Building Data Streaming Applications With Apache Kafka Design Develop And Streamline Applications Using Apache Kafka Storm Heron And Spark

Thank you very much for downloading **Building Data Streaming Applications With Apache Kafka Design Develop And Streamline Applications Using Apache Kafka Storm Heron And Spark** .Maybe you have knowledge that, people have look numerous times for their favorite books in the manner of this Building Data Streaming Applications With Apache Kafka Design Develop And Streamline Applications Using Apache Kafka Storm Heron And Spark , but end happening in harmful downloads.

Rather than enjoying a good book with a mug of coffee in the afternoon, on the other hand they juggled following some harmful virus inside their computer. **Building Data Streaming Applications With Apache Kafka Design Develop And Streamline Applications Using Apache Kafka Storm Heron And Spark** is open in our digital library an online entrance to it is set as public in view of that you can download it instantly. Our digital library saves in fused countries, allowing you to acquire the most less latency era to download any of our books bearing in mind this one. Merely said, the Building Data Streaming Applications With Apache Kafka Design Develop And Streamline

Applications Using Apache Kafka Storm Heron And Spark is universally compatible subsequent to any devices to read.

Big Data SMACK - Raul Estrada 2016-09-29
Learn how to integrate full-stack open source big data architecture and to choose the correct technology—Scala/Spark, Mesos, Akka, Cassandra, and Kafka—in every layer. Big data architecture is becoming a requirement for many different enterprises. So far, however, the focus has largely been on collecting, aggregating, and crunching large data sets in a timely manner. In many cases now, organizations need more than one paradigm to perform efficient analyses. Big Data SMACK explains each of the full-stack technologies and, more importantly, how to best integrate them. It provides detailed coverage of the practical benefits of these technologies and incorporates real-world examples in every situation. This book focuses on the problems and scenarios solved by the architecture, as well as the solutions provided by every technology. It covers the six main concepts of big data architecture and how integrate, replace, and reinforce every layer: The language: Scala The engine: Spark (SQL, MLib, Streaming, GraphX) The container: Mesos, Docker The view: Akka The storage: Cassandra The message broker: Kafka What You Will Learn: Make big data architecture without using complex Greek letter architectures Build a cheap but effective cluster infrastructure Make queries, reports, and graphs that business demands Manage and exploit unstructured and No-SQL data sources Use tools to monitor the performance of your architecture Integrate all technologies and decide which ones replace and which ones reinforce Who This Book Is For: Developers, data architects, and data scientists looking to integrate the most

successful big data open stack architecture and to choose the correct technology in every layer

Real-Time Streaming with Apache Kafka, Spark, and Storm - Brindha Priyadarshini Jeyaraman 2021-08-20

Build a platform using Apache Kafka, Spark, and Storm to generate real-time data insights and view them through Dashboards. KEY FEATURES ● Extensive practical demonstration of Apache Kafka concepts, including producer and consumer examples. ● Includes graphical examples and explanations of implementing Kafka Producer and Kafka Consumer commands and methods. ● Covers integration and implementation of Spark-Kafka and Kafka-Storm architectures. DESCRIPTION Real-Time Streaming with Apache Kafka, Spark, and Storm is a book that provides an overview of the real-time streaming concepts and architectures of Apache Kafka, Storm, and Spark. The readers will learn how to build systems that can process data streams in real time using these technologies. They will be able to process a large amount of real-time data and perform analytics or generate insights as a result of this. The architecture of Kafka and its various components are described in detail. A Kafka Cluster installation and configuration will be demonstrated. The Kafka publisher-subscriber system will be implemented in the Eclipse IDE using the Command Line and Java. The book discusses the architecture of Apache Storm, the concepts of Spout and Bolt, as well as their applications in a Transaction Alert System. It also describes Spark's core concepts, applications, and the use of Spark to implement a microservice. To learn about the process of integrating Kafka and Storm, two approaches to Spark and Kafka integration will be discussed. This book will assist a software engineer to transition to a Big Data engineer and Big Data architect by providing knowledge of big data

processing and the architectures of Kafka, Storm, and Spark Streaming. WHAT YOU WILL LEARN ● Creation of Kafka producers, consumers, and brokers using command line. ● End-to-end implementation of Kafka messaging system with Java in Eclipse. ● Perform installation and creation of a Storm Cluster and execute Storm Management commands. ● Implement Spouts, Bolts and a Topology in Storm for Transaction alert application system. ● Perform the implementation of a microservice using Spark in Scala IDE. ● Learn about the various approaches of integrating Kafka and Spark. ● Perform integration of Kafka and Storm using Java in the Eclipse IDE. WHO THIS BOOK IS FOR This book is intended for Software Developers, Data Scientists, and Big Data Architects who want to build software systems to process data streams in real time. To understand the concepts in this book, knowledge of any programming language such as Java, Python, etc. is needed. TABLE OF CONTENTS 1. Chapter Two: Installing Kafka 2. Chapter Three: Kafka Messaging 3. Chapter Four: Kafka Producers 4. Chapter Five: Kafka Consumers 5. Chapter Six: Introduction to Storm 6. Chapter Seven: Installation and Configuration 7. Chapter Eight: Spouts and Bolts 8. Chapter Nine: Introduction to Spark 9. Chapter Ten: Spark Streaming 10. Chapter Eleven: Kafka Integration with Storm 11. Chapter Twelve: Kafka Integration with Spark

Kafka Streams in Action - Bill Bejeck 2018-08-29

Summary Kafka Streams in Action teaches you everything you need to know to implement stream processing on data flowing into your Kafka platform, allowing you to focus on getting more from your data without sacrificing time or effort. Foreword by Neha Narkhede, Cocreator of Apache Kafka Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications.

About the Technology Not all stream-based applications require a dedicated processing cluster. The lightweight Kafka Streams library provides exactly the power and simplicity you need for message handling in microservices and real-time event processing. With the Kafka Streams API, you filter and transform data streams with just Kafka and your application. About the Book Kafka Streams in Action teaches you to implement stream processing within the Kafka platform. In this easy-to-follow book, you'll explore real-world examples to collect, transform, and aggregate data, work with multiple processors, and handle real-time events. You'll even dive into streaming SQL with KSQL! Practical to the very end, it finishes with testing and operational aspects, such as monitoring and debugging. What's inside Using the KStreams API Filtering, transforming, and splitting data Working with the Processor API Integrating with external systems About the Reader Assumes some experience with distributed systems. No knowledge of Kafka or streaming applications required. About the Author Bill Bejeck is a Kafka Streams contributor and Confluent engineer with over 15 years of software development experience. Table of Contents PART 1 - GETTING STARTED WITH KAFKA STREAMS Welcome to Kafka Streams Kafka quicklyPART 2 - KAFKA STREAMS DEVELOPMENT Developing Kafka Streams Streams and state The KTable API The Processor APIPART 3 - ADMINISTERING KAFKA STREAMS Monitoring and performance Testing a Kafka Streams applicationPART 4 - ADVANCED CONCEPTS WITH KAFKA STREAMS Advanced applications with Kafka StreamsAPPENDIXES Appendix A - Additional configuration information Appendix B - Exactly once semantics

Kafka Streams in Action, Second Edition - Bill Bejeck 2022-12-27

Everything you need to implement stream processing on Apache Kafka? using Kafka Streams and the kqsIDB event streaming database. This totally revised new edition of Kafka Streams in Action has been expanded to cover more of the Kafka platform used for building event- based applications. You'll also find full coverage of ksqlDB, an event streaming database purpose-built for stream processing applications. Kafka Streams in Action, Second Edition teaches you to implement stream processing within the Kafka platform. In this easy-to-follow book, you'll explore real-world examples to collect, transform, and aggregate data, work with multiple processors, and handle real-time events. You'll also dive into processing event data with ksqlDB. Practical to the very end, it finishes with testing and operational aspects, such as monitoring, debugging, and gives you the opportunity to explore a few end-to-end projects. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications.

Kafka in Action - Dylan Scott 2022-02-15

Kafka in Action is a practical, hands-on guide to building Kafka-based data pipelines. Filled with real-world use cases and scenarios, this book probes Kafka's most common use cases, ranging from simple logging through managing streaming data systems for message routing, analytics, and more. In systems that handle big data, streaming data, or fast data, it's important to get your data pipelines right. Apache Kafka is a wicked-fast distributed streaming platform that operates as more than just a persistent log or a flexible message queue. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications.

Building Event-Driven Microservices - Adam Bellemare 2020-07-02

Organizations today often struggle to balance business requirements with ever-

increasing volumes of data. Additionally, the demand for leveraging large-scale, real-time data is growing rapidly among the most competitive digital industries. Conventional system architectures may not be up to the task. With this practical guide, you'll learn how to leverage large-scale data usage across the business units in your organization using the principles of event-driven microservices. Author Adam Bellemare takes you through the process of building an event-driven microservice-powered organization. You'll reconsider how data is produced, accessed, and propagated across your organization. Learn powerful yet simple patterns for unlocking the value of this data. Incorporate event-driven design and architectural principles into your own systems. And completely rethink how your organization delivers value by unlocking near-real-time access to data at scale. You'll learn: How to leverage event-driven architectures to deliver exceptional business value The role of microservices in supporting event-driven designs Architectural patterns to ensure success both within and between teams in your organization Application patterns for developing powerful event-driven microservices Components and tooling required to get your microservice ecosystem off the ground

Encyclopedia of Information Science and Technology, Fifth Edition - Khosrow-Pour D.B.A., Mehdi 2020-07-24

The rise of intelligence and computation within technology has created an eruption of potential applications in numerous professional industries. Techniques such as data analysis, cloud computing, machine learning, and others have altered the traditional processes of various disciplines including healthcare, economics, transportation, and politics. Information technology in today's world is beginning to uncover opportunities for experts in these fields that they are not

yet aware of. The exposure of specific instances in which these devices are being implemented will assist other specialists in how to successfully utilize these transformative tools with the appropriate amount of discretion, safety, and awareness. Considering the level of diverse uses and practices throughout the globe, the fifth edition of the Encyclopedia of Information Science and Technology series continues the enduring legacy set forth by its predecessors as a premier reference that contributes the most cutting-edge concepts and methodologies to the research community. The Encyclopedia of Information Science and Technology, Fifth Edition is a three-volume set that includes 136 original and previously unpublished research chapters that present multidisciplinary research and expert insights into new methods and processes for understanding modern technological tools and their applications as well as emerging theories and ethical controversies surrounding the field of information science. Highlighting a wide range of topics such as natural language processing, decision support systems, and electronic government, this book offers strategies for implementing smart devices and analytics into various professional disciplines. The techniques discussed in this publication are ideal for IT professionals, developers, computer scientists, practitioners, managers, policymakers, engineers, data analysts, and programmers seeking to understand the latest developments within this field and who are looking to apply new tools and policies in their practice. Additionally, academicians, researchers, and students in fields that include but are not limited to software engineering, cybersecurity, information technology, media and communications, urban planning, computer science, healthcare, economics, environmental science, data management, and political science will benefit from the

extensive knowledge compiled within this publication.

Apache Hive Essentials - Dayong Du 2018-06-30

This book takes you on a fantastic journey to discover the attributes of big data using Apache Hive. Key Features Grasp the skills needed to write efficient Hive queries to analyze the Big Data Discover how Hive can coexist and work with other tools within the Hadoop ecosystem Uses practical, example-oriented scenarios to cover all the newly released features of Apache Hive 2.3.3 Book Description In this book, we prepare you for your journey into big data by frstly introducing you to backgrounds in the big data domain, alongwith the process of setting up and getting familiar with your Hive working environment. Next, the book guides you through discovering and transforming the values of big data with the help of examples. It also hones your skills in using the Hive language in an effcient manner. Toward the end, the book focuses on advanced topics, such as performance, security, and extensions in Hive, which will guide you on exciting adventures on this worthwhile big data journey. By the end of the book, you will be familiar with Hive and able to work effeciently to find solutions to big data problems What you will learn Create and set up the Hive environment Discover how to use Hive's definition language to describe data Discover interesting data by joining and filtering datasets in Hive Transform data by using Hive sorting, ordering, and functions Aggregate and sample data in different ways Boost Hive query performance and enhance data security in Hive Customize Hive to your needs by using user-defined functions and integrate it with other tools Who this book is for If you are a data analyst, developer, or simply someone who wants to quickly get started with Hive to explore and analyze Big Data in Hadoop, this is the book for you. Since Hive is an SQL-like language, some previous experience with SQL will be

useful to get the most out of this book.

Kafka: The Definitive Guide - Gwen Shapira 2021-11-05
Every enterprise application creates data, whether it consists of log messages, metrics, user activity, or outgoing messages. Moving all this data is just as important as the data itself. With this updated edition, application architects, developers, and production engineers new to the Kafka streaming platform will learn how to handle data in motion. Additional chapters cover Kafka's AdminClient API, transactions, new security features, and tooling changes. Engineers from Confluent and LinkedIn responsible for developing Kafka explain how to deploy production Kafka clusters, write reliable event-driven microservices, and build scalable stream processing applications with this platform. Through detailed examples, you'll learn Kafka's design principles, reliability guarantees, key APIs, and architecture details, including the replication protocol, the controller, and the storage layer. You'll examine: Best practices for deploying and configuring Kafka Kafka producers and consumers for writing and reading messages Patterns and use-case requirements to ensure reliable data delivery Best practices for building data pipelines and applications with Kafka How to perform monitoring, tuning, and maintenance tasks with Kafka in production The most critical metrics among Kafka's operational measurements Kafka's delivery capabilities for stream processing systems

Mastering Apache Pulsar - Jowanza Joseph 2021-12-06
Every enterprise application creates data, including log messages, metrics, user activity, and outgoing messages. Learning how to move these items is almost as important as the data itself. If you're an application architect, developer, or production engineer new to Apache Pulsar, this practical guide shows you how to use this open source event streaming platform to handle real-time data feeds.

Jowanza Joseph, staff software engineer at Finicity, explains how to deploy production Pulsar clusters, write reliable event streaming applications, and build scalable real-time data pipelines with this platform. Through detailed examples, you'll learn Pulsar's design principles, reliability guarantees, key APIs, and architecture details, including the replication protocol, the load manager, and the storage layer. This book helps you: Understand how event streaming fits in the big data ecosystem Explore Pulsar producers, consumers, and readers for writing and reading events Build scalable data pipelines by connecting Pulsar with external systems Simplify event-streaming application building with Pulsar Functions Manage Pulsar to perform monitoring, tuning, and maintenance tasks Use Pulsar's operational measurements to secure a production cluster Process event streams using Flink and query event streams using Presto

Big Data Processing with Apache Spark - Srini Penchikala 2018-03-13 Apache Spark is a popular open-source big-data processing framework thatÕs built around speed, ease of use, and unified distributed computing architecture. Not only it supports developing applications in different languages like Java, Scala, Python, and R, itÕs also hundred times faster in memory and ten times faster even when running on disk compared to traditional data processing frameworks. Whether you are currently working on a big data project or interested in learning more about topics like machine learning, streaming data processing, and graph data analytics, this book is for you. You can learn about Apache Spark and develop Spark programs for various use cases in big data analytics using the code examples provided. This book covers all the libraries in Spark ecosystem: Spark Core, Spark SQL, Spark Streaming, Spark ML, and Spark GraphX.

Data Lake for Enterprises -

Tomcy John 2017-05-31
A practical guide to implementing your enterprise data lake using Lambda Architecture as the base About This Book Build a full-fledged data lake for your organization with popular big data technologies using the Lambda architecture as the base Delve into the big data technologies required to meet modern day business strategies A highly practical guide to implementing enterprise data lakes with lots of examples and real-world use-cases Who This Book Is For Java developers and architects who would like to implement a data lake for their enterprise will find this book useful. If you want to get hands-on experience with the Lambda Architecture and big data technologies by implementing a practical solution using these technologies, this book will also help you. What You Will Learn Build an enterprise-level data lake using the relevant big data technologies Understand the core of the Lambda architecture and how to apply it in an enterprise Learn the technical details around Sqoop and its functionalities Integrate Kafka with Hadoop components to acquire enterprise data Use flume with streaming technologies for stream-based processing Understand stream- based processing with reference to Apache Spark Streaming Incorporate Hadoop components and know the advantages they provide for enterprise data lakes Build fast, streaming, and high-performance applications using ElasticSearch Make your data ingestion process consistent across various data formats with configurability Process your data to derive intelligence using machine learning algorithms In Detail The term "Data Lake" has recently emerged as a prominent term in the big data industry. Data scientists can make use of it in deriving meaningful insights that can be used by businesses to redefine or transform the way they operate. Lambda architecture is also emerging as one of the very eminent

patterns in the big data landscape, as it not only helps to derive useful information from historical data but also correlates real-time data to enable business to take critical decisions. This book tries to bring these two important aspects — data lake and lambda architecture—together. This book is divided into three main sections. The first introduces you to the concept of data lakes, the importance of data lakes in enterprises, and getting you up-to-speed with the Lambda architecture. The second section delves into the principal components of building a data lake using the Lambda architecture. It introduces you to popular big data technologies such as Apache Hadoop, Spark, Sqoop, Flume, and ElasticSearch. The third section is a highly practical demonstration of putting it all together, and shows you how an enterprise data lake can be implemented, along with several real-world use-cases. It also shows you how other peripheral components can be added to the lake to make it more efficient. By the end of this book, you will be able to choose the right big data technologies using the lambda architectural patterns to build your enterprise data lake. Style and approach The book takes a pragmatic approach, showing ways to leverage big data technologies and lambda architecture to build an enterprise-level data lake.

Mastering Hadoop 3 - Chanchal Singh 2019-02-28

A comprehensive guide to mastering the most advanced Hadoop 3 concepts Key FeaturesGet to grips with the newly introduced features and capabilities of Hadoop 3Crunch and process data using MapReduce, YARN, and a host of tools within the Hadoop ecosystemSharpen your Hadoop skills with real-world case studies and codeBook Description Apache Hadoop is one of the most popular big data solutions for distributed storage and for processing large chunks of data. With Hadoop 3, Apache promises to provide a high-performance,

more fault-tolerant, and highly efficient big data processing platform, with a focus on improved scalability and increased efficiency. With this guide, you'll understand advanced concepts of the Hadoop ecosystem tool. You'll learn how Hadoop works internally, study advanced concepts of different ecosystem tools, discover solutions to real-world use cases, and understand how to secure your cluster. It will then walk you through HDFS, YARN, MapReduce, and Hadoop 3 concepts. You'll be able to address common challenges like using Kafka efficiently, designing low latency, reliable message delivery Kafka systems, and handling high data volumes. As you advance, you'll discover how to address major challenges when building an enterprise-grade messaging system, and how to use different stream processing systems along with Kafka to fulfil your enterprise goals. By the end of this book, you'll have a complete understanding of how components in the Hadoop ecosystem are effectively integrated to implement a fast and reliable data pipeline, and you'll be equipped to tackle a range of real-world problems in data pipelines. What you will learnGain an in-depth understanding of distributed computing using Hadoop 3Develop enterprise-grade applications using Apache Spark, Flink, and moreBuild scalable and high-performance Hadoop data pipelines with security, monitoring, and data governanceExplore batch data processing patterns and how to model data in HadoopMaster best practices for enterprises using, or planning to use, Hadoop 3 as a data platformUnderstand security aspects of Hadoop, including authorization and authenticationWho this book is for If you want to become a big data professional by mastering the advanced concepts of Hadoop, this book is for you. You'll also find this book useful if you're a Hadoop professional looking to strengthen your knowledge of the Hadoop

ecosystem. Fundamental knowledge of the Java programming language and basics of Hadoop is necessary to get started with this book.

Enterprise Integration Patterns - Gregor Hohpe 2012-03-09

Enterprise Integration Patterns provides an invaluable catalog of sixty-five patterns, with real-world solutions that demonstrate the formidable of messaging and help you to design effective messaging solutions for your enterprise. The authors also include examples covering a variety of different integration technologies, such as JMS, MSMQ, TIBCO ActiveEnterprise, Microsoft BizTalk, SOAP, and XSL. A case study describing a bond trading system illustrates the patterns in practice, and the book offers a look at emerging standards, as well as insights into what the future of enterprise integration might hold. This book provides a consistent vocabulary and visual notation framework to describe large-scale integration solutions across many technologies. It also explores in detail the advantages and limitations of asynchronous messaging architectures. The authors present practical advice on designing code that connects an application to a messaging system, and provide extensive information to help you determine when to send a message, how to route it to the proper destination, and how to monitor the health of a messaging system. If you want to know how to manage, monitor, and maintain a messaging system once it is in use, get this book.

Effective Kafka - Emil Koutanov 2020-03-17

The software architecture landscape has evolved dramatically over the past decade. Microservices have displaced monoliths. Data and applications are increasingly becoming distributed and decentralised. But composing disparate systems is a hard problem. More recently, software practitioners have been rapidly converging on event-driven architecture as a sustainable way of dealing with

complexity - integrating systems without increasing their coupling.In Effective Kafka, Emil Koutanov explores the fundamentals of Event-Driven Architecture - using Apache Kafka - the world's most popular and supported open-source event streaming platform.You'll learn: - The fundamentals of event-driven architecture and event streaming platforms- The background and rationale behind Apache Kafka, its numerous potential uses and applications- The architecture and core concepts - the underlying software components, partitioning and parallelism, load-balancing, record ordering and consistency modes- Installation of Kafka and related tooling - using standalone deployments, clusters, and containerised deployments with Docker- Using CLI tools to interact with and administer Kafka classes, as well as publishing data and browsing topics- Using third-party web-based tools for monitoring a cluster and gaining insights into the event streams- Building stream processing applications in Java 11 using off-the-shelf client libraries- Patterns and best-practice for organising the application architecture, with emphasis on maintainability and testability of the resulting code- The numerous gotchas that lurk in Kafka's client and broker configuration, and how to counter them- Theoretical background on distributed and concurrent computing, exploring factors affecting their liveness and safety- Best-practices for running multi-tenanted clusters across diverse engineering teams, how teams collaborate to build complex systems at scale and equitably share the cluster with the aid of quotas- Operational aspects of running Kafka clusters at scale, performance tuning and methods for optimising network and storage utilisation- All aspects of Kafka security -including network segregation, encryption, certificates, authentication and authorization.The coverage is progressively delivered and

carefully aimed at giving you a journey-like experience into becoming proficient with Apache Kafka and Event-Driven Architecture. The goal is to get you designing and building applications. And by the conclusion of this book, you will be a confident practitioner and a Kafka evangelist within your organisation - wielding the knowledge necessary to teach others.

Building Data Streaming Applications with Apache Kafka - Manish Kumar 2017-08-18

Design and administer fast, reliable enterprise messaging systems with Apache Kafka About This Book Build efficient real-time streaming applications in Apache Kafka to process data streams of data Master the core Kafka APIs to set up Apache Kafka clusters and start writing message producers and consumers A comprehensive guide to help you get a solid grasp of the Apache Kafka concepts in Apache Kafka with pracitcalpractical examples Who This Book Is For If you want to learn how to use Apache Kafka and the different tools in the Kafka ecosystem in the easiest possible manner, this book is for you. Some programming experience with Java is required to get the most out of this book What You Will Learn Learn the basics of Apache Kafka from scratch Use the basic building blocks of a streaming application Design effective streaming applications with Kafka using Spark, Storm &, and Heron Understand the importance of a low -latency , high-throughput, and fault-tolerant messaging system Make effective capacity planning while deploying your Kafka Application Understand and implement the best security practices In Detail Apache Kafka is a popular distributed streaming platform that acts as a messaging queue or an enterprise messaging system. It lets you publish and subscribe to a stream of records, and process them in a fault-tolerant way as they occur. This book is a comprehensive guide to

designing and architecting enterprise-grade streaming applications using Apache Kafka and other big data tools. It includes best practices for building such applications, and tackles some common challenges such as how to use Kafka efficiently and handle high data volumes with ease. This book first takes you through understanding the type messaging system and then provides a thorough introduction to Apache Kafka and its internal details. The second part of the book takes you through designing streaming application using various frameworks and tools such as Apache Spark, Apache Storm, and more. Once you grasp the basics, we will take you through more advanced concepts in Apache Kafka such as capacity planning and security. By the end of this book, you will have all the information you need to be comfortable with using Apache Kafka, and to design efficient streaming data applications with it. Style and approach A step-by –step, comprehensive guide filled with practical and real- world examples

Kafka: The Definitive Guide - Neha Narkhede 2017-08-31

Every enterprise application creates data, whether it’s log messages, metrics, user activity, outgoing messages, or something else. And how to move all of this data becomes nearly as important as the data itself. If you’re an application architect, developer, or production engineer new to Apache Kafka, this practical guide shows you how to use this open source streaming platform to handle real-time data feeds. Engineers from Confluent and LinkedIn who are responsible for developing Kafka explain how to deploy production Kafka clusters, write reliable event-driven microservices, and build scalable stream-processing applications with this platform. Through detailed examples, you’ll learn Kafka’s design principles, reliability guarantees, key APIs, and architecture details, including the replication protocol, the controller, and the storage

layer. Understand publish-subscribe messaging and how it fits in the big data ecosystem. Explore Kafka producers and consumers for writing and reading messages Understand Kafka patterns and use-case requirements to ensure reliable data delivery Get best practices for building data pipelines and applications with Kafka Manage Kafka in production, and learn to perform monitoring, tuning, and maintenance tasks Learn the most critical metrics among Kafka's operational measurements Explore how Kafka's stream delivery capabilities make it a perfect source for stream processing systems

Event Streams in Action - Valentin Crettaz 2019-05-10 Summary Event Streams in Action is a foundational book introducing the ULP paradigm and presenting techniques to use it effectively in data-rich environments. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Many high-profile applications, like LinkedIn and Netflix, deliver nimble, responsive performance by reacting to user and system events as they occur. In large-scale systems, this requires efficiently monitoring, managing, and reacting to multiple event streams. Tools like Kafka, along with innovative patterns like unified log processing, help create a coherent data processing architecture for event-based applications. About the Book Event Streams in Action teaches you techniques for aggregating, storing, and processing event streams using the unified log processing pattern. In this hands-on guide, you'll discover important application designs like the lambda architecture, stream aggregation, and event reprocessing. You'll also explore scaling, resiliency, advanced stream patterns, and much more! By the time you're finished, you'll be designing large-scale data-driven applications that are easier to build, deploy, and maintain.

What's inside Validating and monitoring event streams Event analytics Methods for event modeling Examples using Apache Kafka and Amazon Kinesis About the Reader For readers with experience coding in Java, Scala, or Python. About the Author Alexander Dean developed Snowplow, an open source event processing and analytics platform. Valentin Crettaz is an independent IT consultant with 25 years of experience. Table of Contents PART 1 - EVENT STREAMS AND UNIFIED LOGS Introducing event streams The unified log 24 Event stream processing with Apache Kafka Event stream processing with Amazon Kinesis Stateful stream processing PART 2- DATA ENGINEERING WITH STREAMS Schemas Archiving events Railway-oriented processing Commands PART 3 - EVENT ANALYTICS Analytics-on-read Analytics-on-write

Kafka Streams - Real-time Stream Processing - Prashant Kumar Pandey 2019-03-26

The book Kafka Streams - Real-time Stream Processing helps you understand the stream processing in general and apply that skill to Kafka streams programming. This book is focusing mainly on the new generation of the Kafka Streams library available in the Apache Kafka 2.x. The primary focus of this book is on Kafka Streams. However, the book also touches on the other Apache Kafka capabilities and concepts that are necessary to grasp the Kafka Streams programming. Who should read this book? Kafka Streams: Real-time Stream Processing is written for software engineers willing to develop a stream processing application using Kafka Streams library. I am also writing this book for data architects and data engineers who are responsible for designing and building the organization’s data-centric infrastructure. Another group of people is the managers and architects who do not directly work with Kafka implementation, but they work with the people who implement Kafka Streams at the ground level. What should you already

know? This book assumes that the reader is familiar with the basics of Java programming language. The source code and examples in this book are using Java 8, and I will be using Java 8 lambda syntax, so experience with lambda will be helpful. Kafka Streams is a library that runs on Kafka. Having a good fundamental knowledge of Kafka is essential to get the most out of Kafka Streams. I will touch base on the mandatory Kafka concepts for those who are new to Kafka. The book also assumes that you have some familiarity and experience in running and working on the Linux operating system.

Streaming Systems - Tyler Akidau 2018-07-16

Streaming data is a big deal in big data these days. As more and more businesses seek to tame the massive unbounded data sets that pervade our world, streaming systems have finally reached a level of maturity sufficient for mainstream adoption. With this practical guide, data engineers, data scientists, and developers will learn how to work with streaming data in a conceptual and platform-agnostic way. Expanded from Tyler Akidau's popular blog posts "Streaming 101" and "Streaming 102", this book takes you from an introductory level to a nuanced understanding of the what, where, when, and how of processing real-time data streams. You'll also dive deep into watermarks and exactly-once processing with co-authors Slava Chernyak and Reuven Lax. You'll explore: How streaming and batch data processing patterns compare The core principles and concepts behind robust out-of-order data processing How watermarks track progress and completeness in infinite datasets How exactly-once data processing techniques ensure correctness How the concepts of streams and tables form the foundations of both batch and streaming data processing The practical motivations behind a powerful persistent state mechanism, driven by a real-world example How time-varying relations provide a link

between stream processing and the world of SQL and relational algebra

Hands-On Big Data Modeling - James Lee 2018-11-30

Solve all big data problems by learning how to create efficient data models Key FeaturesCreate effective models that get the most out of big dataApply your knowledge to datasets from Twitter and weather data to learn big dataTackle different data modeling challenges with expert techniques presented in this bookBook Description Modeling and managing data is a central focus of all big data projects. In fact, a database is considered to be effective only if you have a logical and sophisticated data model. This book will help you develop practical skills in modeling your own big data projects and improve the performance of analytical queries for your specific business requirements. To start with, you'll get a quick introduction to big data and understand the different data modeling and data management platforms for big data. Then you'll work with structured and semi-structured data with the help of real-life examples. Once you've got to grips with the basics, you'll use the SQL Developer Data Modeler to create your own data models containing different file types such as CSV, XML, and JSON. You'll also learn to create graph data models and explore data modeling with streaming data using real-world datasets. By the end of this book, you'll be able to design and develop efficient data models for varying data sizes easily and efficiently. What you will learnGet insights into big data and discover various data modelsExplore conceptual, logical, and big data modelsUnderstand how to model data containing different file typesRun through data modeling with examples of Twitter, Bitcoin, IMDB and weather data modelingCreate data models such as Graph Data and Vector SpaceModel structured and unstructured data using Python and RWho this book is for This book is

great for programmers, geologists, biologists, and every professional who deals with spatial data. If you want to learn how to handle GIS, GPS, and remote sensing data, then this book is for you. Basic knowledge of R and QGIS would be helpful.

Apache Kafka 1.0 Cookbook - Raúl Estrada 2017-12-22

Simplify real-time data processing by leveraging the power of Apache Kafka 1.0 Key Features Use Kafka 1.0 features such as Confluent platforms and Kafka streams to build efficient streaming data applications to handle and process your data Integrate Kafka with other Big Data tools such as Apache Hadoop, Apache Spark, and more Hands-on recipes to help you design, operate, maintain, and secure your Apache Kafka cluster with ease Book Description Apache Kafka provides a unified, high-throughput, low-latency platform to handle real-time data feeds. This book will show you how to use Kafka efficiently, and contains practical solutions to the common problems that developers and administrators usually face while working with it. This practical guide contains easy-to-follow recipes to help you set up, configure, and use Apache Kafka in the best possible manner. You will use Apache Kafka Consumers and Producers to build effective real-time streaming applications. The book covers the recently released Kafka version 1.0, the Confluent Platform and Kafka Streams. The programming aspect covered in the book will teach you how to perform important tasks such as message validation, enrichment and composition.Recipes focusing on optimizing the performance of your Kafka cluster, and integrate Kafka with a variety of third-party tools such as Apache Hadoop, Apache Spark, and Elasticsearch will help ease your day to day collaboration with Kafka greatly. Finally, we cover tasks related to monitoring and securing your Apache Kafka cluster using tools such as

Ganglia and Graphite. If you're looking to become the go-to person in your organization when it comes to working with Apache Kafka, this book is the only resource you need to have. What you will learn -Install and configure Apache Kafka 1.0 to get optimal performance - Create and configure Kafka Producers and Consumers - Operate your Kafka clusters efficiently by implementing the mirroring technique -Work with the new Confluent platform and Kafka streams, and achieve high availability with Kafka - Monitor Kafka using tools such as Graphite and Ganglia - Integrate Kafka with third-party tools such as Elasticsearch, Logstash, Apache Hadoop, Apache Spark, and more Who this book is for This book is for developers and Kafka administrators who are looking for quick, practical solutions to problems encountered while operating, managing or monitoring Apache Kafka. If you are a developer, some knowledge of Scala or Java will help, while for administrators, some working knowledge of Kafka will be useful.

Stream Processing with Apache Spark - Gerard Maas 2019-06-05

Before you can build analytics tools to gain quick insights, you first need to know how to process data in real time. With this practical guide, developers familiar with Apache Spark will learn how to put this in-memory framework to use for streaming data. You'll discover how Spark enables you to write streaming jobs in almost the same way you write batch jobs. Authors Gerard Maas and François Garillot help you explore the theoretical underpinnings of Apache Spark. This comprehensive guide features two sections that compare and contrast the streaming APIs Spark now supports: the original Spark Streaming library and the newer Structured Streaming API. Learn fundamental stream processing concepts and examine different streaming architectures Explore Structured Streaming through practical examples; learn

different aspects of stream processing in detail Create and operate streaming jobs and applications with Spark Streaming; integrate Spark Streaming with other Spark APIs Learn advanced Spark Streaming techniques, including approximation algorithms and machine learning algorithms Compare Apache Spark to other stream processing projects, including Apache Storm, Apache Flink, and Apache Kafka Streams

Streaming Data - Andrew Psaltis 2017-05-31

Summary Streaming Data introduces the concepts and requirements of streaming and real-time data systems. The book is an idea-rich tutorial that teaches you to think about how to efficiently interact with fast-flowing data. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology As humans, we're constantly filtering and deciphering the information streaming toward us. In the same way, streaming data applications can accomplish amazing tasks like reading live location data to recommend nearby services, tracking faults with machinery in real time, and sending digital receipts before your customers leave the shop. Recent advances in streaming data technology and techniques make it possible for any developer to build these applications if they have the right mindset. This book will let you join them. About the Book Streaming Data is an idea-rich tutorial that teaches you to think about efficiently interacting with fast-flowing data. Through relevant examples and illustrated use cases, you'll explore designs for applications that read, analyze, share, and store streaming data. Along the way, you'll discover the roles of key technologies like Spark, Storm, Kafka, Flink, RabbitMQ, and more. This book offers the perfect balance between big-picture thinking and implementation details. What's Inside The right way to collect real-time data Architecting a streaming pipeline Analyzing

the data Which technologies to use and when About the Reader Written for developers familiar with relational database concepts. No experience with streaming or real-time applications required. About the Author Andrew Psaltis is a software engineer focused on massively scalable real-time analytics. Table of Contents PART 1 - A NEW HOLISTIC APPROACH Introducing streaming data Getting data from clients: data ingestion Transporting the data from collection tier: decoupling the data pipeline Analyzing streaming data Algorithms for data analysis Storing the analyzed or collected data Making the data available Consumer device capabilities and limitations accessing the data PART 2 - TAKING IT REAL WORLD Analyzing Meetup RSVPs in real time

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 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

Streaming Architecture - Ted Dunning 2016-05-10

More and more data-driven companies are looking to adopt stream processing and streaming analytics. With this concise ebook, you'll learn best practices for designing a reliable architecture that supports this emerging big-data paradigm. Authors Ted Dunning and Ellen Friedman (Real World Hadoop) help you explore some of the best technologies to handle stream processing and analytics, with a focus on the upstream queuing or message-passing layer. To illustrate the effectiveness of these technologies, this book also includes specific use cases. Ideal for developers and non-technical people alike, this book describes: Key elements in good design for streaming analytics, focusing on the essential characteristics of the messaging layer New messaging technologies, including Apache Kafka and MapR Streams, with links to sample code Technology choices for streaming analytics: Apache Spark Streaming, Apache Flink, Apache Storm, and Apache Apex How stream-based architectures are helpful to support microservices Specific use cases such as fraud detection and geo-distributed data streams Ted Dunning is Chief Applications Architect at MapR Technologies, and active in the open source community. He currently serves as VP for Incubator at the Apache Foundation, as a champion and mentor for a large number of projects, and as committer and PMC member of the Apache ZooKeeper and Drill projects. Ted is on Twitter as @ted_dunning. Ellen Friedman, a committer for the Apache Drill and Apache Mahout projects, is a solutions consultant and well-known speaker and author, currently writing mainly about big data topics. With a PhD in Biochemistry, she has years of experience as a research scientist and has written about a variety of technical topics. Ellen is on Twitter as

@Ellen_Friedman.

Practical Real-time Data Processing and Analytics - Shilpi Saxena 2017-09-28

A practical guide to help you tackle different real-time data processing and analytics problems using the best tools for each scenario About This Book Learn about the various challenges in real-time data processing and use the right tools to overcome them This book covers popular tools and frameworks such as Spark, Flink, and Apache Storm to solve all your distributed processing problems A practical guide filled with examples, tips, and tricks to help you perform efficient Big Data processing in real-time Who This Book Is For If you are a Java developer who would like to be equipped with all the tools required to devise an end-to-end practical solution on real-time data streaming, then this book is for you. Basic knowledge of real-time processing would be helpful, and knowing the fundamentals of Maven, Shell, and Eclipse would be great. What You Will Learn Get an introduction to the established real-time stack Understand the key integration of all the components Get a thorough understanding of the basic building blocks for real-time solution designing Garnish the search and visualization aspects for your real-time solution Get conceptually and practically acquainted with real-time analytics Be well equipped to apply the knowledge and create your own solutions In Detail With the rise of Big Data, there is an increasing need to process large amounts of data continuously, with a shorter turnaround time. Real-time data processing involves continuous input, processing and output of data, with the condition that the time required for processing is as short as possible. This book covers the majority of the existing and evolving open source technology stack for real-time processing and analytics. You will get to know about all the real-time solution aspects, from the source to the presentation to persistence.

Through this practical book, you'll be equipped with a clear understanding of how to solve challenges on your own. We'll cover topics such as how to set up components, basic executions, integrations, advanced use cases, alerts, and monitoring. You'll be exposed to the popular tools used in real-time processing today such as Apache Spark, Apache Flink, and Storm. Finally, you will put your knowledge to practical use by implementing all of the techniques in the form of a practical, real-world use case. By the end of this book, you will have a solid understanding of all the aspects of real-time data processing and analytics, and will know how to deploy the solutions in production environments in the best possible manner. Style and Approach In this practical guide to real-time analytics, each chapter begins with a basic high-level concept of the topic, followed by a practical, hands-on implementation of each concept, where you can see the working and execution of it. The book is written in a DIY style, with plenty of practical use cases, well-explained code examples, and relevant screenshots and diagrams.

Streaming Architecture - Ted Dunning 2016-05-10

More and more data-driven companies are looking to adopt stream processing and streaming analytics. With this concise ebook, you'll learn best practices for designing a reliable architecture that supports this emerging big-data paradigm. Authors Ted Dunning and Ellen Friedman (Real World Hadoop) help you explore some of the best technologies to handle stream processing and analytics, with a focus on the upstream queuing or message-passing layer. To illustrate the effectiveness of these technologies, this book also includes specific use cases. Ideal for developers and non-technical people alike, this book describes: Key elements in good design for streaming analytics, focusing on the essential characteristics of the

messaging layer New messaging technologies, including Apache Kafka and MapR Streams, with links to sample code Technology choices for streaming analytics: Apache Spark Streaming, Apache Flink, Apache Storm, and Apache Apex How stream-based architectures are helpful to support microservices Specific use cases such as fraud detection and geo-distributed data streams Ted Dunning is Chief Applications Architect at MapR Technologies, and active in the open source community. He currently serves as VP for Incubator at the Apache Foundation, as a champion and mentor for a large number of projects, and as committer and PMC member of the Apache ZooKeeper and Drill projects. Ted is on Twitter as @ted_dunning. Ellen Friedman, a committer for the Apache Drill and Apache Mahout projects, is a solutions consultant and well-known speaker and author, currently writing mainly about big data topics. With a PhD in Biochemistry, she has years of experience as a research scientist and has written about a variety of technical topics. Ellen is on Twitter as @Ellen_Friedman.

Mastering Kafka Streams and ksqlDB - Mitch Seymour 2021-02-04

Working with unbounded and fast-moving data streams has historically been difficult. But with Kafka Streams and ksqlDB, building stream processing applications is easy and fun. This practical guide shows data engineers how to use these tools to build highly scalable stream processing applications for moving, enriching, and transforming large amounts of data in real time. Mitch Seymour, data services engineer at Mailchimp, explains important stream processing concepts against a backdrop of several interesting business problems. You'll learn the strengths of both Kafka Streams and ksqlDB to help you choose the best tool for each unique stream processing project. Non-Java developers will find the ksqlDB

path to be an especially gentle introduction to stream processing. Learn the basics of Kafka and the pub/sub communication pattern Build stateless and stateful stream processing applications using Kafka Streams and ksqlDB Perform advanced stateful operations, including windowed joins and aggregations Understand how stateful processing works under the hood Learn about ksqlDB's data integration features, powered by Kafka Connect Work with different types of collections in ksqlDB and perform push and pull queries Deploy your Kafka Streams and ksqlDB applications to production

Kafka: the Definitive Guide - Gwen Shapira 2022-01-18

Every enterprise application creates data, whether it consists of log messages, metrics, user activity, outgoing messages, or something else. Moving all of this data is just as important as the data itself. This bookâ??s updated second edition shows application architects, developers, and production engineers new to the Kafka open source streaming platform how to handle real-time data feeds. Additional chapters cover Kafkaâ??s AdminClient API, new security features, and tooling changes. Engineers from Confluent and LinkedIn responsible for developing Kafka explain how to deploy production Kafka clusters, write reliable event-driven microservices, and build scalable stream processing applications with this platform. Through detailed examples, youâ??ll learn Kafkaâ??s design principles, reliability guarantees, key APIs, and architecture details, including the replication protocol, the controller, and the storage layer. Youâ??ll examine: How publish-subscribe messaging fits in the big data ecosystem Kafka producers and consumers for writing and reading messages Patterns and use-case requirements to ensure reliable data delivery Best practices for building data pipelines and applications with Kafka How to perform

monitoring, tuning, and maintenance tasks with Kafka in production The most critical metrics among Kafkaâ??s operational measurements Kafkaâ??s delivery capabilities for stream processing systems

Graph Algorithms - Mark Needham 2019-05-16

Discover how graph algorithms can help you leverage the relationships within your data to develop more intelligent solutions and enhance your machine learning models. You'll learn how graph analytics are uniquely suited to unfold complex structures and reveal difficult-to-find patterns lurking in your data. Whether you are trying to build dynamic network models or forecast real-world behavior, this book illustrates how graph algorithms deliver value—from finding vulnerabilities and bottlenecks to detecting communities and improving machine learning predictions. This practical book walks you through hands-on examples of how to use graph algorithms in Apache Spark and Neo4j—two of the most common choices for graph analytics. Also included: sample code and tips for over 20 practical graph algorithms that cover optimal pathfinding, importance through centrality, and community detection. Learn how graph analytics vary from conventional statistical analysis Understand how classic graph algorithms work, and how they are applied Get guidance on which algorithms to use for different types of questions Explore algorithm examples with working code and sample datasets from Spark and Neo4j See how connected feature extraction can increase machine learning accuracy and precision Walk through creating an ML workflow for link prediction combining Neo4j and Spark

Learning Apache Kafka Second Edition - Nishant Garg 2015-02-26

This book is for readers who want to know more about Apache Kafka at a hands-on level; the key audience is those with software development experience but no prior exposure to Apache Kafka or

similar technologies. It is also useful for enterprise application developers and big data enthusiasts who have worked with other publisher-subscriber-based systems and want to explore Apache Kafka as a futuristic solution.

Stream Processing with Apache Flink - Fabian Hueske 2019-04-11

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

Apache Kafka Quick Start Guide - Raúl Estrada 2018-12-27

Process large volumes of data in real-time while building high performance and robust data stream processing pipeline using the latest Apache Kafka

2.0 Key FeaturesSolve practical large data and processing challenges with KafkaTackle data processing challenges like late events, windowing, and watermarkingUnderstand real-time streaming applications processing using Schema registry, Kafka connect, Kafka streams, and KSQLBook Description Apache Kafka is a great open source platform for handling your real-time data pipeline to ensure high-speed filtering and pattern matching on the fly. In this book, you will learn how to use Apache Kafka for efficient processing of distributed applications and will get familiar with solving everyday problems in fast data and processing pipelines. This book focuses on programming rather than the configuration management of Kafka clusters or DevOps. It starts off with the installation and setting up the development environment, before quickly moving on to performing fundamental messaging operations such as validation and enrichment. Here you will learn about message composition with pure Kafka API and Kafka Streams. You will look into the transformation of messages in different formats, such asext, binary, XML, JSON, and AVRO. Next, you will learn how to expose the schemas contained in Kafka with the Schema Registry. You will then learn how to work with all relevant connectors with Kafka Connect. While working with Kafka Streams, you will perform various interesting operations on streams, such as windowing, joins, and aggregations. Finally, through KSQL, you will learn how to retrieve, insert, modify, and delete data streams, and how to manipulate watermarks and windows. What you will learnHow to validate data with KafkaAdd information to existing data flowsGenerate new information through message compositionPerform data validation and versioning with the Schema RegistryHow to perform message Serialization and DeserializationHow to perform message Serialization and

DeserializationProcess data streams with Kafka StreamsUnderstand the duality between tables and streams with KSQLWho this book is for This book is for developers who want to quickly master the practical concepts behind Apache Kafka. The audience need not have come across Apache Kafka previously; however, a familiarity of Java or any JVM language will be helpful in understanding the code in this book.

Mastering Apache Storm - Ankit Jain 2017-08-16

Master the intricacies of Apache Storm and develop real-time stream processing applications with ease About This Book Exploit the various real-time processing functionalities offered by Apache Storm such as parallelism, data partitioning, and more Integrate Storm with other Big Data technologies like Hadoop, HBase, and Apache Kafka An easy-to-understand guide to effortlessly create distributed applications with Storm Who This Book Is For If you are a Java developer who wants to enter into the world of real-time stream processing applications using Apache Storm, then this book is for you. No previous experience in Storm is required as this book starts from the basics. After finishing this book, you will be able to develop not-so-complex Storm applications. What You Will Learn Understand the core concepts of Apache Storm and real-time processing Follow the steps to deploy multiple nodes of Storm Cluster Create Trident topologies to support various message-processing semantics Make your cluster sharing effective using Storm scheduling Integrate Apache Storm with other Big Data technologies such as Hadoop, HBase, Kafka, and more Monitor the health of your Storm cluster In Detail Apache Storm is a real-time Big Data processing framework that processes large amounts of data reliably, guaranteeing that every message will be processed. Storm allows you to scale your data as it grows, making it an excellent platform

to solve your big data problems. This extensive guide will help you understand right from the basics to the advanced topics of Storm. The book begins with a detailed introduction to real-time processing and where Storm fits in to solve these problems. You'll get an understanding of deploying Storm on clusters by writing a basic Storm Hello World example. Next we'll introduce you to Trident and you'll get a clear understanding of how you can develop and deploy a trident topology. We cover topics such as monitoring, Storm Parallelism, scheduler and log processing, in a very easy to understand manner. You will also learn how to integrate Storm with other well-known Big Data technologies such as HBase, Redis, Kafka, and Hadoop to realize the full potential of Storm. With real-world examples and clear explanations, this book will ensure you will have a thorough mastery of Apache Storm. You will be able to use this knowledge to develop efficient, distributed real-time applications to cater to your business needs. Style and approach This easy-to-follow guide is full of examples and real-world applications to help you get an in-depth understanding of Apache Storm. This book covers the basics thoroughly and also delves into the intermediate and slightly advanced concepts of application development with Apache Storm.

Introduction to Apache Flink - Ellen Friedman 2016-10-19

There’s growing interest in learning how to analyze streaming data in large-scale systems such as web traffic, financial transactions, machine logs, industrial sensors, and many others. But analyzing data streams at scale has been difficult to do well—until now. This practical book delivers a deep introduction to Apache Flink, a highly innovative open source stream processor with a surprising range of capabilities. Authors Ellen Friedman and Kostas Tzoumas show technical and

nontechnical readers alike how Flink is engineered to overcome significant tradeoffs that have limited the effectiveness of other approaches to stream processing. You'll also learn how Flink has the ability to handle both stream and batch data processing with one technology. Learn the consequences of not doing streaming well—in retail and marketing, IoT, telecom, and banking and finance Explore how to design data architecture to gain the best advantage from stream processing Get an overview of Flink's capabilities and features, along with examples of how companies use Flink, including in production Take a technical dive into Flink, and learn how it handles time and stateful computation Examine how Flink processes both streaming (unbounded) and batch (bounded) data without sacrificing performance

Big Data Analytics - Venkat Ankam 2016-09-28

A handy reference guide for data analysts and data scientists to help to obtain value from big data analytics using Spark on Hadoop clusters About This Book This book is based on the latest 2.0 version of Apache Spark and 2.7 version of Hadoop integrated with most commonly used tools. Learn all Spark stack components including latest topics such as DataFrames, DataSets, GraphFrames, Structured Streaming, DataFrame based ML Pipelines and SparkR. Integrations with frameworks such as HDFS, YARN and tools such as Jupyter, Zeppelin, NiFi, Mahout, HBase Spark Connector, GraphFrames, H2O and Hivemall. Who This Book Is For Though this book is primarily aimed at data analysts and data scientists, it will also help architects, programmers, and practitioners. Knowledge of either Spark or Hadoop would be beneficial. It is assumed that you have basic programming background in Scala, Python, SQL, or R programming with basic Linux experience. Working

experience within big data environments is not mandatory. What You Will Learn Find out and implement the tools and techniques of big data analytics using Spark on Hadoop clusters with wide variety of tools used with Spark and Hadoop Understand all the Hadoop and Spark ecosystem components Get to know all the Spark components: Spark Core, Spark SQL, DataFrames, DataSets, Conventional and Structured Streaming, MLLib, ML Pipelines and Graphx See batch and real-time data analytics using Spark Core, Spark SQL, and Conventional and Structured Streaming Get to grips with data science and machine learning using MLLib, ML Pipelines, H2O, Hivemall, Graphx, SparkR and Hivemall. In Detail Big Data Analytics book aims at providing the fundamentals of Apache Spark and Hadoop. All Spark components – Spark Core, Spark SQL, DataFrames, Data sets, Conventional Streaming, Structured Streaming, MLlib, Graphx and Hadoop core components – HDFS, MapReduce and Yarn are explored in greater depth with implementation examples on Spark + Hadoop clusters. It is moving away from MapReduce to Spark. So, advantages of Spark over MapReduce are explained at great depth to reap benefits of in-memory speeds. DataFrames API, Data Sources API and new Data set API are explained for building Big Data analytical applications. Real-time data analytics using Spark Streaming with Apache Kafka and HBase is covered to help building streaming applications. New Structured streaming concept is explained with an IOT (Internet of Things) use case. Machine learning techniques are covered using MLLib, ML Pipelines and SparkR and Graph Analytics are covered with GraphX and GraphFrames components of Spark. Readers will also get an opportunity to get started with web based notebooks such as Jupyter, Apache Zeppelin and data flow tool Apache NiFi to analyze and visualize data. Style and

approach This step-by-step pragmatic guide will make life easy no matter what your level of experience. You will deep dive into Apache Spark on Hadoop clusters through ample exciting real-life examples. Practical tutorial explains data science in simple terms to help programmers and data analysts get started with Data Science

Mastering Kafka Streams and Ksqldb - Mitch Seymour 2021-04-13

Working with unbounded and fast-moving data streams has historically been difficult. But with Kafka Streams and ksqlDB, building stream processing applications is easy and fun. This practical guide explores the world of real-time data systems through the lens of these popular technologies and explains important stream processing concepts against a backdrop of interesting business problems. Mitch Seymour, senior data systems engineer at Mailchimp, introduces you to both Kafka Streams and ksqlDB so that you can choose the best tool for each unique stream processing project. Non-Java developers will find the ksqlDB path to be an especially gentle introduction to stream processing. In this book, you'll learn: Basic and advanced uses of Kafka Streams and ksqlDB How to transform, enrich, and process event streams How to build both stateless and stateful stream processing applications The different notions of time and the role it plays in stream processing How to to build event-driven microservices on top of continuous event streams Features, operational characteristics, deployment patterns, and configuration tips for both technologies

Beginning Apache Spark 2 - Hien Luu 2018-08-16

Develop applications for the big data landscape with Spark and Hadoop. This book also explains the role of Spark in developing scalable machine learning and analytics applications with Cloud technologies. Beginning Apache Spark 2 gives you an introduction to Apache Spark and shows you how to work

with it. Along the way, you'll discover resilient distributed datasets (RDDs); use Spark SQL for structured data; and learn stream processing and build real-time applications with Spark Structured Streaming. Furthermore, you'll learn the fundamentals of Spark ML for machine learning and much more. After you read this book, you will have the fundamentals to become proficient in using Apache Spark and know when and how to apply it to your big data applications. What You Will Learn Understand Spark unified data processing platform How to run Spark in Spark Shell or Databricks Use and manipulate RDDs Deal with structured data using Spark SQL through its operations and advanced functions Build real-time applications using Spark Structured Streaming Develop intelligent applications with the Spark Machine Learning library Who This Book Is For Programmers and developers active in big data, Hadoop, and Java but who are new to the Apache Spark platform.

Practical Apache Spark - Subhashini Chellappan 2018-12-12

Work with Apache Spark using Scala to deploy and set up single-node, multi-node, and high-availability clusters. This book discusses various components of Spark such as Spark Core, DataFrames, Datasets and SQL, Spark Streaming, Spark MLib, and R on Spark with the help of practical code snippets for each topic. Practical Apache Spark also covers the integration of Apache Spark with Kafka with examples. You'll follow a learn-to-do-by-yourself approach to learning – learn the concepts, practice the code snippets in Scala, and complete the assignments given to get an overall exposure. On completion, you'll have knowledge of the functional programming aspects of Scala, and hands-on expertise in various Spark components. You'll also become familiar with machine learning algorithms with real-time usage. What You Will

LearnDiscover the functional programming features of Scala Understand the complete architecture of Spark and its componentsIntegrate Apache Spark with Hive and Kafka Use Spark SQL, DataFrames, and Datasets to process data using traditional SQL queries Work with different machine learning concepts and libraries using Spark's MLlib packages Who This Book Is For Developers and professionals who deal with batch and stream data processing.