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

This book presents a selection of papers from the 2017 World Conference on Information Systems and Technologies (WorldCIST'17), held between the 11st and 13th of April 2017 at Porto Santo Island, Madeira, Portugal. WorldCIST is a global forum for researchers and practitioners to present and discuss recent results and innovations, current trends, professional experiences and challenges involved in modern Information Systems and Technologies research, together with technological developments and applications. The main topics covered are: Information and Knowledge Management; Organizational Models and Information Systems; Software and Systems Modeling; Software Systems, Architectures, Applications and Tools; Multimedia Systems and Applications; Computer Networks, Mobility and Pervasive Systems; Intelligent and Decision Support Systems; Big Data Analytics and Applications; Human–Computer Interaction; Ethics, Computers & Security; Health Informatics; Information Technologies in Education; and Information Technologies in Radiocommunications.
In the data stream scenario, input arrives very rapidly and there is limited memory to store the input. Algorithms

have to work with one or few passes over the data, space less than linear in the input size or time significantly less than the input size. In the past few years, a new theory has emerged for reasoning about algorithms that work within these constraints on space, time, and number of passes. Some of the methods rely on metric embeddings, pseudo-random computations, sparse approximation theory and communication complexity. The applications for this scenario include IP network traffic analysis, mining text message streams and processing massive data sets in general. Researchers in Theoretical Computer Science, Databases, IP Networking and Computer Systems are working on the data stream challenges.

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 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

This practical guide takes a hands-on approach to implementation and associated methodologies to have you up and running with all that Amazon Kinesis has to offer. You'll work with use cases and practical examples to be able to ingest, process, analyze, and stream real-time data in no time.

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.

This book constitutes the refereed proceedings of the 14th International Conference on Software Architecture, ECSA 2020, held in A'quila, Italy, in September 2020. In the Research Track, 12 full papers presented together with 5 short papers were carefully reviewed and selected from 103 submissions. They are organized in topical

sections as follows: microservices; uncertainty; self-adaptive, and open systems; model-based approaches; performance and security engineering; architectural smells and source code analysis; education and training; experiences and learnings from industrial case studies; and architecting contemporary distributed systems. In the Industrial Track, 11 submissions were received and 6 were accepted to form part of these proceedings. In addition the book contains 3 keynote talks. Due to the Corona pandemic ECSA 2020 was held as an virtual event.

Written to be tool-agnostic, Grokking Streaming Systems helps you unravel what streaming systems are, how they work, and whether they’re right for your business. Every action by a user or a system process generates valuable data for your application or organization. Streaming systems capture and process these events, turning disconnected bits into coherent, useful information sources. Grokking Streaming Systems is a simple guide to the complex concepts you need to start building your own streaming systems. Written to be tool-agnostic, Grokking Streaming Systems helps you unravel what streaming systems are, how they work, and whether they’re right for your business. You’ll start with the key concepts and then work your way through increasingly complex examples. You’ll even be able to easily experiment with your own streaming system. By the time you’re done, you’ll be able to easily assess the capabilities of streaming frameworks, and solve common challenges that arise when building streaming systems. Purchase of the print book includes a free eBook in PDF,

Kindle, and ePub formats from Manning Publications. 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

This book constitutes the proceedings of the 10th International Conference on Advanced Data Mining and Applications, ADMA 2014, held in Guilin, China during December 2014. The 48 regular papers and 10 workshop papers presented in this volume were carefully reviewed and selected from 90 submissions. They deal with the following topics: data mining, social network and social media, recommend systems, database, dimensionality reduction, advance machine learning techniques, classification, big data and applications, clustering methods, machine learning, and data mining and database.

A hands-on approach to tasks and techniques in data stream mining and real-time analytics, with examples in MOA, a popular freely available open-source software framework. Today many information sources—including sensor networks, financial markets, social networks, and healthcare monitoring—are so-called data streams, arriving sequentially and at high speed. Analysis must take

place in real time, with partial data and without the capacity to store the entire data set. This book presents algorithms and techniques used in data stream mining and real-time analytics. Taking a hands-on approach, the book demonstrates the techniques using MOA (Massive Online Analysis), a popular, freely available open-source software framework, allowing readers to try out the techniques after reading the explanations. The book first offers a brief introduction to the topic, covering big data mining, basic methodologies for mining data streams, and a simple example of MOA. More detailed discussions follow, with chapters on sketching techniques, change, classification, ensemble methods, regression, clustering, and frequent pattern mining. Most of these chapters include exercises, an MOA-based lab session, or both. Finally, the book discusses the MOA software, covering the MOA graphical user interface, the command line, use of its API, and the development of new methods within MOA. The book will be an essential reference for readers who want to use data stream mining as a tool, researchers in innovation or data stream mining, and programmers who want to create new algorithms for MOA.

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 layerNew messaging technologies, including Apache Kafka and MapR Streams, with links to sample codeTechnology choices for streaming analytics: Apache Spark Streaming, Apache Flink, Apache Storm, and Apache ApexHow stream-based architectures are helpful to support microservicesSpecific 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.

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.

Design and administer fast, reliable enterprise messaging systems with Apache KafkaAbout 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 examplesWho This Book Is ForIf 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 bookWhat 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 practicesIn DetailApache 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 approachA step-by -step, comprehensive guide filled with practical and real- world examples

Since the beginning of the Internet age and the increased use of ubiquitous computing devices, the large volume and continuous flow of distributed data have imposed new constraints on the design of learning algorithms. Exploring how to extract knowledge structures from evolving and time-changing data, Knowledge Discovery from Data Streams presents a coherent overview of state-of-the-art research in learning from data streams. The book covers the fundamentals that are imperative to understanding data streams and describes important applications, such as TCP/IP traffic, GPS data,

sensor networks, and customer click streams. It also addresses several challenges of data mining in the future, when stream mining will be at the core of many applications. These challenges involve designing useful and efficient data mining solutions applicable to real-world problems. In the appendix, the author includes examples of publicly available software and online data sets. This practical, up-to-date book focuses on the new requirements of the next generation of data mining. Although the concepts presented in the text are mainly about data streams, they also are valid for different areas of machine learning and data mining.

"Spark is one of todays most popular distributed computation engines for processing and analyzing big data. This course provides data engineers, data scientist and data analysts interested in exploring the technology of data streaming with practical experience in using Spark. You'll learn about the Spark Structured Streaming API, the powerful Catalyst query optimizer, the Tungsten execution engine, and more in this hands-on course where youll build small several applications that leverage all the aspects of Spark 2.0. While not a requirement, the course works best for those with some Scala experience."--Resource description page.

This volume focuses on the theory and practice of data stream management, and the novel challenges this emerging domain poses for data-management

algorithms, systems, and applications. The collection of chapters, contributed by authorities in the field, offers a comprehensive introduction to both the algorithmic/theoretical foundations of data streams, as well as the streaming systems and applications built in different domains. A short introductory chapter provides a brief summary of some basic data streaming concepts and models, and discusses the key elements of a generic stream query processing architecture. Subsequently, Part I focuses on basic streaming algorithms for some key analytics functions (e.g., quantiles, norms, join aggregates, heavy hitters) over streaming data. Part II then examines important techniques for basic stream mining tasks (e.g., clustering, classification, frequent itemsets). Part III discusses a number of advanced topics on stream processing algorithms, and Part IV focuses on system and language aspects of data stream processing with surveys of influential system prototypes and language designs. Part V then presents some representative applications of streaming techniques in different domains (e.g., network management, financial analytics). Finally, the volume concludes with an overview of current data streaming products and new application domains (e.g. cloud computing, big data analytics, and complex event processing), and a discussion of future directions in this exciting field. The book provides a comprehensive overview of

core concepts and technological foundations, as well as various systems and applications, and is of particular interest to students, lecturers and researchers in the area of data stream management. This book primarily discusses issues related to the mining aspects of data streams and it is unique in its primary focus on the subject. This volume covers mining aspects of data streams comprehensively: each contributed chapter contains a survey on the topic, the key ideas in the field for that particular topic, and future research directions. The book is intended for a professional audience composed of researchers and practitioners in industry. This book is also appropriate for advanced-level students in computer science.

Construct a robust end-to-end solution for analyzing andvisualizing streaming data Real-time analytics is the hottest topic in data analyticstoday. In Real-Time Analytics: Techniques to Analyze andVisualize Streaming Data, expert Byron Ellis teaches dataanalysts technologies to build an effective real-time analyticsplatform. This platform can then be used to make sense of theconstantly changing data that is beginning to outpace traditionalbatch-based analysis platforms. The author is among a very few leading experts in the field. Hehas a prestigious background in research, development, analytics,real-time visualization, and Big Data streaming and is uniquelyqualified to help you explore this revolutionary field. Moving froma description of the overall analytic architecture of real-

timeanalytics to using specific tools to obtain targeted results,Real-Time Analytics leverages open source and moderncommercial tools to construct robust, efficient systems that canprovide real-time analysis in a cost-effective manner. The bookincludes: A deep discussion of streaming data systems andarchitectures Instructions for analyzing, storing, and delivering streamingdata Tips on aggregating data and working with sets Information on data warehousing options and techniques Real-Time Analytics includes in-depth case studies forwebsite analytics, Big Data, visualizing streaming and mobile data,and mining and visualizing operational data flows. The book's"recipe" layout lets readers quickly learn and implement differenttechniques. All of the code examples presented in the book, alongwith their related data sets, are available on the companionwebsite.

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

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

Tips, techniques, and trends on how to use dashboard technology tooptimize business performance Business performance management is a hot new management disciplinethat delivers tremendous value when supported by informationtechnology. Through case studies and industry research, this bookshows how leading companies are using performance dashboards toexecute strategy, optimize business processes, and improveperformance. Wayne W. Eckerson (Hingham, MA) is the Director of Research for TheData Warehousing Institute (TDWI), the leading association ofbusiness intelligence and data warehousing professionals worldwidethat provide high-quality, in-depth education, training, andresearch. He is a columnist for SearchCIO.com, DM Review,Application Development Trends, the Business Intelligence Journal,and TDWI Case Studies & Solution.

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

Building Data Streaming Applications with Apache KafkaPackt Publishing Ltd

Many of the technologies discussed in the book - Spark, Storm, Kafka, Impala, RabbitMQ, etc. - are covered individually in other books. Throughout this book, readers will get a clear picture of how these technologies work individually and together, gain insight on how to choose the correct technologies, and discover how to fuse them together to architect a robust system. Streaming Data introduces the concepts and requirements of streaming and real-time data systems. Readers will develop a foundation to understand the

challenges and solutions of building in-the-moment data systems before committing to specific technologies. Using lots of diagrams, this book systematically builds up the blueprint for an in-the-moment system concept by concept. This book focuses on the big ideas of streaming and real time data systems rather than the implementation details. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications.

Big data modeling is very challenging to handle using traditional database modeling and management systems. This book will teach you how to model big data using the latest and more efficient tools such as ERWIN, ANACONDA (Python), and WEKA to model data.

A comprehensive guide to mastering the most advanced Hadoop 3 concepts Key Features Get to grips with the newly introduced features and capabilities of Hadoop 3 Crunch and process data using MapReduce, YARN, and a host of tools within the Hadoop ecosystem Sharpen your Hadoop skills with real-world case studies and code Book 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 learn Gain an in-depth understanding of distributed computing using Hadoop 3 Develop enterprise-grade applications using Apache Spark, Flink, and more Build scalable and high-performance Hadoop data pipelines with security, monitoring, and data governance Explore batch data processing patterns and how to model data in Hadoop Master best practices for enterprises using, or planning to use, Hadoop 3 as a data platform Understand security aspects of Hadoop, including authorization and authentication Who 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.
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
Get well-versed with FastAPI features and best practices for testing, monitoring, and deployment to run high-quality and robust data science applications Key Features Cover the concepts of the FastAPI framework, including aspects relating to asynchronous programming, type hinting, and dependency injection Develop efficient RESTful APIs for data science with modern Python Build, test, and deploy high performing data science and machine learning systems with FastAPI Book

Description FastAPI is a web framework for building APIs with Python 3.6 and its later versions based on standard Python-type hints. With this book, you'll be able to create fast and reliable data science API backends using practical examples. This book starts with the basics of the FastAPI framework and associated modern Python programming language concepts. You'll be taken through all the aspects of the framework, including its powerful dependency injection system and how you can use it to communicate with databases, implement authentication and integrate machine learning models. Later, you'll cover best practices relating to testing and deployment to run a high-quality and robust application. You'll also be introduced to the extensive ecosystem of Python data science packages. As you progress, you'll learn how to build data science applications in Python using FastAPI. The book also demonstrates how to develop fast and efficient machine learning prediction backends and test them to achieve the best performance. Finally, you'll see how to implement a real-time face detection system using WebSockets and a web browser as a client. By the end of this FastAPI book, you'll have not only learned how to implement Python in data science projects but also how to maintain and design them to meet high programming standards with the help of FastAPI. What you will learn Explore the basics of modern Python and async I/O programming Get to grips with basic and advanced concepts of the FastAPI framework Implement a FastAPI dependency to efficiently run a machine learning model Integrate a simple face detection algorithm in a FastAPI backend Integrate common Python data science libraries in a web backend Deploy a performant and reliable web backend for a data science application Who this book is for This Python data science book is for data scientists and software developers interested in gaining knowledge of FastAPI and its ecosystem to build data science applications. Basic

knowledge of data science and machine learning concepts and how to apply them in Python is recommended.
Data has cemented itself as a building block of daily life. However, surrounding oneself with great quantities of information heightens risks to one's personal privacy. Additionally, the presence of massive amounts of information prompts researchers into how best to handle and disseminate it. Research is necessary to understand how to cope with the current technological requirements. Large-Scale Data Streaming, Processing, and Blockchain Security is a collection of innovative research that explores the latest methodologies, modeling, and simulations for coping with the generation and management of large-scale data in both scientific and individual applications. Featuring coverage on a wide range of topics including security models, internet of things, and collaborative filtering, this book is ideally designed for entrepreneurs, security analysts, IT consultants, security professionals, programmers, computer technicians, data scientists, technology developers, engineers, researchers, academicians, and students.
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

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

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

Data is at the center of many challenges in system design today. Difficult issues need to be figured out, such as scalability, consistency, reliability, efficiency, and maintainability. In addition, we have an overwhelming variety of tools, including relational databases, NoSQL datastores, stream or batch processors, and message brokers. What are the right choices for your application? How do you make sense of all these buzzwords? In this practical and comprehensive guide, author Martin Kleppmann helps you navigate this diverse landscape by examining the pros and cons of various technologies for processing and storing data. Software keeps changing, but the fundamental principles remain the same. With this book, software engineers and architects will learn how to apply those ideas in practice, and how to make full use of data in modern applications. Peer under the hood of the systems you already use, and learn how to use and operate them more effectively Make informed

decisions by identifying the strengths and weaknesses of different tools Navigate the trade-offs around consistency, scalability, fault tolerance, and complexity Understand the distributed systems research upon which modern databases are built Peek behind the scenes of major online services, and learn from their architectures

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.

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

Software development today is embracing functional programming (FP), whether it's for writing concurrent programs or for managing Big Data. Where does that leave Java developers? This concise book offers a pragmatic, approachable introduction to FP for Java developers or anyone who uses an object-oriented language. Dean Wampler, Java expert and author of Programming Scala (O'Reilly), shows you how to apply FP principles such as immutability, avoidance of side-effects, and higher-order functions to your Java code. Each chapter provides exercises to help you practice what you've learned. Once you grasp the benefits of functional programming, you'll discover that it improves all of the code you write. Learn basic FP principles and apply them to object-oriented programming Discover how FP is more concise and modular than OOP Get useful FP lessons for your Java type design—such as avoiding nulls Design data structures and algorithms using functional

programming principles Write concurrent programs using the Actor model and software transactional memory Use functional libraries and frameworks for Java—and learn where to go next to deepen your functional programming skills

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.

In this IBM® Redbooks® publication, we discuss and describe the positioning, functions, capabilities, and advanced programming techniques for IBM InfoSphereTM Streams (V2), a new paradigm and key component of IBM Big Data platform. Data has traditionally been stored in files or databases, and then analyzed by queries and applications. With stream computing, analysis is performed moment by moment as the data is in motion. In fact, the data might never be stored (perhaps only the analytic results). The ability to analyze data in motion is called real-time analytic processing (RTAP). IBM InfoSphere Streams takes a fundamentally different approach to Big Data analytics and differentiates itself with its distributed runtime platform, programming model, and tools for developing and debugging analytic applications that have a high volume and variety of data types. Using in-memory techniques and analyzing record by record enables high velocity. Volume, variety and velocity are the key attributes of Big Data. The data streams that are

consumable by IBM InfoSphere Streams can originate from sensors, cameras, news feeds, stock tickers, and a variety of other sources, including traditional databases. It provides an execution platform and services for applications that ingest, filter, analyze, and correlate potentially massive volumes of continuous data streams. This book is intended for professionals that require an understanding of how to process high volumes of streaming data or need information about how to implement systems to satisfy those requirements. See: http://www.redbooks.ibm.com/abstracts/sg247865.html for the IBM InfoSphere Streams (V1) release.

Build data-intensive applications locally and deploy at scale using the combined powers of Python and Spark 2.0 About This Book Learn why and how you can efficiently use Python to process data and build machine learning models in Apache Spark 2.0 Develop and deploy efficient, scalable real-time Spark solutions Take your understanding of using Spark with Python to the next level with this jump start guide Who This Book Is For If you are a Python developer who wants to learn about the Apache Spark 2.0 ecosystem, this book is for you. A firm understanding of Python is expected to get the best out of the book. Familiarity with Spark would be useful, but is not mandatory. What You Will Learn Learn about Apache Spark and the Spark 2.0 architecture Build and interact with Spark DataFrames using Spark SQL Learn how to solve graph and deep learning problems using GraphFrames and TensorFrames respectively Read, transform, and understand data and use it to train machine learning models Build machine learning models with MLlib and ML Learn how to submit your applications programmatically using spark-submit Deploy locally built applications to a cluster In Detail Apache Spark is an open source framework for efficient cluster computing with a strong interface for data parallelism and fault tolerance. This book will show you how to leverage the

power of Python and put it to use in the Spark ecosystem. You will start by getting a firm understanding of the Spark 2.0 architecture and how to set up a Python environment for Spark. You will get familiar with the modules available in PySpark. You will learn how to abstract data with RDDs and DataFrames and understand the streaming capabilities of PySpark. Also, you will get a thorough overview of machine learning capabilities of PySpark using ML and MLlib, graph processing using GraphFrames, and polyglot persistence using Blaze. Finally, you will learn how to deploy your applications to the cloud using the spark-submit command. By the end of this book, you will have established a firm understanding of the Spark Python API and how it can be used to build data-intensive applications. Style and approach This book takes a very comprehensive, step-by-step approach so you understand how the Spark ecosystem can be used with Python to develop efficient, scalable solutions. Every chapter is standalone and written in a very easy-to-understand manner, with a focus on both the hows and the whys of each concept.

Learn about the fastest-growing open source project in the world, and find out how it revolutionizes big data analytics About This Book Exclusive guide that covers how to get up and running with fast data processing using Apache Spark Explore and exploit various possibilities with Apache Spark using real-world use cases in this book Want to perform efficient data processing at real time? This book will be your one-stop solution. Who This Book Is For This guide appeals to big data engineers, analysts, architects, software engineers, even technical managers who need to perform efficient data processing on Hadoop at real time. Basic familiarity with Java or Scala will be helpful. The assumption is that readers will be from a mixed background, but would be typically people with background in engineering/data science

with no prior Spark experience and want to understand how Spark can help them on their analytics journey. What You Will Learn Get an overview of big data analytics and its importance for organizations and data professionals Delve into Spark to see how it is different from existing processing platforms Understand the intricacies of various file formats, and how to process them with Apache Spark. Realize how to deploy Spark with YARN, MESOS or a Stand-alone cluster manager. Learn the concepts of Spark SQL, SchemaRDD, Caching and working with Hive and Parquet file formats Understand the architecture of Spark MLLib while discussing some of the off-the-shelf algorithms that come with Spark. Introduce yourself to the deployment and usage of SparkR. Walk through the importance of Graph computation and the graph processing systems available in the market Check the real world example of Spark by building a recommendation engine with Spark using ALS. Use a Telco data set, to predict customer churn using Random Forests. In Detail Spark juggernaut keeps on rolling and getting more and more momentum each day. Spark provides key capabilities in the form of Spark SQL, Spark Streaming, Spark ML and Graph X all accessible via Java, Scala, Python and R. Deploying the key capabilities is crucial whether it is on a Standalone framework or as a part of existing Hadoop installation and configuring with Yarn and Mesos. The next part of the journey after installation is using key components, APIs, Clustering, machine learning APIs, data pipelines, parallel programming. It is important to understand why each framework component is key, how widely it is being used, its stability and pertinent use cases. Once we understand the individual components, we will take a couple of real life advanced analytics examples such as 'Building a Recommendation system', 'Predicting customer churn' and so on. The objective of these real life examples is to give the reader confidence of using Spark for

real-world problems. Style and approach With the help of practical examples and real-world use cases, this guide will take you from scratch to building efficient data applications using Apache Spark. You will learn all about this excellent data processing engine in a step-by-step manner, taking one aspect of it at a time. This highly practical guide will include how to work with data pipelines, dataframes, clustering, SparkSQL, parallel programming, and such insightful topics with the help of real-world use cases.