

airflow in data engineering

Airflow in Data Engineering: Revolutionizing Data Pipelines

By Dr. Anya Sharma, PhD, Data Engineering Lead at DataWeave Solutions

(DataWeave Solutions is a leading provider of data integration and analytics solutions, recognized for its innovation and expertise in the field.)

(Edited by Michael Davis, Senior Editor at DataWeave Solutions, with 15+ years of experience in technical writing and data engineering publications.)

Abstract: This article delves into the transformative impact of Apache Airflow in modern data engineering. We explore its core functionalities, benefits, and implications for the industry, examining its role in building robust, scalable, and maintainable data pipelines. We will also discuss best practices and future trends related to airflow in data engineering.

Introduction: The explosion of data in recent years has placed immense pressure on data engineers to build efficient and reliable data pipelines. This is where Apache Airflow, an open-source workflow management platform, shines. Airflow in data engineering has become a cornerstone, enabling the creation of complex data pipelines with ease, scalability, and maintainability that were previously challenging to achieve.

H1: Understanding Apache Airflow in Data Engineering

Apache Airflow is a platform designed to programmatically author, schedule, and monitor workflows. In the context of data engineering, these workflows represent the series of steps required to ingest, process, transform, and load (ETL) data. Unlike traditional scheduling tools, Airflow allows for the creation of complex Directed Acyclic Graphs (DAGs) that define the dependencies between tasks within a workflow. This DAG-based approach provides unparalleled flexibility and control over the entire data pipeline. Airflow in data engineering transcends simple scheduling; it offers a powerful framework for managing the entire lifecycle of data processing.

H2: Key Features of Airflow for Data Engineering

DAGs: The core of Airflow is its DAGs, which visually represent the workflow. This allows for easy understanding, debugging, and modification of complex pipelines.

Task Dependencies: Airflow explicitly defines dependencies between tasks,

ensuring that tasks execute in the correct order. This is crucial for avoiding errors and ensuring data integrity.

Extensibility: Airflow supports a wide range of operators and hooks, allowing integration with various data sources and technologies (databases, cloud storage, machine learning platforms, etc.). This makes airflow in data engineering extremely versatile.

Monitoring and Logging: Airflow provides comprehensive monitoring and logging capabilities, enabling efficient troubleshooting and performance analysis.

Scalability: Airflow can be scaled to handle very large and complex workflows, making it suitable for even the most demanding data engineering projects.

H3: Benefits of Using Airflow in Data Engineering

The adoption of airflow in data engineering offers numerous advantages:

Increased Efficiency: Automation of data pipelines leads to significant time savings and improved resource utilization.

Improved Reliability: Explicit task dependencies and robust monitoring minimize errors and downtime.

Enhanced Maintainability: The DAG-based approach makes it easier to understand, modify, and maintain complex workflows.

Better Collaboration: Airflow facilitates collaboration among data engineers by providing a central platform for managing and monitoring workflows.

Cost Optimization: Automated workflows can optimize resource usage, reducing cloud computing costs.

H4: Real-World Applications of Airflow in Data Engineering

Airflow's versatility makes it applicable across a wide range of data engineering scenarios:

ETL Processes: Building and managing complex ETL pipelines for various data sources.

Data Warehousing: Loading data into data warehouses and data lakes.

Real-time Data Processing: Handling streaming data using tools like Kafka and Spark.

Machine Learning Pipelines: Orchestrating the training and deployment of machine learning models.

Data Quality Monitoring: Implementing data quality checks and alerts.

H5: Challenges and Best Practices for Airflow in Data Engineering

While Airflow offers numerous advantages, there are challenges to consider:

Complexity: Designing and managing complex DAGs can be challenging, especially for large and complex workflows. Best practice involves modular design and breaking down large DAGs into smaller, manageable components.

Monitoring and Alerting: Effective monitoring and alerting are crucial for ensuring the reliability of data pipelines. Implementing robust monitoring and alerting systems is paramount.

Debugging: Debugging complex DAGs can be time-consuming. Thorough testing and logging are essential for efficient debugging.

H6: The Future of Airflow in Data Engineering

The future of airflow in data engineering looks bright. Continuous development and community support ensure its continued relevance and improvement. We can expect further advancements in areas such as:

Improved Scalability: Enhanced scalability to handle even larger and more complex workflows.

Enhanced Integration: Improved integration with newer technologies and cloud platforms.

AI-powered Automation: Integration of AI and machine learning capabilities for automated workflow optimization and troubleshooting.

Conclusion:

Airflow in data engineering has emerged as a critical tool, empowering data engineers to build and manage robust, scalable, and maintainable data pipelines. Its DAG-based approach, extensibility, and monitoring capabilities have revolutionized the way data is processed and analyzed. While challenges exist, adopting best practices and leveraging the ongoing development of Airflow ensure its continued role in shaping the future of data engineering.

FAQs:

1. What are the system requirements for running Apache Airflow? Airflow's system requirements vary depending on the scale of your workflows. Consult the official documentation for detailed specifications.
2. How does Airflow handle task failures? Airflow offers retry mechanisms and alerting capabilities for handling task failures. The behavior can be customized based on specific needs.
3. Can Airflow be integrated with cloud platforms like AWS and Azure? Yes, Airflow offers excellent integration with major cloud platforms, leveraging their services for storage, compute, and other functionalities.
4. What programming languages are supported by Airflow? Primarily Python, although other languages can be integrated through custom operators.
5. How do I monitor the performance of my Airflow DAGs? Airflow provides a web UI for monitoring DAG execution, including visualization of

progress, logs, and performance metrics.

6. What are some popular Airflow operators? Popular operators include those for interacting with databases (e.g., `PostgresOperator`, `MySQLOperator`), cloud storage (e.g., `S3Operator`, `AzureBlobStorageOperator`), and big data frameworks (e.g., `SparkOperator`, `HadoopOperator`).
7. Is Airflow suitable for real-time data processing? While primarily designed for batch processing, Airflow can be integrated with real-time streaming technologies to manage parts of real-time data pipelines.
8. How secure is Airflow? Airflow's security can be enhanced through various methods, including access control, encryption, and integration with security tools. Proper configuration is essential.
9. What are the alternatives to Airflow? Other workflow management tools include Prefect, Dagster, and Luigi, each with its strengths and weaknesses.

Related Articles:

1. "Building Scalable ETL Pipelines with Apache Airflow": This article focuses on building and scaling ETL pipelines using Airflow, covering best practices and common challenges.
2. "Monitoring and Alerting in Apache Airflow": A deep dive into setting up effective monitoring and alerting systems for Airflow DAGs to ensure pipeline reliability.
3. "Integrating Airflow with Cloud Platforms": This article explores the integration of Airflow with AWS, Azure, and GCP, showcasing the benefits of cloud-based deployments.
4. "Airflow Best Practices for Data Engineering Teams": This article provides practical tips and best practices for managing and optimizing Airflow workflows in a team environment.
5. "Troubleshooting Common Airflow Errors": This article provides guidance on diagnosing and resolving common issues encountered when working with Airflow.
6. "Advanced Airflow Concepts: Branching, Dynamic DAGs, and Sensors": This article explores more advanced features of Airflow, including dynamic DAG generation and conditional branching.
7. "Comparing Airflow with Other Workflow Management Tools": A comparative analysis of Airflow against other popular workflow management platforms.

8. "Using Airflow for Machine Learning Pipelines": This article discusses how Airflow can be leveraged to build and manage complex machine learning workflows.
9. "Security Best Practices for Apache Airflow": A detailed guide on securing your Airflow deployments to protect sensitive data and prevent unauthorized access.

airflow in data engineering: Data Pipelines with Apache Airflow Bas P. Harenslak, Julian de Ruiter, 2021-04-27 This book 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. --

airflow in data engineering: Data Pipelines Pocket Reference James Densmore, 2021-02-10 Data pipelines are the foundation for success in data analytics. Moving data from numerous diverse sources and transforming it to provide context is the difference between having data and actually gaining value from it. This pocket reference defines data pipelines and explains how they work in today's modern data stack. You'll learn common considerations and key decision points when implementing pipelines, such as batch versus streaming data ingestion and build versus buy. This book addresses the most common decisions made by data professionals and discusses foundational concepts that apply to open source frameworks, commercial products, and homegrown solutions. You'll learn: What a data pipeline is and how it works How data is moved and processed on modern data infrastructure, including cloud platforms Common tools and products used by data engineers to build pipelines How pipelines support analytics and reporting needs Considerations for pipeline maintenance, testing, and alerting

airflow in data engineering: Data Engineering with Python Paul Crickard, 2020-10-23 Build, monitor, and manage real-time data pipelines to create data engineering infrastructure efficiently using open-source Apache projects Key Features Become well-versed in data architectures, data preparation, and data optimization skills with the help of practical examples Design data models and learn how to extract, transform, and load (ETL) data using Python Schedule, automate, and monitor complex data pipelines in production Book DescriptionData 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 learn Understand how data engineering supports data science workflows Discover how to extract data from files and databases and then clean, transform, and enrich it Configure processors for handling different file formats as well as both relational and NoSQL databases Find out how to implement a data pipeline and dashboard to visualize results Use staging and validation to check data before landing in the warehouse Build real-time pipelines with staging areas that perform validation and handle failures Get to grips with

deploying pipelines in the production environment Who 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.

airflow in data engineering: *Data Engineering with Google Cloud Platform* Adi Wijaya, 2022-03-31 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 Features Understand data engineering concepts, the role of a data engineer, and the benefits of using GCP for building your solution Learn how to use the various GCP products to ingest, consume, and transform data and orchestrate pipelines Discover tips to prepare for and pass the Professional Data Engineer exam Book 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 learn Load data into BigQuery and materialize its output for downstream consumption Build data pipeline orchestration using Cloud Composer Develop Airflow jobs to orchestrate and automate a data warehouse Build a Hadoop data lake, create ephemeral clusters, and run jobs on the Dataproc cluster Leverage Pub/Sub for messaging and ingestion for event-driven systems Use Dataflow to perform ETL on streaming data Unlock the power of your data with Data Studio Calculate the GCP cost estimation for your end-to-end data solutions Who 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.

airflow in data engineering: *Data Engineering with Apache Spark, Delta Lake, and Lakehouse* Manoj Kukreja, Danil Zburivsky, 2021-10-22 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 learnDiscover the challenges you may face in the data engineering worldAdd ACID transactions to Apache Spark using Delta LakeUnderstand effective design strategies to build enterprise-grade data lakesExplore architectural and design patterns for building efficient data ingestion pipelinesOrchestrate a data pipeline for preprocessing data using Apache Spark and Delta Lake APIsAutomate deployment and monitoring of data pipelines in productionGet to grips with securing, monitoring, and managing data pipelines models efficientlyWho 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.

airflow in data engineering: Data Engineering with AWS Gareth Eagar, 2023-10-31

Looking to revolutionize your data transformation game with AWS? Look no further! From strong foundations to hands-on building of data engineering pipelines, our expert-led manual has got you covered. Key Features Delve into robust AWS tools for ingesting, transforming, and consuming data, and for orchestrating pipelines Stay up to date with a comprehensive revised chapter on Data Governance Build modern data platforms with a new section covering transactional data lakes and data mesh Book DescriptionThis book, authored by a seasoned Senior Data Architect with 25 years of experience, aims to help you achieve proficiency in using the AWS ecosystem for data engineering. This revised edition provides updates in every chapter to cover the latest AWS services and features, takes a refreshed look at data governance, and includes a brand-new section on building modern data platforms which covers; implementing a data mesh approach, open-table formats (such as Apache Iceberg), and using DataOps for automation and observability. You'll begin by reviewing the key concepts and essential AWS tools in a data engineer's toolkit and getting acquainted with modern data management approaches. You'll then architect a data pipeline, review raw data sources, transform the data, and learn how that transformed data is used by various data consumers. You'll learn how to ensure strong data governance, and about populating data marts and data warehouses along with how a data lakehouse fits into the picture. After that, you'll be introduced to AWS tools for analyzing data, including those for ad-hoc SQL queries and creating visualizations. Then, you'll explore how the power of machine learning and artificial intelligence can be used to draw new insights from data. In the final chapters, you'll discover transactional data lakes, data meshes, and how to build a cutting-edge data platform on AWS. By the end of this AWS book, you'll be able to execute data engineering tasks and implement a data pipeline on AWS like a pro!What you will learn Seamlessly 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 Load data into a Redshift data warehouse and run queries with ease Visualize and explore data using Amazon QuickSight Extract sentiment data from a dataset using Amazon Comprehend Build transactional data lakes using Apache Iceberg with Amazon Athena Learn how a data mesh approach can be implemented on AWS Who this book is forThis 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.

airflow in data engineering: Fundamentals of Data Engineering Joe Reis, Matt Housley, 2022-06-22 Data engineering has grown rapidly in the past decade, leaving many software

engineers, data scientists, and analysts looking for a comprehensive view of this practice. With this practical book, you'll learn how to plan and build systems to serve the needs of your organization and customers by evaluating the best technologies available through the framework of the data engineering lifecycle. Authors Joe Reis and Matt Housley walk you through the data engineering lifecycle and show you how to stitch together a variety of cloud technologies to serve the needs of downstream data consumers. You'll understand how to apply the concepts of data generation, ingestion, orchestration, transformation, storage, and governance that are critical in any data environment regardless of the underlying technology. This book will help you: Get a concise overview of the entire data engineering landscape Assess data engineering problems using an end-to-end framework of best practices Cut through marketing hype when choosing data technologies, architecture, and processes Use the data engineering lifecycle to design and build a robust architecture Incorporate data governance and security across the data engineering lifecycle

airflow in data engineering: Data Science on AWS Chris Fregly, Antje Barth, 2021-04-07 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 up your 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

airflow in data engineering: Data Engineering for Machine Learning Pipelines Pavan Kumar Narayanan,

airflow in data engineering: *Official Google Cloud Certified Professional Data Engineer Study Guide* Dan Sullivan, 2020-05-11 The proven Study Guide that prepares you for this new Google Cloud exam The Google Cloud Certified Professional Data Engineer Study Guide, provides everything you need to prepare for this important exam and master the skills necessary to land that coveted Google Cloud Professional Data Engineer certification. Beginning with a pre-book assessment quiz to evaluate what you know before you begin, each chapter features exam objectives and review questions, plus the online learning environment includes additional complete practice tests. Written by Dan Sullivan, a popular and experienced online course author for machine learning, big data, and Cloud topics, Google Cloud Certified Professional Data Engineer Study Guide is your ace in the hole for deploying and managing analytics and machine learning applications. Build and operationalize storage systems, pipelines, and compute infrastructure Understand machine learning models and learn how to select pre-built models Monitor and troubleshoot machine learning models Design analytics and machine learning applications that are secure, scalable, and highly available. This exam guide is designed to help you develop an in depth understanding of data engineering and machine learning on Google Cloud Platform.

airflow in data engineering: **Building Machine Learning Pipelines** Hannes Hapke, Catherine Nelson, 2020-07-13 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

airflow in data engineering: Cracking the Data Engineering Interview Kedeisha Bryan, Taamir Ransome, 2023-11-07 Get to grips with the fundamental concepts of data engineering, and solve mock interview questions while building a strong resume and a personal brand to attract the right employers Key Features Develop your own brand, projects, and portfolio with expert help to stand out in the interview round Get a quick refresher on core data engineering topics, such as Python, SQL, ETL, and data modeling Practice with 50 mock questions on SQL, Python, and more to ace the behavioral and technical rounds Purchase of the print or Kindle book includes a free PDF eBook Book DescriptionPreparing for a data engineering interview can often get overwhelming due to the abundance of tools and technologies, leaving you struggling to prioritize which ones to focus on. This hands-on guide provides you with the essential foundational and advanced knowledge needed to simplify your learning journey. The book begins by helping you gain a clear understanding of the nature of data engineering and how it differs from organization to organization. As you progress through the chapters, you'll receive expert advice, practical tips, and real-world insights on everything from creating a resume and cover letter to networking and negotiating your salary. The chapters also offer refresher training on data engineering essentials, including data modeling, database architecture, ETL processes, data warehousing, cloud computing, big data, and machine learning. As you advance, you'll gain a holistic view by exploring continuous integration/continuous development (CI/CD), data security, and privacy. Finally, the book will help you practice case studies, mock interviews, as well as behavioral questions. By the end of this book, you will have a clear understanding of what is required to succeed in an interview for a data engineering role. What you will learn Create maintainable and scalable code for unit testing Understand the fundamental concepts of core data engineering tasks Prepare with over 100 behavioral and technical interview questions Discover data engineer archetypes and how they can help you prepare for the interview Apply the essential concepts of Python and SQL in data engineering Build your personal brand to noticeably stand out as a candidate Who this book is for If you're an aspiring data engineer looking for guidance on how to land, prepare for, and excel in data engineering interviews, this book is for you. Familiarity with the fundamentals of data engineering, such as data modeling, cloud warehouses, programming (python and SQL), building data pipelines, scheduling your workflows (Airflow), and APIs, is a prerequisite.

airflow in data engineering: Data Engineering with Scala and Spark Eric Tome, Rupam Bhattacharjee, David Radford, 2024-01-31 Take your data engineering skills to the next level by learning how to utilize Scala and functional programming to create continuous and scheduled pipelines that ingest, transform, and aggregate data Key Features Transform data into a clean and trusted source of information for your organization using Scala Build streaming and batch-processing pipelines with step-by-step explanations Implement and orchestrate your pipelines by following CI/CD best practices and test-driven development (TDD) Purchase of the print or Kindle book includes a free PDF eBook Book DescriptionMost data engineers know that performance issues in a distributed computing environment can easily lead to issues impacting the overall efficiency and effectiveness of data engineering tasks. While Python remains a popular choice for data engineering due to its ease of use, Scala shines in scenarios where the performance of distributed data processing is paramount. This book will teach you how to leverage the Scala programming language on the Spark framework and use the latest cloud technologies to build continuous and triggered data pipelines. You'll do this by setting up a data engineering environment for local development and

scalable distributed cloud deployments using data engineering best practices, test-driven development, and CI/CD. You'll also get to grips with DataFrame API, Dataset API, and Spark SQL API and its use. Data profiling and quality in Scala will also be covered, alongside techniques for orchestrating and performance tuning your end-to-end pipelines to deliver data to your end users. By the end of this book, you will be able to build streaming and batch data pipelines using Scala while following software engineering best practices. What you will learn Set up your development environment to build pipelines in Scala Get to grips with polymorphic functions, type parameterization, and Scala implicits Use Spark DataFrames, Datasets, and Spark SQL with Scala Read and write data to object stores Profile and clean your data using Deeplake Performance tune your data pipelines using Scala Who this book is for This book is for data engineers who have experience in working with data and want to understand how to transform raw data into a clean, trusted, and valuable source of information for their organization using Scala and the latest cloud technologies.

airflow in data engineering: Data Science in Production Ben Weber, 2020 Putting predictive models into production is one of the most direct ways that data scientists can add value to an organization. By learning how to build and deploy scalable model pipelines, data scientists can own more of the model production process and more rapidly deliver data products. This book provides a hands-on approach to scaling up Python code to work in distributed environments in order to build robust pipelines. Readers will learn how to set up machine learning models as web endpoints, serverless functions, and streaming pipelines using multiple cloud environments. It is intended for analytics practitioners with hands-on experience with Python libraries such as Pandas and scikit-learn, and will focus on scaling up prototype models to production. From startups to trillion dollar companies, data science is playing an important role in helping organizations maximize the value of their data. This book helps data scientists to level up their careers by taking ownership of data products with applied examples that demonstrate how to: Translate models developed on a laptop to scalable deployments in the cloud Develop end-to-end systems that automate data science workflows Own a data product from conception to production The accompanying Jupyter notebooks provide examples of scalable pipelines across multiple cloud environments, tools, and libraries (github.com/bgweber/DS_Production). Book Contents Here are the topics covered by Data Science in Production: Chapter 1: Introduction - This chapter will motivate the use of Python and discuss the discipline of applied data science, present the data sets, models, and cloud environments used throughout the book, and provide an overview of automated feature engineering. Chapter 2: Models as Web Endpoints - This chapter shows how to use web endpoints for consuming data and hosting machine learning models as endpoints using the Flask and Gunicorn libraries. We'll start with scikit-learn models and also set up a deep learning endpoint with Keras. Chapter 3: Models as Serverless Functions - This chapter will build upon the previous chapter and show how to set up model endpoints as serverless functions using AWS Lambda and GCP Cloud Functions. Chapter 4: Containers for Reproducible Models - This chapter will show how to use containers for deploying models with Docker. We'll also explore scaling up with ECS and Kubernetes, and building web applications with Plotly Dash. Chapter 5: Workflow Tools for Model Pipelines - This chapter focuses on scheduling automated workflows using Apache Airflow. We'll set up a model that pulls data from BigQuery, applies a model, and saves the results. Chapter 6: PySpark for Batch Modeling - This chapter will introduce readers to PySpark using the community edition of Databricks. We'll build a batch model pipeline that pulls data from a data lake, generates features, applies a model, and stores the results to a No SQL database. Chapter 7: Cloud Dataflow for Batch Modeling - This chapter will introduce the core components of Cloud Dataflow and implement a batch model pipeline for reading data from BigQuery, applying an ML model, and saving the results to Cloud Datastore. Chapter 8: Streaming Model Workflows - This chapter will introduce readers to Kafka and PubSub for streaming messages in a cloud environment. After working through this material, readers will learn how to use these message brokers to create streaming model pipelines with PySpark and Dataflow that provide near real-time predictions. Excerpts of these chapters are available on Medium (@bgweber), and a book sample is available on Leanpub.

airflow in data engineering: Modern Data Engineering with Apache Spark Scott Haines, 2022-03-23 Leverage Apache Spark within a modern data engineering ecosystem. This hands-on guide will teach you how to write fully functional applications, follow industry best practices, and learn the rationale behind these decisions. With Apache Spark as the foundation, you will follow a step-by-step journey beginning with the basics of data ingestion, processing, and transformation, and ending up with an entire local data platform running Apache Spark, Apache Zeppelin, Apache Kafka, Redis, MySQL, Minio (S3), and Apache Airflow. Apache Spark applications solve a wide range of data problems from traditional data loading and processing to rich SQL-based analysis as well as complex machine learning workloads and even near real-time processing of streaming data. Spark fits well as a central foundation for any data engineering workload. This book will teach you to write interactive Spark applications using Apache Zeppelin notebooks, write and compile reusable applications and modules, and fully test both batch and streaming. You will also learn to containerize your applications using Docker and run and deploy your Spark applications using a variety of tools such as Apache Airflow, Docker and Kubernetes. Reading this book will empower you to take advantage of Apache Spark to optimize your data pipelines and teach you to craft modular and testable Spark applications. You will create and deploy mission-critical streaming spark applications in a low-stress environment that paves the way for your own path to production. What You Will Learn Simplify data transformation with Spark Pipelines and Spark SQL Bridge data engineering with machine learning Architect modular data pipeline applications Build reusable application components and libraries Containerize your Spark applications for consistency and reliability Use Docker and Kubernetes to deploy your Spark applications Speed up application experimentation using Apache Zeppelin and Docker Understand serializable structured data and data contracts Harness effective strategies for optimizing data in your data lakes Build end-to-end Spark structured streaming applications using Redis and Apache Kafka Embrace testing for your batch and streaming applications Deploy and monitor your Spark applications Who This Book Is For Professional software engineers who want to take their current skills and apply them to new and exciting opportunities within the data ecosystem, practicing data engineers who are looking for a guiding light while traversing the many challenges of moving from batch to streaming modes, data architects who wish to provide clear and concise direction for how best to harness and use Apache Spark within their organization, and those interested in the ins and outs of becoming a modern data engineer in today's fast-paced and data-hungry world

airflow in data engineering: Learning Spark Jules S. Damji, Brooke Wenig, Tathagata Das, Denny Lee, 2020-07-16 Data is bigger, arrives faster, and comes in a variety of formats—and it all needs to be processed at scale for analytics or machine learning. But how can you process such varied workloads efficiently? Enter Apache Spark. Updated to include Spark 3.0, this second edition shows data engineers and data scientists why structure and unification in Spark matters. Specifically, this book explains how to perform simple and complex data analytics and employ machine learning algorithms. Through step-by-step walk-throughs, code snippets, and notebooks, you'll be able to: Learn Python, SQL, Scala, or Java high-level Structured APIs Understand Spark operations and SQL Engine Inspect, tune, and debug Spark operations with Spark configurations and Spark UI Connect to data sources: JSON, Parquet, CSV, Avro, ORC, Hive, S3, or Kafka Perform analytics on batch and streaming data using Structured Streaming Build reliable data pipelines with open source Delta Lake and Spark Develop machine learning pipelines with MLlib and productionize models using MLflow

airflow in data engineering: Natural Ventilation for Infection Control in Health-care Settings Y. Chartier, C. L Pessoa-Silva, 2009 This guideline defines ventilation and then natural ventilation. It explores the design requirements for natural ventilation in the context of infection control, describing the basic principles of design, construction, operation and maintenance for an effective natural ventilation system to control infection in health-care settings.

airflow in data engineering: Urban Climates T. R. Oke, G. Mills, A. Christen, J. A. Voogt, 2017-09-14 Urban Climates is the first full synthesis of modern scientific and applied research on

urban climates. The book begins with an outline of what constitutes an urban ecosystem. It develops a comprehensive terminology for the subject using scale and surface classification as key constructs. It explains the physical principles governing the creation of distinct urban climates, such as airflow around buildings, the heat island, precipitation modification and air pollution, and it then illustrates how this knowledge can be applied to moderate the undesirable consequences of urban development and help create more sustainable and resilient cities. With urban climate science now a fully-fledged field, this timely book fulfills the need to bring together the disparate parts of climate research on cities into a coherent framework. It is an ideal resource for students and researchers in fields such as climatology, urban hydrology, air quality, environmental engineering and urban design.

airflow in data engineering: Spark: The Definitive Guide Bill Chambers, Matei Zaharia, 2018-02-08 Learn how to use, deploy, and maintain Apache Spark with this comprehensive guide, written by the creators of the open-source cluster-computing framework. With an emphasis on improvements and new features in Spark 2.0, authors Bill Chambers and Matei Zaharia break down Spark topics into distinct sections, each with unique goals. 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

airflow in data engineering: Aeronautical Engineer's Data Book Cliff Matthews, 2001-10-17 Aeronautical Engineer's Data Book is an essential handy guide containing useful up to date information regularly needed by the student or practising engineer. Covering all aspects of aircraft, both fixed wing and rotary craft, this pocket book provides quick access to useful aeronautical engineering data and sources of information for further in-depth information. - Quick reference to essential data - Most up to date information available

airflow in data engineering: Data Engineering on Azure Vlad Riscutia, 2021-08-17 Build a data platform to the industry-leading standards set by Microsoft's own infrastructure. Summary In Data Engineering on Azure you will learn how to: Pick the right Azure services for different data scenarios Manage data inventory Implement production quality data modeling, analytics, and machine learning workloads Handle data governance Using DevOps to increase reliability Ingesting, storing, and distributing data Apply best practices for compliance and access control Data Engineering on Azure reveals the data management patterns and techniques that support Microsoft's own massive data infrastructure. Author Vlad Riscutia, a data engineer at Microsoft, teaches you to bring an engineering rigor to your data platform and ensure that your data prototypes function just as well under the pressures of production. You'll implement common data modeling patterns, stand up cloud-native data platforms on Azure, and get to grips with DevOps for both analytics and machine learning. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Build secure, stable data platforms that can scale to loads of any size. When a project moves from the lab into production, you need confidence that it can stand up to real-world challenges. This book teaches you to design and implement cloud-based data infrastructure that you can easily monitor, scale, and modify. About the book In Data Engineering on Azure you'll learn the skills you need to build and maintain big data platforms in massive enterprises. This invaluable guide includes clear, practical guidance for setting up infrastructure, orchestration, workloads, and governance. As you go, you'll set up efficient machine learning pipelines, and then master time-saving automation and DevOps solutions. The Azure-based examples are easy to reproduce on other cloud platforms. What's inside Data inventory

and data governance Assure data quality, compliance, and distribution Build automated pipelines to increase reliability Ingest, store, and distribute data Production-quality data modeling, analytics, and machine learning About the reader For data engineers familiar with cloud computing and DevOps. About the author Vlad Riscutia is a software architect at Microsoft. Table of Contents 1 Introduction PART 1 INFRASTRUCTURE 2 Storage 3 DevOps 4 Orchestration PART 2 WORKLOADS 5 Processing 6 Analytics 7 Machine learning PART 3 GOVERNANCE 8 Metadata 9 Data quality 10 Compliance 11 Distributing data

airflow in data engineering: Agile Data Science 2.0 Russell Journey, 2017-06-07 Data science teams looking to turn research into useful analytics applications require not only the right tools, but also the right approach if they're to succeed. With the revised second edition of this hands-on guide, up-and-coming data scientists will learn how to use the Agile Data Science development methodology to build data applications with Python, Apache Spark, Kafka, and other tools. Author Russell Journey demonstrates how to compose a data platform for building, deploying, and refining analytics applications with Apache Kafka, MongoDB, Elasticsearch, d3.js, scikit-learn, and Apache Airflow. You'll learn an iterative approach that lets you quickly change the kind of analysis you're doing, depending on what the data is telling you. Publish data science work as a web application, and affect meaningful change in your organization. Build value from your data in a series of agile sprints, using the data-value pyramid Extract features for statistical models from a single dataset Visualize data with charts, and expose different aspects through interactive reports Use historical data to predict the future via classification and regression Translate predictions into actions Get feedback from users after each sprint to keep your project on track

airflow in data engineering: Fundamentals of Analytics Engineering Dumky De Wilde, Fanny Kassapian, Jovan Gligorevic, Juan Manuel Perafan, Lasse Benninga, Ricardo Angel Granados Lopez, Taís Laurindo Pereira, 2024-03-29 Gain a holistic understanding of the analytics engineering lifecycle by integrating principles from both data analysis and engineering Key Features Discover how analytics engineering aligns with your organization's data strategy Access insights shared by a team of seven industry experts Tackle common analytics engineering problems faced by modern businesses Purchase of the print or Kindle book includes a free PDF eBook Book DescriptionWritten by a team of 7 industry experts, Fundamentals of Analytics Engineering will introduce you to everything from foundational concepts to advanced skills to get started as an analytics engineer. After conquering data ingestion and techniques for data quality and scalability, you'll learn about techniques such as data cleaning transformation, data modeling, SQL query optimization and reuse, and serving data across different platforms. Armed with this knowledge, you will implement a simple data platform from ingestion to visualization, using tools like Airbyte Cloud, Google BigQuery, dbt, and Tableau. You'll also get to grips with strategies for data integrity with a focus on data quality and observability, along with collaborative coding practices like version control with Git. You'll learn about advanced principles like CI/CD, automating workflows, gathering, scoping, and documenting business requirements, as well as data governance. By the end of this book, you'll be armed with the essential techniques and best practices for developing scalable analytics solutions from end to end. What you will learn Design and implement data pipelines from ingestion to serving data Explore best practices for data modeling and schema design Scale data processing with cloud based analytics platforms and tools Understand the principles of data quality management and data governance Streamline code base with best practices like collaborative coding, version control, reviews and standards Automate and orchestrate data pipelines Drive business adoption with effective scoping and prioritization of analytics use cases Who this book is for This book is for data engineers and data analysts considering pivoting their careers into analytics engineering. Analytics engineers who want to upskill and search for gaps in their knowledge will also find this book helpful, as will other data professionals who want to understand the value of analytics engineering in their organization's journey toward data maturity. To get the most out of this book, you should have a basic understanding of data analysis and engineering concepts such as data cleaning, visualization, ETL and data warehousing.

airflow in data engineering: Designing Cloud Data Platforms Danil Zburivsky, Lynda Partner, 2021-04-20 Centralized data warehouses, the long-time defacto standard for housing data for analytics, are rapidly giving way to multi-faceted cloud data platforms. Companies that embrace modern cloud data platforms benefit from an integrated view of their business using all of their data and can take advantage of advanced analytic practices to drive predictions and as yet unimagined data services. Designing Cloud Data Platforms is an hands-on guide to envisioning and designing a modern scalable data platform that takes full advantage of the flexibility of the cloud. As you read, you'll learn the core components of a cloud data platform design, along with the role of key technologies like Spark and Kafka Streams. You'll also explore setting up processes to manage cloud-based data, keep it secure, and using advanced analytic and BI tools to analyse it. about the technology Access to affordable, dependable, serverless cloud services has revolutionized the way organizations can approach data management, and companies both big and small are raring to migrate to the cloud. But without a properly designed data platform, data in the cloud can remain just as siloed and inaccessible as it is today for most organizations. Designing Cloud Data Platforms lays out the principles of a well-designed platform that uses the scalable resources of the public cloud to manage all of an organization's data, and present it as useful business insights. about the book In Designing Cloud Data Platforms, you'll learn how to integrate data from multiple sources into a single, cloud-based, modern data platform. Drawing on their real-world experiences designing cloud data platforms for dozens of organizations, cloud data experts Danil Zburivsky and Lynda Partner take you through a six-layer approach to creating cloud data platforms that maximizes flexibility and manageability and reduces costs. Starting with foundational principles, you'll learn how to get data into your platform from different databases, files, and APIs, the essential practices for organizing and processing that raw data, and how to best take advantage of the services offered by major cloud vendors. As you progress past the basics you'll take a deep dive into advanced topics to get the most out of your data platform, including real-time data management, machine learning analytics, schema management, and more. what's inside The tools of different public cloud for implementing data platforms Best practices for managing structured and unstructured data sets Machine learning tools that can be used on top of the cloud Cost optimization techniques about the reader For data professionals familiar with the basics of cloud computing and distributed data processing systems like Hadoop and Spark. about the authors Danil Zburivsky has over 10 years experience designing and supporting large-scale data infrastructure for enterprises across the globe. Lynda Partner is the VP of Analytics-as-a-Service at Pythian, and has been on the business side of data for over 20 years.

airflow in data engineering: Data Engineering with AWS Cookbook Trâm Ngọc Phạm, Gonzalo Herreros González, Viqar Khan, Huda Nofal, 2024-11-29 Master AWS data engineering services and techniques for orchestrating pipelines, building layers, and managing migrations Key Features Get up to speed with the different AWS technologies for data engineering Learn the different aspects and considerations of building data lakes, such as security, storage, and operations Get hands on with key AWS services such as Glue, EMR, Redshift, QuickSight, and Athena for practical learning Purchase of the print or Kindle book includes a free PDF eBook Book Description Performing data engineering with Amazon Web Services (AWS) combines AWS's scalable infrastructure with robust data processing tools, enabling efficient data pipelines and analytics workflows. This comprehensive guide to AWS data engineering will teach you all you need to know about data lake management, pipeline orchestration, and serving layer construction. Through clear explanations and hands-on exercises, you'll master essential AWS services such as Glue, EMR, Redshift, QuickSight, and Athena. Additionally, you'll explore various data platform topics such as data governance, data quality, DevOps, CI/CD, planning and performing data migration, and creating Infrastructure as Code. As you progress, you will gain insights into how to enrich your platform and use various AWS cloud services such as AWS EventBridge, AWS DataZone, and AWS SCT and DMS to solve data platform challenges. Each recipe in this book is tailored to a daily challenge that a data engineer team faces while building a cloud platform. By the end of this book,

you will be well-versed in AWS data engineering and have gained proficiency in key AWS services and data processing techniques. You will develop the necessary skills to tackle large-scale data challenges with confidence. What you will learn Define your centralized data lake solution, and secure and operate it at scale Identify the most suitable AWS solution for your specific needs Build data pipelines using multiple ETL technologies Discover how to handle data orchestration and governance Explore how to build a high-performing data serving layer Delve into DevOps and data quality best practices Migrate your data from on-premises to AWS Who this book is for If you're involved in designing, building, or overseeing data solutions on AWS, this book provides proven strategies for addressing challenges in large-scale data environments. Data engineers as well as big data professionals looking to enhance their understanding of AWS features for optimizing their workflow, even if they're new to the platform, will find value. Basic familiarity with AWS security (users and roles) and command shell is recommended.

airflow in data engineering: Data Engineering with Google Cloud Platform Adi Wijaya, 2024-04-30 Become a successful data engineer by building and deploying your own data pipelines on Google Cloud, including making key architectural decisions Key Features Get up to speed with data governance on Google Cloud Learn how to use various Google Cloud products like Dataform, DLP, Dataplex, Dataproc Serverless, and Datastream Boost your confidence by getting Google Cloud data engineering certification guidance from real exam experiences Purchase of the print or Kindle book includes a free PDF eBook Book DescriptionThe second edition of Data Engineering with Google Cloud builds upon the success of the first edition by offering enhanced clarity and depth to data professionals navigating the intricate landscape of data engineering. Beyond its foundational lessons, this new edition delves into the essential realm of data governance within Google Cloud, providing you with invaluable insights into managing and optimizing data resources effectively. Written by a Data Strategic Cloud Engineer at Google, this book helps you stay ahead of the curve by guiding you through the latest technological advancements in the Google Cloud ecosystem. You'll cover essential aspects, from exploring Cloud Composer 2 to the evolution of Airflow 2.5. Additionally, you'll explore how to work with cutting-edge tools like Dataform, DLP, Dataplex, Dataproc Serverless, and Datastream to perform data governance on datasets. By the end of this book, you'll be equipped to navigate the ever-evolving world of data engineering on Google Cloud, from foundational principles to cutting-edge practices. What you will learn Load data into BigQuery and materialize its output Focus on data pipeline orchestration using Cloud Composer Formulate Airflow jobs to orchestrate and automate a data warehouse Establish a Hadoop data lake, generate ephemeral clusters, and execute jobs on the Dataproc cluster Harness Pub/Sub for messaging and ingestion for event-driven systems Apply Dataflow to conduct ETL on streaming data Implement data governance services on Google Cloud Who this book is for Data analysts, IT practitioners, software engineers, or any data enthusiasts looking to have a successful data engineering career will find this book invaluable. Additionally, experienced data professionals who want to start using Google Cloud to build data platforms will get clear insights on how to navigate the path. Whether you're a beginner who wants to explore the fundamentals or a seasoned professional seeking to learn the latest data engineering concepts, this book is for you.

airflow in data engineering: Data Science on AWS Chris Fregly, Antje Barth, 2021-04-07 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 up your 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

airflow in data engineering: Google BigQuery: The Definitive Guide Valliappa

Lakshmanan, Jordan Tigani, 2019-10-23 Work with petabyte-scale datasets while building a collaborative, agile workplace in the process. This practical book is the canonical reference to Google BigQuery, the query engine that lets you conduct interactive analysis of large datasets. BigQuery enables enterprises to efficiently store, query, ingest, and learn from their data in a convenient framework. With this book, you'll examine how to analyze data at scale to derive insights from large datasets efficiently. Valliappa Lakshmanan, tech lead for Google Cloud Platform, and Jordan Tigani, engineering director for the BigQuery team, provide best practices for modern data warehousing within an autoscaled, serverless public cloud. Whether you want to explore parts of BigQuery you're not familiar with or prefer to focus on specific tasks, this reference is indispensable.

airflow in data engineering: Machine Learning with PySpark Pramod Singh, 2018-12-14

Build machine learning models, natural language processing applications, and recommender systems with PySpark to solve various business challenges. This book starts with the fundamentals of Spark and its evolution and then covers the entire spectrum of traditional machine learning algorithms along with natural language processing and recommender systems using PySpark. Machine Learning with PySpark shows you how to build supervised machine learning models such as linear regression, logistic regression, decision trees, and random forest. You'll also see unsupervised machine learning models such as K-means and hierarchical clustering. A major portion of the book focuses on feature engineering to create useful features with PySpark to train the machine learning models. The natural language processing section covers text processing, text mining, and embedding for classification. After reading this book, you will understand how to use PySpark's machine learning library to build and train various machine learning models. Additionally you'll become comfortable with related PySpark components, such as data ingestion, data processing, and data analysis, that you can use to develop data-driven intelligent applications. What You Will Learn Build a spectrum of supervised and unsupervised machine learning algorithms Implement machine learning algorithms with Spark MLlib libraries Develop a recommender system with Spark MLlib libraries Handle issues related to feature engineering, class balance, bias and variance, and cross validation for building an optimal fit model Who This Book Is For Data science and machine learning professionals.

airflow in data engineering: Agile Data Science 2.0 Russell Journey, 2017-06-07 Data science teams looking to turn research into useful analytics applications require not only the right tools, but also the right approach if they're to succeed. With the revised second edition of this hands-on guide, up-and-coming data scientists will learn how to use the Agile Data Science development methodology to build data applications with Python, Apache Spark, Kafka, and other tools. Author Russell Journey demonstrates how to compose a data platform for building, deploying, and refining analytics applications with Apache Kafka, MongoDB, Elasticsearch, d3.js, scikit-learn, and Apache Airflow. You'll learn an iterative approach that lets you quickly change the kind of analysis you're doing, depending on what the data is telling you. Publish data science work as a web application, and affect meaningful change in your organization. Build value from your data in a series of agile sprints, using the data-value pyramid Extract features for statistical models from a single dataset Visualize data with charts, and expose different aspects through interactive reports Use historical data to predict the future via classification and regression Translate predictions into actions Get feedback from users after each sprint to keep your project on track

airflow in data engineering: Transformers for Natural Language Processing Denis

Rothman, 2021-01-29 Publisher's Note: A new edition of this book is out now that includes working with GPT-3 and comparing the results with other models. It includes even more use cases, such as casual language analysis and computer vision tasks, as well as an introduction to OpenAI's Codex.

Key Features Build and implement state-of-the-art language models, such as the original Transformer, BERT, T5, and GPT-2, using concepts that outperform classical deep learning models Go through hands-on applications in Python using Google Colaboratory Notebooks with nothing to install on a local machine Test transformer models on advanced use cases

Book Description The transformer architecture has proved to be revolutionary in outperforming the classical RNN and CNN models in use today. With an apply-as-you-learn approach, Transformers for Natural Language Processing investigates in vast detail the deep learning for machine translations, speech-to-text, text-to-speech, language modeling, question answering, and many more NLP domains with transformers. The book takes you through NLP with Python and examines various eminent models and datasets within the transformer architecture created by pioneers such as Google, Facebook, Microsoft, OpenAI, and Hugging Face. The book trains you in three stages. The first stage introduces you to transformer architectures, starting with the original transformer, before moving on to RoBERTa, BERT, and DistilBERT models. You will discover training methods for smaller transformers that can outperform GPT-3 in some cases. In the second stage, you will apply transformers for Natural Language Understanding (NLU) and Natural Language Generation (NLG). Finally, the third stage will help you grasp advanced language understanding techniques such as optimizing social network datasets and fake news identification. By the end of this NLP book, you will understand transformers from a cognitive science perspective and be proficient in applying pretrained transformer models by tech giants to various datasets. What you will learn Use the latest pretrained transformer models Grasp the workings of the original Transformer, GPT-2, BERT, T5, and other transformer models Create language understanding Python programs using concepts that outperform classical deep learning models Use a variety of NLP platforms, including Hugging Face, Trax, and AllenNLP Apply Python, TensorFlow, and Keras programs to sentiment analysis, text summarization, speech recognition, machine translations, and more Measure the productivity of key transformers to define their scope, potential, and limits in production Who this book is for Since the book does not teach basic programming, you must be familiar with neural networks, Python, PyTorch, and TensorFlow in order to learn their implementation with Transformers. Readers who can benefit the most from this book include experienced deep learning & NLP practitioners and data analysts & data scientists who want to process the increasing amounts of language-driven data.

airflow in data engineering: Energy Efficient Thermal Management of Data Centers

Yogendra Joshi, Pramod Kumar, 2012-03-20 Energy Efficient Thermal Management of Data Centers examines energy flow in today's data centers. Particular focus is given to the state-of-the-art thermal management and thermal design approaches now being implemented across the multiple length scales involved. The impact of future trends in information technology hardware, and emerging software paradigms such as cloud computing and virtualization, on thermal management are also addressed. The book explores computational and experimental characterization approaches for determining temperature and air flow patterns within data centers. Thermodynamic analyses using the second law to improve energy efficiency are introduced and used in proposing improvements in cooling methodologies. Reduced-order modeling and robust multi-objective design of next generation data centers are discussed.

airflow in data engineering: Data Analysis with Python and PySpark Jonathan Rioux,

2022-03-22 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.

airflow in data engineering: Analytics Engineering with SQL and Dbt Rui Pedro Machado, Helder Russa, 2023-12-08 With the shift from data warehouses to data lakes, data now lands in repositories before it's been transformed, enabling engineers to model raw data into clean, well-defined datasets. dbt (data build tool) helps you take data further. This practical book shows data analysts, data engineers, BI developers, and data scientists how to create a true self-service transformation platform through the use of dynamic SQL. Authors Rui Machado from Monstarlab and Hélder Russa from Jumia show you how to quickly deliver new data products by focusing more on value delivery and less on architectural and engineering aspects. If you know your business well and have the technical skills to model raw data into clean, well-defined datasets, you'll learn how to design and deliver data models without any technical influence. With this book, you'll learn: What dbt is and how a dbt project is structured How dbt fits into the data engineering and analytics worlds How to collaborate on building data models The main tools and architectures for building useful, functional data models How to fit dbt into data warehousing and laking architecture How to build tests for data transformations

airflow in data engineering: Terraform in Action Scott Winkler, 2021-08-24 An outstanding source of knowledge for Terraform enthusiasts of all levels. - Anton Babenko, Betajob Terraform in Action shows you how to automate and scale infrastructure programmatically using the Terraform toolkit. Summary In Terraform in Action you will learn: Cloud architecture with Terraform Terraform module sharing and the private module registry Terraform security in a multitenant environment Strategies for performing blue/green deployments Refactoring for code maintenance and reusability Running Terraform at scale Creating your own Terraform provider Using Terraform as a continuous development/continuous delivery platform Terraform in Action introduces the infrastructure-as-code (IaC) model that lets you instantaneously create new components and respond efficiently to changes in demand. You'll use the Terraform automation tool to design and manage servers that can be provisioned, shared, changed, tested, and deployed with a single command. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Provision, deploy, scale, and clone your entire stack to the cloud at the touch of a button. In Terraform, you create a collection of simple declarative scripts that define and manage application infrastructure. This powerful infrastructure-as-code approach automates key tasks like versioning and testing for everything from low-level networking to cloud services. About the book Terraform in Action shows you how to automate and scale infrastructure programmatically using the Terraform toolkit. Using practical, relevant examples, you'll use Terraform to provision a Kubernetes cluster, deploy a multiplayer game, and configure other hands-on projects. As you progress to advanced techniques like zero-downtime deployments, you'll discover how to think in Terraform rather than just copying and pasting scripts. What's inside Cloud architecture with Terraform Terraform module sharing and the private module registry Terraform security in a multitenant environment Strategies for performing blue/green deployments About the reader For readers experienced with a major cloud platform such as AWS. Examples in JavaScript and Golang. About the author Scott Winkler is a DevOps engineer and a distinguished Terraform expert. He has spoken multiple times at HashiTalks and HashiConf, and was selected as a HashiCorp Ambassador and Core Contributor in 2020. Table of Contents PART 1 TERRAFORM BOOTCAMP 1 Getting started with Terraform 2 Life cycle of a Terraform resource 3 Functional programming 4 Deploying a multi-tiered web application in AWS PART 2 TERRAFORM IN THE WILD 5 Serverless made easy 6 Terraform

with friends 7 CI/CD pipelines as code 8 A multi-cloud MMORPG PART 3 MASTERING TERRAFORM
9 Zero-downtime deployments 10 Testing and refactoring 11 Extending Terraform by writing a
custom provider 12 Automating Terraform 13 Security and secrets management

airflow in data engineering: Data Analytics with Spark Using Python Jeffrey Aven, 2018-06-18
Solve Data Analytics Problems with Spark, PySpark, and Related Open Source Tools Spark is at the heart of today's Big Data revolution, helping data professionals supercharge efficiency and performance in a wide range of data processing and analytics tasks. In this guide, Big Data expert Jeffrey Aven covers all you need to know to leverage Spark, together with its extensions, subprojects, and wider ecosystem. Aven combines a language-agnostic introduction to foundational Spark concepts with extensive programming examples utilizing the popular and intuitive PySpark development environment. This guide's focus on Python makes it widely accessible to large audiences of data professionals, analysts, and developers—even those with little Hadoop or Spark experience. Aven's broad coverage ranges from basic to advanced Spark programming, and Spark SQL to machine learning. You'll learn how to efficiently manage all forms of data with Spark: streaming, structured, semi-structured, and unstructured. Throughout, concise topic overviews quickly get you up to speed, and extensive hands-on exercises prepare you to solve real problems. Coverage includes: • Understand Spark's evolving role in the Big Data and Hadoop ecosystems • Create Spark clusters using various deployment modes • Control and optimize the operation of Spark clusters and applications • Master Spark Core RDD API programming techniques • Extend, accelerate, and optimize Spark routines with advanced API platform constructs, including shared variables, RDD storage, and partitioning • Efficiently integrate Spark with both SQL and nonrelational data stores • Perform stream processing and messaging with Spark Streaming and Apache Kafka • Implement predictive modeling with SparkR and Spark MLlib

airflow in data engineering: Automated Machine Learning on AWS Trenton Potgieter, Jonathan Dahlberg, 2022-04-15 Automate the process of building, training, and deploying machine learning applications to production with AWS solutions such as SageMaker Autopilot, AutoGluon, Step Functions, Amazon Managed Workflows for Apache Airflow, and more Key FeaturesExplore the various AWS services that make automated machine learning easierRecognize the role of DevOps and MLOps methodologies in pipeline automationGet acquainted with additional AWS services such as Step Functions, MWAA, and more to overcome automation challengesBook Description AWS provides a wide range of solutions to help automate a machine learning workflow with just a few lines of code. With this practical book, you'll learn how to automate a machine learning pipeline using the various AWS services. Automated Machine Learning on AWS begins with a quick overview of what the machine learning pipeline/process looks like and highlights the typical challenges that you may face when building a pipeline. Throughout the book, you'll become well versed with various AWS solutions such as Amazon SageMaker Autopilot, AutoGluon, and AWS Step Functions to automate an end-to-end ML process with the help of hands-on examples. The book will show you how to build, monitor, and execute a CI/CD pipeline for the ML process and how the various CI/CD services within AWS can be applied to a use case with the Cloud Development Kit (CDK). You'll understand what a data-centric ML process is by working with the Amazon Managed Services for Apache Airflow and then build a managed Airflow environment. You'll also cover the key success criteria for an MLSDLC implementation and the process of creating a self-mutating CI/CD pipeline using AWS CDK from the perspective of the platform engineering team. By the end of this AWS book, you'll be able to effectively automate a complete machine learning pipeline and deploy it to production. What you will learnEmploy SageMaker Autopilot and Amazon SageMaker SDK to automate the machine learning processUnderstand how to use AutoGluon to automate complicated model building tasksUse the AWS CDK to codify the machine learning processCreate, deploy, and rebuild a CI/CD pipeline on AWSBuild an ML workflow using AWS Step Functions and the Data Science SDKLeverage the Amazon SageMaker Feature Store to automate the machine learning software development life cycle (MLSDLC)Discover how to use Amazon MWAA for a data-centric ML processWho this book is for This book is for the novice as well as experienced machine learning

practitioners looking to automate the process of building, training, and deploying machine learning-based solutions into production, using both purpose-built and other AWS services. A basic understanding of the end-to-end machine learning process and concepts, Python programming, and AWS is necessary to make the most out of this book.

airflow in data engineering: Building Big Data Pipelines with Apache Beam Jan Lukavsky, 2022-01-21 Implement, run, operate, and test data processing pipelines using Apache Beam Key Features Understand how to improve usability and productivity when implementing Beam pipelines Learn how to use stateful processing to implement complex use cases using Apache Beam Implement, test, and run Apache Beam pipelines with the help of expert tips and techniques Book 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 learn Understand the core concepts and architecture of Apache Beam Implement stateless and stateful data processing pipelines Use state and timers for processing real-time event processing Structure your code for reusability Use streaming SQL to process real-time data for increasing productivity and data accessibility Run a pipeline using a portable runner and implement data processing using the Apache Beam Python SDK Implement Apache Beam I/O connectors using the Splittable DoFn API Who 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.

airflow in data engineering: I Heart Logs Jay Kreps, 2014-09-23 Why a book about logs? That's easy: the humble log is an abstraction that lies at the heart of many systems, from NoSQL databases to cryptocurrencies. Even though most engineers don't think much about them, this short book shows you why logs are worthy of your attention. Based on his popular blog posts, LinkedIn principal engineer Jay Kreps shows you how logs work in distributed systems, and then delivers practical applications of these concepts in a variety of common uses—data integration, enterprise architecture, real-time stream processing, data system design, and abstract computing models. Go ahead and take the plunge with logs; you're going to love them. Learn how logs are used for programmatic access in databases and distributed systems Discover solutions to the huge data integration problem when more data of more varieties meet more systems Understand why logs are at the heart of real-time stream processing Learn the role of a log in the internals of online data systems Explore how Jay Kreps applies these ideas to his own work on data infrastructure systems at LinkedIn

airflow in data engineering: *Business Intelligence Career Master Plan* Eduardo Chavez, Danny Moncada, 2023-08-31 Learn the foundations of business intelligence, sector trade-offs, organizational structures, and technology stacks while mastering coursework, certifications, and interview success strategies Purchase of the print or Kindle book includes a free PDF eBook Key Features Identify promising job opportunities and ideal entry point into BI Build, design, implement, and maintain BI systems successfully Ace your BI interview with author's expert guidance on certifications, trainings, and courses Book Description Navigating the challenging path of a business intelligence career requires you to consider your expertise, interests, and skills. Business Intelligence Career Master Plan explores key skills like stacks, coursework, certifications, and interview advice, enabling you to make informed decisions about your BI journey. You'll start by assessing the different roles in BI and matching your skills and career with the tech stack. You'll

then learn to build taxonomy and a data story using visualization types. Additionally, you'll explore the fundamentals of programming, frontend development, backend development, software development lifecycle, and project management, giving you a broad view of the end-to-end BI process. With the help of the author's expert advice, you'll be able to identify what subjects and areas of study are crucial and would add significant value to your skill set. By the end of this book, you'll be well-equipped to make an informed decision on which of the myriad paths to choose in your business intelligence journey based on your skill set and interests. What you will learn Understand BI roles, roadmap, and technology stack Accelerate your career and land your first job in the BI industry Build the taxonomy of various data sources for your organization Use the AdventureWorks database and PowerBI to build a robust data model Create compelling data stories using data visualization Automate, templatize, standardize, and monitor systems for productivity Who this book is for This book is for BI developers and business analysts who are passionate about data and are looking to advance their proficiency and career in business intelligence. While foundational knowledge of tools like Microsoft Excel is required, having a working knowledge of SQL, Python, Tableau, and major cloud providers such as AWS or GCP will be beneficial.

Additional Resources

How to create a conditional task in Airflow - Stack Overflow

Apr 28, 2017 · Airflow 1.x. Airflow has a BranchPythonOperator that can be used to express the branching dependency more directly. The docs describe its use: The BranchPythonOperator is ...

First time login to Apache Airflow asks for username and ...

Feb 12, 2021 · user: airflow password: airflow If you want to create another account with docker use this (you have to be in the same folder of the docker-compose.yaml file): docker-compose ...

python - How to control the parallelism or concurrency of an ...

May 30, 2019 · Options that are specified across an entire Airflow setup: core.parallelism: maximum number of tasks running across an entire Airflow installation; core.dag_concurrency: ...

openssl - How to enable SSL on Apache Airflow? - Stack Overflow

Dec 19, 2017 · Go to AIRFLOW_HOME -> airflow.cfg. It has section named [webserver], under that there are two config properties like below: web_server_ssl_cert = web_server_ssl_key = if ...

Proper way to create dynamic workflows in Airflow

Jan 7, 2017 · Problem Is there any way in Airflow to create a workflow such that the number of tasks B.* is unknown until completion of Task A? I have looked at subdags but it looks like it ...

Python script scheduling in airflow - Stack Overflow

Jan 19, 2017 · In my airflow configuration dags_folder is specified as /home/amit/airflow/dags so I have placed the DAG definition there only.. is that right ? – Abhishek Pansotra Commented ...

How to install packages in Airflow (docker-compose)?

Jun 8, 2021 · 1. Create new Airflow docker image with installed Python

requirements. Check what Airflow image your docker-compose.yaml is using and use that image, in my case it's: ...

[python - How to branch multiple paths in Airflow DAG using ...](#)

May 6, 2021 · This is what I want, but I don't know how to achieve this in airflow, as both of the tasks are being executed. To summarize: T1 executes; T2 executes; Based on the output of ...

[How to use apache airflow in a virtual environment?](#)

Jul 4, 2019 · Edit the airflow.cfg file and set: load_examples = False
dags_folder = /path/to/your/dag/files If your airflow directory is not set to the default, you should set this env ...

Airflow: How to SSH and run BashOperator from a different server

If anyone is using 1.10 then new import should be from
airflow.contrib.hooks.ssh_hook import SSHHook and from
airflow.contrib.operators.ssh_operator import SSHOperator. – Biranjan ...

[How to create a conditional task in Airflow - Stack Overflow](#)

Apr 28, 2017 · Airflow 1.x. Airflow has a BranchPythonOperator that can be used to express the branching dependency more directly. The docs describe its use: The BranchPythonOperator is ...

First time login to Apache Airflow asks for username and ...

Feb 12, 2021 · user: airflow password: airflow If you want to create another account with docker use this (you have to be in the same folder of the docker-compose.yaml file): docker-compose ...

[python - How to control the parallelism or concurrency of an ...](#)

May 30, 2019 · Options that are specified across an entire Airflow setup:
core.parallelism: maximum number of tasks running across an entire Airflow installation; core.dag_concurrency: ...

[openssl - How to enable SSL on Apache Airflow? - Stack Overflow](#)

Dec 19, 2017 · Go to AIRFLOW_HOME -> airflow.cfg. It has section named [webserver], under that there are two config properties like below:
web_server_ssl_cert = web_server_ssl_key = if ...

Proper way to create dynamic workflows in Airflow

Jan 7, 2017 · Problem Is there any way in Airflow to create a workflow such that the number of tasks B.* is unknown until completion of Task A? I have looked at subdags but it looks like it ...

Python script scheduling in airflow - Stack Overflow

Jan 19, 2017 · In my airflow configuration dags_folder is specified as /home/amit/airflow/dags so I have placed the DAG definition there only.. is that right ? – Abhishek Pansotra Commented ...

How to install packages in Airflow (docker-compose)?

Jun 8, 2021 · 1. Create new Airflow docker image with installed Python requirements. Check what Airflow image your docker-compose.yaml is using and

use that image, in my case it's: ...

python - How to branch multiple paths in Airflow DAG using ...

May 6, 2021 · This is what I want, but I don't know how to achieve this in airflow, as both of the tasks are being executed. To summarize: T1 executes; T2 executes; Based on the output of T2 ...

How to use apache airflow in a virtual environment?

Jul 4, 2019 · Edit the airflow.cfg file and set: load_examples = False
dags_folder = /path/to/your/dag/files If your airflow directory is not set to the default, you should set this env ...

Airflow: How to SSH and run BashOperator from a different server

If anyone is using 1.10 then new import should be from
airflow.contrib.hooks.ssh_hook import SSHHook and from
airflow.contrib.operators.ssh_operator import SSHOperator. – Biranjan ...

Airflow In Data Engineering

Airflow In Data Engineering

Related Articles

- [al roker interview with biden](#)
- [air force rotc field training 2023 dates](#)
- [airpods max user guide](#)

[Back to Home](#)