Big Data on Kubernetes : A Practical Guide to Building Efficient and Scalable Data Solutions /

Crepalde, Neylson,
author

Monografía

Gain hands-on experience in building efficient and scalable big data architecture on Kubernetes, utilizing leading technologies such as Spark, Airflow, Kafka, and Trino Key Features Leverage Kubernetes in a cloud environment to integrate seamlessly with a variety of tools Explore best practices for optimizing the performance of big data pipelines Build end-to-end data pipelines and discover real-world use cases using popular tools like Spark, Airflow, and Kafka Purchase of the print or Kindle book includes a free PDF eBook Book Description In today's data-driven world, organizations across different sectors need scalable and efficient solutions for processing large volumes of data. Kubernetes offers an open-source and cost-effective platform for deploying and managing big data tools and workloads, ensuring optimal resource utilization and minimizing operational overhead. If you want to master the art of building and deploying big data solutions using Kubernetes, then this book is for you. Written by an experienced data specialist, Big Data on Kubernetes takes you through the entire process of developing scalable and resilient data pipelines, with a focus on practical implementation. Starting with the basics, you'll progress toward learning how to install Docker and run your first containerized applications. You'll then explore Kubernetes architecture and understand its core components. This knowledge will pave the way for exploring a variety of essential tools for big data processing such as Apache Spark and Apache Airflow. You'll also learn how to install and configure these tools on Kubernetes clusters. Throughout the book, you'll gain hands-on experience building a complete big data stack on Kubernetes. By the end of this Kubernetes book, you'll be equipped with the skills and knowledge you need to tackle real-world big data challenges with confidence. What you will learn Install and use Docker to run containers and build concise images Gain a deep understanding of Kubernetes architecture and its components Deploy and manage Kubernetes clusters on different cloud platforms Implement and manage data pipelines using Apache Spark and Apache Airflow Deploy and configure Apache Kafka for real-time data ingestion and processing Build and orchestrate a complete big data pipeline using open-source tools Deploy Generative AI applications on a Kubernetes-based architecture Who this book is for If you're a data engineer, BI analyst, data team leader, data architect, or tech manager with a basic understanding of big data technologies, then this big data book is for you. Familiarity with the basics of Python programming, SQL queries, and YAML is required to understand the topics discussed in this book

Título: Big Data on Kubernetes A Practical Guide to Building Efficient and Scalable Data Solutions Neylson Crepalde

Edición: First edition

Editorial: Birmingham, England Packt Publishing [2024] 2024

Descripción física: 1 online resource (297 pages)

Contenido: Cover -- Title page -- Copyright and credits -- Dedication -- Contributors -- Table of Contents -- Preface -- Part 1: Docker and Kubernetes -- Chapter 1: Getting Started with Containers -- Technical requirements -- Container architecture -- Installing Docker -- Windows -- macOS -- Linux -- Getting started with Docker images -- hello-world -- NGINX -- Julia -- Building your own image -- Batch processing job -- API service -- Summary -- Chapter 2: Kubernetes Architecture -- Technical requirements -- Kubernetes architecture -- Control plane -- Node components -- Pods -- Deployments -- StatefulSets -- Jobs -- Services -- ClusterIP Service -- NodePort Service -- LoadBalancer Service -- Ingress and Ingress Controller -- Gateway -- Persistent Volumes -- StorageClasses -- ConfigMaps and Secrets -- ConfigMaps -- Secrets -- Summary -- Chapter 3: Getting Hands-On with Kubernetes -- Technical requirements -- Installing kubectl -- Deploying a local cluster using Kind -- Installing kind -- Deploying the cluster -- Deploying an AWS EKS cluster -- Deploying a Google Cloud GKE cluster -- Deploying an Azure AKS cluster -- Running your API on Kubernetes -- Creating the deployment -- Creating a service -- Using an ingress to access the API -- Running a data processing job in Kubernetes -- Summary -- Part 2: Big Data Stack -- Chapter 4: The Modern Data Stack -- Data architectures -- The Lambda architecture -- The Kappa architecture -- Comparing Lambda and Kappa -- Data lake design for big data -- Data warehouses -- The rise of big data and data lakes -- The rise of the data lakehouse -- Implementing the lakehouse architecture -- Batch ingestion -- Storage -- Batch processing -- Orchestration -- Batch serving -- Data visualization -- Real-time ingestion -- Real-time processing -- Real-time serving -- Real-time data visualization -- Summary Chapter 5: Big Data Processing with Apache Spark -- Technical requirements -- Getting started with Spark -- Installing Spark locally -- Spark architecture -- Spark executors -- Components of execution -- Starting a Spark program -- The DataFrame API and the Spark SQL API -- Transformations -- Actions -- Lazy evaluation -- Data partitioning -- Narrow versus wide transformations -- Analyzing the titanic dataset -- Working with real data -- How Spark performs joins -- Joining IMDb tables -- Summary -- Chapter 6: Building Pipelines with Apache Airflow -- Technical requirements -- Getting started with Airflow -- Installing Airflow with Astro -- Airflow architecture -- Airflow's distributed architecture -- Building a data pipeline -- Airflow integration with other tools -- Summary -- Chapter 7: Apache Kafka for Real-Time Events and Data Ingestion -- Technical requirements -- Getting started with Kafka -- Exploring the Kafka architecture -- The PubSub design -- How Kafka delivers exactly-once semantics -- First producer and consumer -- Streaming from a database with Kafka Connect -- Real-time data processing with Kafka and Spark -- Summary -- Part 3: Connecting It All Together -- Chapter 8: Deploying the Big Data Stack on Kubernetes -- Technical requirements -- Deploying Spark on Kubernetes -- Deploying Airflow on Kubernetes -- Deploying Kafka on Kubernetes -- Summary -- Chapter 9: Data Consumption Layer -- Technical requirements -- Getting started with SQL query engines -- The limitations of traditional data warehouses -- The rise of SQL query engines -- The architecture of SQL query engines -- Deploying Trino in Kubernetes -- Connecting DBeaver with Trino -- Deploying Elasticsearch in Kubernetes -- How Elasticsearch stores, indexes and manages data -- Elasticsearch deployment -- Summary -- Chapter 10: Building a Big Data Pipeline on Kubernetes Technical requirements -- Checking the deployed tools -- Building a batch pipeline -- Building the Airflow DAG -- Creating SparkApplication jobs -- Creating a Glue crawler -- Building a real-time pipeline -- Deploying Kafka Connect and Elasticsearch -- Real-time processing with Spark -- Deploying the Elasticsearch sink connector -- Summary -- Chapter 11: Generative AI on Kubernetes -- Technical requirements -- What generative AI is and what it is not -- The power of large neural networks -- Challenges and limitations -- Using Amazon Bedrock to work with foundational models -- Building a generative AI application on Kubernetes -- Deploying the Streamlit app -- Building RAG with Knowledge Bases for Amazon Bedrock -- Adjusting the code for RAG retrieval -- Building action models with agents -- Creating a DynamoDB table -- Configuring the agent -- Deploying the application on Kubernetes -- Summary -- Chapter 12: Where to Go from Here -- Important topics for big data in Kubernetes -- Kubernetes monitoring and application monitoring -- Building a service mesh -- Security considerations --

Automated scalability -- GitOps and CI/CD for Kubernetes -- Kubernetes cost control -- What about team skills? -- Key skills for monitoring -- Building a service mesh -- Security considerations -- Automated scalability -- Skills for GitOps and CI/CD -- Cost control skills -- Summary -- Index -- Other Books You May Enjoy

ISBN: 9781835468999 electronic bk.) 9781835462140

Materia Título preferido: Kubernetes

Materia: Application software- Development Application program interfaces (Computer software) Datos masivos

Enlace a formato físico adicional: Print version Crepalde, Neylson. Big Data on Kubernetes Birmingham : Packt Publishing, Limited,c2024 9781835462140

Baratz Innovación Documental

- Gran Vía, 59 28013 Madrid
- (+34) 91 456 03 60
- informa@baratz.es