expert-led manual for the AWS ecosystem — go from foundations to building data engineering pipelines effortlessly Purchase of the print or Kindle book includes a free eBook in the PDF format. Key Features Learn about common data architectures and modern approaches to generating value from big data Explore AWS tools for ingesting, transforming, and consuming data, and for orchestrating pipelines Learn how to architect and implement data lakes and data lakehouses for big data analytics from a data lakes expert Book Description Written by a Senior Data Architect with over twenty-five years of experience in the business, Data Engineering for AWS is a book whose sole aim is to make you proficient in using the AWS ecosystem. Using a thorough and hands-on approach to data, this book will give aspiring and new data engineers a solid theoretical and practical foundation to succeed with AWS. As you progress, you'll be taken through the services and the skills you need to architect and implement data pipelines on AWS. You'll begin by reviewing important data engineering concepts and some of the core AWS services that form a part of the data engineer's toolkit. You'll then architect a data pipeline, review raw data sources, transform the data, and learn how the transformed data is used by various data consumers. You'll also learn about populating data marts and data warehouses along with how a data lakehouse fits into the picture. Later, you'll be introduced to AWS tools for analyzing data, including those for ad-hoc SQL queries and creating visualizations. In the final chapters, you'll

understand how the power of machine learning and artificial intelligence can be used to draw new insights from data. By the end of this AWS book, you'll be able to carry out data engineering tasks and implement a data pipeline on AWS independently. What you will learn Understand data engineering concepts and emerging technologies Ingest streaming data with Amazon Kinesis Data Firehose Optimize, denormalize, and join datasets with AWS Glue Studio Use Amazon S3 events to trigger a Lambda process to transform a file Run complex SQL queries on data lake data using Amazon Athena Load data into a Redshift data warehouse and run queries Create a visualization of your data using Amazon QuickSight Extract sentiment data from a dataset using Amazon Comprehend Who this book is for This book is for data engineers, data analysts, and data architects who are new to AWS and looking to extend their skills to the AWS cloud. Anyone new to data engineering who wants to learn about the foundational concepts while gaining practical experience with common data engineering services on AWS will also find this book useful. A basic understanding of big data-related topics and Python coding will help you get the most out of this book but it's not a prerequisite. Familiarity with the AWS console and core services will also help you follow along.

Mastering Spark with R Apr 18 2020 If you’re like most R users, you have deep knowledge and love for statistics. But as your organization continues to collect huge amounts of data, adding tools such as Apache Spark makes a lot of sense. With this practical book, data scientists and professionals working with large-scale data applications will

learn how to use Spark from R to tackle big data and big compute problems. Authors Javier Luraschi, Kevin Kuo, and Edgar Ruiz show you how to use R with Spark to solve different data analysis problems. This book covers relevant data science topics, cluster computing, and issues that should interest even the most advanced users. Analyze, explore, transform, and visualize data in Apache Spark with R Create statistical models to extract information and predict outcomes; automate the process in production-ready workflows Perform analysis and modeling across many machines using distributed computing techniques Use large-scale data from multiple sources and different formats with ease from within Spark Learn about alternative modeling frameworks for graph processing, geospatial analysis, and genomics at scale Dive into advanced topics including custom transformations, real-time data processing, and creating custom Spark extensions

Unlock Complex and Streaming Data with Declarative Data Pipelines Jan 16 2020 Unlocking the value of modern data is critical for data-driven companies. This report provides a concise, practical guide to building a data architecture that efficiently delivers big, complex, and streaming data to both internal users and customers. Authors Ori Rafael, Roy Hasson, and Rick Bilodeau from Upsolver examine how modern data pipelines can improve business outcomes. Tech leaders and data engineers will explore the role these pipelines play in the data architecture and learn how to intelligently consider tradeoffs between different data architecture patterns and data pipeline development approaches.