Learning Apache Kafka Second Edition

The book will follow a step-by-step tutorial approach which will show the readers how to use Apache Kafka for messaging from scratch.Apache Kafka is for readers with software development experience, but no prior exposure to Apache Kafka or similar technologies is assumed. This book is also for enterprise application developers and big data enthusiasts who have worked with other publisher-subscriber based systems and now want to explore Apache Kafka as a futuristic scalable solution.

Understand how to apply auto machine learning to data streams and create transactional machine learning (TML) solutions that are frictionless (require minimal to no human intervention) and elastic (machine learning solutions that can scale up or down by controlling the number of data streams, algorithms, and users of the insights). This book will strengthen your knowledge of the inner workings of TML solutions using data streams with auto machine learning integrated with Apache Kafka. Transactional Machine Learning with Data Streams and AutoML introduces the industry challenges with applying machine learning to data streams. You will learn the framework that will help you in choosing business problems that are best suited for TML. You will also see how to measure the business value of TML solutions. You will then learn the technical components of TML solutions, including the reference and technical architecture of a TML solution. This book also presents a TML solution template that will make it easy for you to quickly start building your own TML solutions. Specifically, you are given access to a TML Python library and integration technologies for download. You will also learn how TML will evolve in the future, and the growing need by organizations for deeper insights from data streams. By the end of the book, you will have a solid understanding of TML. You will know how to build TML solutions with all the necessary details, and all the resources at your fingertips. What You Will Learn Discover transactional machine learning Measure the business value of TML Choose TML use cases Design technical architecture of TML solutions with Apache Kafka Work with the technologies used to build TML solutions Build transactional machine learning solutions with hands-on code together with Apache Kafka in the cloud Who This Book Is For Data scientists, machine learning engineers and architects, and AI and machine learning business leaders.

Build a scalable, fault-tolerant and highly available data layer for your applications using Apache CassandraAbout This Book* Install Cassandra and use it to set up multi-node clusters* Design rich schemas that capture the relationships between different data types* Master the advanced features available in Cassandra 3.x through a step-by-step tutorial and build a scalable, high performance database layerWho This Book Is ForIf you are a first-time user of Apache Cassandra who wants to learn the basic of it, as well as some not-so-basic features, this book is for you. It does not assume any prior experience in coding or any framework.What you will learn* Install Cassandra and create your first keyspace* Create tables with multiple clustering columns to organize related data* Use secondary indexes and materialized views to avoid denormalization of data* Effortlessly handle concurrent updates with collection columns* Ensure data integrity with lightweight transactions and logged batches* Understand eventual consistency and use the right consistency level for your situation* Understand data distribution with Cassandra and get to know ways to implement application-level optimizationsIn DetailCassandra is a distributed database that stands out thanks to its robust feature set and intuitive interface, while still providing the high availability and scalability of a distributed store. This book will introduce you to the rich features offered by Cassandra, and empower you to create and manage a highly performant, fault-tolerant database layer.The book starts by explaining the new features implemented in Cassandra 3.x, you'll see how to install Cassandra, and you'll understand Lightweight Transactions. Next you'll learn to create tables with composite partition keys, and get to know different methods to avoid denormalization of data. You will then proceed to create user-defined functions and data distribution in Cassandra. Finally, you will set up a multi node cluster and implement application-level optimization using a Java client.By the end of this book, you'll be fully equipped to build powerful, scalable Cassandra database layers for your applications.

Summary Camel in Action, Second Edition is the most complete Camel book on the market. Written by core developers of Camel and the authors of the highly acclaimed first edition, this book distills their experience and practical insights so that you can tackle integration tasks like a pro. Forewords by James Strachan and Dr. Mark Little Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Apache Camel is a Java framework that implements enterprise integration patterns (EIPs) and comes with over 200 adapters to third-party systems. A concise DSL lets you build integration logic into your app with just a few lines of Java or XML. By using Camel, you benefit from the testing and experience of a large and vibrant open source community. About the Book Camel in Action, Second Edition is the definitive guide to the Camel framework. It starts with core concepts like sending, receiving, routing, and transforming data. It then goes in depth on many topics such as how to develop, debug, test, deal with errors, secure, scale, cluster, deploy, and monitor your Camel applications. The book also discusses how to run Camel with microservices, reactive systems, containers, and in the cloud. What's Inside Coverage of all relevant EIPs Camel microservices with Spring Boot Camel on Docker and Kubernetes Error handling, testing, security, clustering, monitoring, and deployment Hundreds of examples in Java and XML About the Reader Readers should be familiar with Java. This book is accessible to beginners and invaluable to experts. About the Author Claus Ibsen is a senior principal engineer working for Red Hat specializing in cloud and integration. He has worked on Apache Camel for the last nine years where he heads the project. Claus lives in Denmark. Jonathan Anstey is an engineering manager at Red Hat and a core Camel contributor. He lives in Newfoundland, Canada. Table of Contents Part 1 - First steps Meeting Camel Routing with Camel Part 2 - Core Camel Transforming data with Camel Using beans with Camel Enterprise integration patterns Using components Part 3 - Developing and testing Microservices Developing Camel projects Testing RESTful web services Part 4 - Going further with Camel Error handling Transactions and idempotency Parallel processing Securing Camel Part 5 - Running and managing Camel Running and deploying Camel Management and monitoring Part 6 - Out in the wild Clustering Microservices with Docker and Kubernetes Camel tooling Bonus online chapters Available at https://www.manning.com/books/camel-in-?action-second-edition and in electronic versions of this book: Reactive Camel Camel and the IoT by Henryk Konsek

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

Transactional Machine Learning with Data Streams and AutoML

Apache Kafka Quick Start Guide

Mastering Kafka Streams and ksqlDB

Learning Apache OpenWhisk

Agile Data Science 2.0

Camel in Action

Build Frictionless and Elastic Machine Learning Solutions with Apache Kafka in the Cloud Using Python

Data in all domains is getting bigger. How can you work with it efficiently? Recently updated for Spark 1.3, this book introduces Apache Spark, the open source cluster computing system that makes data analytics fast to write and fast to run. With Spark, you can tackle big datasets quickly through simple APIs in Python, Java, and Scala. This edition includes new information on Spark SQL, Spark Streaming, setup, and Maven coordinates. Written by the developers of Spark, this book will have data scientists and engineers up and running in no time. You'll learn how to express parallel jobs with just a few lines of code, and cover applications from simple batch jobs to stream processing and machine learning. Quickly dive into Spark capabilities such as distributed datasets, in-memory caching, and the interactive shell Leverage Spark's powerful built-in libraries, including Spark SQL, Spark Streaming, and MLlib Use one programming paradigm instead of mixing and matching tools like Hive, Hadoop, Mahout, and Storm Learn how to deploy interactive, batch, and streaming applications Connect to data sources including HDFS, Hive, JSON, and S3 Master advanced topics like data partitioning and shared variables

Learning Apache Kafka Second EditionPackt Pub Limited

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

Definitive guide to lightning fast data processing for distributed systems with Apache FlinkAbout This Book* Build your expertise in processing realtime data with Apache Flink and its ecosystem* Gain insights into the working of all components of Apache Flink such as FlinkML, Gelly, and Table APIFilled with real world use cases,* Your guide to take advantage of Apache Flink for solving real world problemsWho This Book Is ForBig data developers who are looking to process batch and real-time data on distributed systems. Basic knowledge of Hadoop and big data is assumed. Reasonable knowledge of Java or Scala is expected.What You Will Learn* Learn how to build end to end real time analytics projects* Integrate with existing big data stack and utilize existing infrastructure.* Build predictive analytics applications using FlinkML* Use graph library to perform graph querying and search.In DetailWith the advent of massive computer systems, organizations in different domains generate large amounts of data at a realtime basis. The latest entrant to big data processing, Apache Flink, is designed to process continuous streams of data at a lightning fast pace.This book will be your definitive guide to batch and stream data processing with Apache Flink. The book begins with introducing the Apache Flink ecosystem, setting it up and using the DataSet and DataStream API for processing batch and streaming datasets. Bringing the power of SQL to Flink, this book will then explore the Table API for querying and manipulating data. In the latter half of the book, readers will get to learn the remaining ecosystem of Apache Flink to achieve complex tasks such as event processing, machine learning, and graph processing. The final part of the book would consist of topics such as scaling Flink solutions, performance optimization and integrating Flink with other tools such as ElasticSearch.Whether you want to dive deeper into Apache Flink, or want to investigate how to get more out of this powerful technology, you'll find everything inside

Get up to speed with Apache Drill, an extensible distributed SQL query engine that reads massive datasets in many popular file formats such as Parquet, JSON, and CSV. Drill reads data in HDFS or in cloud-native storage such as S3 and works with Hive metastores along with distributed databases such as HBase, MongoDB, and relational databases. Drill works everywhere: on your laptop or in your largest cluster. In this practical book, Drill committers Charles Givre and Paul Rogers show analysts and data scientists how to query and analyze raw data using this powerful tool. Data scientists today spend about 80% of their time just gathering and cleaning data. With this book, you'll learn how Drill helps you analyze data more effectively to drive down time to insight. Use Drill to clean, prepare, and summarize delimited data for further analysis Query file types including logfiles, Parquet, JSON, and other complex formats Query Hadoop, relational databases, MongoDB, and Kafka with standard SQL Connect to Drill programmatically using a variety of languages Use Drill even with challenging or ambiguous file formats Perform sophisticated analysis by extending Drill's functionality with user-defined functions Facilitate data analysis for network security, image metadata, and machine learning

Kafka: The Definitive Guide

Machine Learning

Cloud Native Applications with Ballerina

A guide for programmers interested in developing cloud native applications using Ballerina Swan Lake

Streaming Architecture

Spark: The Definitive Guide

Mastering Apache Pulsar

Summary The Spark distributed data processing platform provides an easy-to-implement tool for ingesting, streaming, and processing data from any source. In Spark in Action, Second Edition, you'll learn to take advantage of Spark's core features and incredible processing speed, with applications including real-time computation, delayed evaluation, and machine learning. Spark skills are a hot commodity in enterprises worldwide, and with Spark's powerful and flexible Java APIs, you can reap all the benefits without first learning Scala or Hadoop. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Analyzing enterprise data starts by reading, filtering, and merging files and streams from many sources. The Spark data processing engine handles this varied volume like a champ, delivering speeds 100 times faster than Hadoop systems. Thanks to SQL support, an intuitive interface, and a straightforward multilanguage API, you can use Spark without learning a complex new ecosystem. About the book Spark in Action, Second Edition, teaches you to create end-to-end analytics applications. In this entirely new book, you'll learn from interesting Java-based examples, including a complete data pipeline for processing NASA satellite data. And you'll discover Java, Python, and Scala code samples hosted on GitHub that you can explore and adapt, plus appendixes that give you a cheat sheet for installing tools and understanding Spark-specific terms. What's inside Writing Spark applications in Java Spark application architecture Ingestion through files, databases, streaming, and Elasticsearch Querying distributed datasets with Spark SQL About the reader This book does not assume previous experience with Spark, Scala, or Hadoop. About the author Jean-Georges Perrin is an experienced data and software architect. He is France's first IBM Champion and has been honored for 12 consecutive years. Table of Contents PART 1 - THE THEORY CRIPPLED BY AWESOME EXAMPLES 1 So, what is Spark, anyway? 2 Architecture and flow 3 The majestic role of the dataframe 4 Fundamentally lazy 5 Building a simple app for deployment 6 Deploying your simple app PART 2 - INGESTION 7 Ingestion from files 8 Ingestion from databases 9 Advanced ingestion: finding data sources and building your own 10 Ingestion through structured streaming PART 3 - TRANSFORMING YOUR DATA 11 Working with SQL 12 Transforming your data 13 Transforming entire documents 14 Extending transformations with user-defined functions 15 Aggregating your data PART 4 - GOING FURTHER 16 Cache and checkpoint: Enhancing Spark's performances 17 Exporting data and building full data pipelines 18 Exploring deployment

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

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

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

Designing and writing a real-time streaming publication with Apache Apex About This Book Get a clear, practical approach to real-time data processing Program Apache Apex streaming applications This book shows you Apex integration with the open source Big Data ecosystem Who This Book Is For This book assumes knowledge of application development with Java and familiarity with distributed systems. Familiarity with other real-time streaming frameworks is not required, but some practical experience with other big data processing utilities might be helpful. What You Will Learn Put together a functioning Apex application from scratch Scale an Apex application and configure it for optimal performance Understand how to deal with failures via the fault tolerance features of the platform Use Apex via other frameworks such as Beam Understand the DevOps implications of deploying Apex In Detail Apache Apex is a next-generation stream processing framework designed to operate on data at large scale, with minimum latency, maximum reliability, and strict correctness guarantees. Half of the book consists of Apex applications, showing you key aspects of data processing pipelines such as connectors for sources and sinks, and common data transformations. The other half of the book is evenly split into explaining the Apex framework, and tuning, testing, and scaling Apex applications. Much of our economic world depends on growing streams of data, such as social media feeds, financial records, data from mobile devices, sensors and machines (the Internet of Things - IoT). The projects in the book show how to process such streams to gain valuable, timely, and actionable insights. Traditional use cases, such as ETL, that currently consume a significant chunk of data engineering resources are also covered. The final chapter shows you future possibilities emerging in the streaming space, and how Apache Apex can contribute to it. Style and approach This book is divided into two major parts: first it explains what Apex is, what its relevant parts are, and how to write well-built Apex applications. The second part is entirely application-driven, walking you through Apex applications of increasing complexity.

Fundamentals, Implementation, and Operation of Streaming Applications

Covers Apache Spark 3 with Examples in Java, Python, and Scala

Big Data Processing Made Simple

Spark in Action

Learning Spark

Apache Flume: Distributed Log Collection for Hadoop - Second Edition

A Guide to Apache Spark, Mesos, Akka, Cassandra, and Kafka

Imagine what you could do if scalability wasn't a problem. With this hands-on guide, you’ll learn how the Cassandra database management system handles hundreds of terabytes of data while remaining highly available across multiple data centers. This expanded second edition—updated for Cassandra 3.0—provides the technical details and practical examples you need to put this database to work in a production environment. Authors Jeff Carpenter and Eben Hewitt demonstrate the advantages of Cassandra’s non-relational design, with special attention to data modeling. If you’re a developer, DBA, or application architect looking to solve a database scaling issue or future-proof your application, this guide helps you harness Cassandra’s speed and flexibility. Understand Cassandra’s distributed and decentralized structure Use the Cassandra Query Language (CQL) and cqlsh—the CQL shell Create a working data model and compare it with an equivalent relational model Develop sample applications using client drivers for languages including Java, Python, and Node.js Explore cluster topology and learn how nodes exchange data Maintain a high level of performance in your cluster Deploy Cassandra on site, in the Cloud, or with Docker Integrate Cassandra with Spark, Hadoop, Elasticsearch, Solr, and Lucene

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.

Ready to build cloud native applications? Get a hands-on introduction to daily life as a developer crafting code on OpenShift, the open source container application platform from Red Hat. Creating and packaging your apps for deployment on modern distributed systems can be daunting. Too often, adding infrastructure value can complicate development. With this practical guide, you'll learn how to build, deploy, and manage a multitiered application on OpenShift. Authors Joshua Wood and Brian Tannous, principal developer advocates at Red Hat, demonstrate how OpenShift speeds application development. With the Kubernetes container orchestrator at its core, OpenShift simplifies and automates the way you build, ship, and run code. You'll learn how to use OpenShift and the Quarkus Java framework to develop and deploy apps using proven enterprise technologies and practices that you can apply to code in any language. Learn the development cycles for building and deploying on OpenShift, and the tools that drive them Use OpenShift to build, deploy, and manage the ongoing lifecycle of an n-tier application Create a continuous integration and deployment pipeline to build and deploy application source code on OpenShift Automate scaling decisions with metrics and trigger lifecycle events with webhooks

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

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

Stream Processing with Apache Spark

Event Data, Stream Processing, and Data Integration

Query and Analyze Distributed Data Sources with SQL

Big Data SMACK

Hadoop Application Architectures

Leverage Apache Kafka 2.0 to simplify real-time data processing for distributed applications

Building Data Streaming Applications with Apache Kafka

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

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

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.

Learn how to build scalable cloud native applications with the new-generation Ballerina language using expert tips and best practices Key FeaturesWork with code samples based on the Ballerina Swan Lake Beta1 versionExplore the in-built networking protocol support in Ballerina to develop secure distributed appsBuild a Ballerina app with an automated CI/CD pipeline with observability to simplify maintenance and deploymentBook Description The Ballerina programming language was created by WSO2 for the modern needs of developers where cloud native development techniques have become ubiquitous. Ballerina simplifies how programmers develop and deploy cloud native distributed apps and microservices. Cloud Native Applications with Ballerina will guide you through Ballerina essentials, including variables, types, functions, flow control, security, and more. You'll explore networking as an in-built feature in Ballerina, which makes it a first-class language for distributed computing. With this app development book, you'll learn about different networking protocols as well as different architectural patterns that you can use to implement services on the cloud. As you advance, you'll explore multiple design patterns used in microservice architecture and use serverless in Amazon Web Services (AWS) and Microsoft Azure platforms. You will also get to grips with Docker, Kubernetes, and serverless platforms to simplify maintenance and the deployment process. Later, you'll focus on the Ballerina testing framework along with deployment tools and monitoring tools to build fully automated observable cloud applications. By the end of this book, you will have learned how to apply the Ballerina language for building scalable, resilient, secured, and easy-to-maintain cloud native Ballerina projects and applications. What you will learnUnderstand the concepts and models in cloud native architectureGet to grips with the high-level concepts of building applications with the Ballerina languageUse cloud native architectural design patterns to develop cloud native Ballerina applicationsDiscover how to automate, maintain, and observe cloud native Ballerina applicationsUse a container to deploy and maintain a Ballerina application with Docker and KubernetesExplore serverless architecture and use Microsoft Azure and the AWS platform to build serverless applicationsWho this book is for This Ballerina Swan Lake book is for cloud developers, integration developers, and microservices developers who are facing challenges with legacy tooling and are looking for the latest tools and technologies to solve them. Beginner-level programming knowledge is required before getting started with this Ballerina book.

Speed up the design and implementation of deep learning solutions using Apache Spark Key FeaturesExplore the world of distributed deep learning with Apache SparkTrain neural networks with deep learning libraries such as BigDL and TensorFlowDevelop Spark deep learning applications to intelligently handle large and complex datasetsBook Description Deep learning is a subset of machine learning where datasets with several layers of complexity can be processed. Hands-On Deep Learning with Apache Spark addresses the sheer complexity of technical and analytical parts and the speed at which deep learning solutions can be implemented on Apache Spark. The book starts with the fundamentals of Apache Spark and deep learning. You will set up Spark for deep learning, learn principles of distributed modeling, and understand different types of neural nets. You will then implement deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) on Spark. As you progress through the book, you will gain hands-on experience of what it takes to understand the complex datasets you are dealing with. During the course of this book, you will use popular deep learning frameworks, such as TensorFlow, Deeplearning4j, and Keras to train your distributed models. By the end of this book, you'll have gained experience with the implementation of your models on a variety of use cases. What you will learnUnderstand the basics of deep learningSet up Apache Spark for deep learningUnderstand the principles of distribution modeling and different types of neural networksObtain an understanding of deep learning algorithmsDiscover textual analysis and deep learning with SparkUse popular deep learning frameworks, such as Deeplearning4j, TensorFlow, and KerasExplore popular deep learning algorithms Who this book is for If you are a Scala developer, data scientist, or data analyst who wants to learn how to use Spark for implementing efficient deep learning models, Hands-On Deep Learning with Apache Spark is for you. Knowledge of the core machine learning concepts and some exposure to Spark will be helpful.

Introduction to Apache Flink

Hands-On for Developers and Technical Professionals

Developing Open Serverless Solutions

Streaming Systems

Kafka Streams - Real-time Stream Processing

Mastering Apache Flink

Build and deploy distributed deep learning applications on Apache Spark

Serverless computing greatly simplifies software development. Your team can focus solely on your application while the cloud provider manages the servers you need. This practical guide shows you step-by-step how to build and deploy complex applications in a flexible multicloud, multilanguage environment using Apache OpenWhisk. You’ll learn how this platform enables you to pursue a vendor-independent approach using preconfigured containers, microservices, and Kubernetes as your cloud operating system. Michele Sciabarrà demonstrates how to build a serverless application using classical design patterns and the programming language or languages that best fit your task. You’ll start by building a simple serverless application hands-on before diving into the more complex aspects of the OpenWhisk platform. Examine how OpenWhisk’s serverless architecture works, including the use of packages, actions, sequences, triggers, rules, and feeds Learn how OpenWhisk compares to existing architectures, such as Java Enterprise Edition Manipulate OpenWhisk features using the command-line interface or a JavaScript API Design applications using common Gang of Four design patterns Use architectural design patterns such as model-view-controller to combine several OpenWhisk actions Learn how to test and debug your code in a serverless environment

Learn how to use, deploy, and maintain Apache Spark with this comprehensive guide, written by the creators of the open-source cluster-computing framework. With an emphasis on improvements and new features in Spark 2.0, authors Bill Chambers and Matei Zaharia break down Spark topics into distinct sections, each with unique goals. You’ll explore the basic operations and common functions of Spark’s structured APIs, as well as Structured Streaming, a new high-level API for building end-to-end streaming applications. Developers and system administrators will learn the fundamentals of monitoring, tuning, and debugging

Spark, and explore machine learning techniques and scenarios for employing MLlib, Spark’s scalable machine-learning library. Get a gentle overview of big data and Spark Learn about DataFrames, SQL, and Datasets—Spark’s core APIs—through worked examples Dive into Spark’s low-level APIs, RDDs, and execution of SQL and DataFrames Understand how Spark runs on a cluster Debug, monitor, and tune Spark clusters and applications Learn the power of Structured Streaming, Spark’s stream-processing engine Learn how you can apply MLlib to a variety of problems, including classification or recommendation

If you are a Hadoop programmer who wants to learn about Flume to be able to move datasets into Hadoop in a timely and replicable manner, then this book is ideal for you. No prior knowledge about Apache Flume is necessary, but a basic knowledge of Hadoop and the Hadoop File System (HDFS) is assumed.

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.

Get expert guidance on architecting end-to-end data management solutions with Apache Hadoop. While many sources explain how to use various components in the Hadoop ecosystem, this practical book takes you through architectural considerations necessary to tie those components together into a complete tailored application, based on your particular use case. To reinforce those lessons, the book’s second section provides detailed examples of architectures used in some of the most commonly found Hadoop applications. Whether you’re designing a new Hadoop application, or planning to integrate Hadoop into your existing data infrastructure, Hadoop Application Architectures will skillfully guide you through the process. This book covers: Factors to consider when using Hadoop to store and model data Best practices for moving data in and out of the system Data processing frameworks, including MapReduce, Spark, and Hive Common Hadoop processing patterns, such as removing duplicate records and using windowing analytics Giraph, GraphX, and other tools for large graph processing on Hadoop Using workflow orchestration and scheduling tools such as Apache Oozie Near-real-time stream processing with Apache Storm, Apache Spark Streaming, and Apache Flume Architecture examples for clickstream analysis, fraud detection, and data warehousing

Real-time streaming applications with Apex

The What, Where, When, and How of Large-Scale Data Processing

Mastering Apache Spark 2.x

Building Full-Stack Data Analytics Applications with Spark

A Hands-On Guide to Building Robust and Scalable Event-Driven Applications with Code Examples in Java

OpenShift for Developers

Learning Apache Drill

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.

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

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

Dig deep into the data with a hands-on guide to machine learning with updated examples and more! Machine Learning: Hands-On for Developers and Technical Professionals provides hands-on instruction and fully-coded working examples for the most common machine learning techniques used by developers and technical professionals. The book contains a breakdown of each ML variant, explaining how it works and how it is used within certain industries, allowing readers to incorporate the presented techniques into their own work as they follow along. A core tenant of machine learning is a strong focus on data preparation, and a full exploration of the various types of learning algorithms illustrates how the proper tools can help any developer extract information and insights from existing data. The book includes a full complement of Instructor's Materials to facilitate use in the classroom, making this resource useful for students and as a professional reference. At its core, machine learning is a mathematical, algorithm-based technology that forms the basis of historical data mining and modern big data science. Scientific analysis of big data requires a working knowledge of machine learning, which forms predictions based on known properties learned from training data. Machine Learning is an accessible, comprehensive guide for the non-mathematician, providing clear guidance that allows readers to: Learn the languages of machine learning including Hadoop, Mahout, and Weka Understand decision trees, Bayesian networks, and artificial neural networks Implement Association Rule, Real Time, and Batch learning Develop a strategic plan for safe, effective, and efficient machine learning By learning to construct a system that can learn from data, readers can increase their utility across industries. Machine learning sits at the core of deep dive data analysis and visualization, which is increasingly in demand as companies discover the goldmine hiding in their existing data. For the tech professional involved in data science, Machine Learning: Hands-On for Developers and Technical Professionals provides the skills and techniques required to dig deeper.

Advanced analytics on your Big Data with latest Apache Spark 2.x About This Book An advanced guide with a combination of instructions and practical examples to extend the most up-to date Spark functionalities. Extend your data processing capabilities to process huge chunk of data in minimum time using advanced concepts in Spark. Master the art of real-time processing with the help of Apache Spark 2.x Who This Book Is For If you are a developer with some experience with Spark and want to strengthen your knowledge of how to get around in the world of Spark, then this book is ideal for you. Basic knowledge of Linux, Hadoop and Spark is assumed. Reasonable knowledge of Scala is expected. What You Will Learn Examine Advanced Machine Learning and DeepLearning with MLlib, SparkML, SystemML, H2O and DeepLearning4J Study highly optimised unified batch and real-time data processing using SparkSQL and Structured Streaming Evaluate large-scale Graph Processing and Analysis using GraphX and GraphFrames Apply Apache Spark in Elastic deployments using Jupyter and Zeppelin Notebooks, Docker, Kubernetes and the IBM Cloud Understand internal details of cost based optimizers used in Catalyst, SystemML and GraphFrames Learn how specific parameter settings affect overall performance of an Apache Spark cluster Leverage Scala, R and python for your data science projects In Detail Apache Spark is an in-memory cluster-based parallel processing system that provides a wide range of functionalities such as graph processing, machine learning, stream processing, and SQL. This book aims to take your knowledge of Spark to the next level by teaching you how to expand Spark's functionality and implement your data flows and machine/deep learning programs on top of the platform. The book commences with an overview of the Spark ecosystem. It will introduce you to Project Tungsten and Catalyst, two of the major advancements of Apache Spark 2.x. You will understand how memory management and binary processing, cache-aware computation, and code generation are used to speed things up dramatically. The book extends to show how to incorporate H20, SystemML, and Deeplearning4j for machine learning, and Jupyter Notebooks and Kubernetes/Docker for cloud-based Spark. During the course of the book, you will learn about the latest enhancements to Apache Spark 2.x, such as interactive querying of live data and unifying DataFrames and Datasets. You will also learn about the updates on the APIs and how DataFrames and Datasets affect SQL, machine learning, graph processing, and streaming. You will learn to use Spark as a big data operating system, understand how to implement advanced analytics on the new APIs, and explore how easy it is to use Spark in day-to-day tasks. Style and approach This book is an extensive guide to Apache Spark modules and tools and shows how Spark's functionality can be extended for real-time processing and storage with worked examples.

Real-Time Data and Stream Processing at Scale

Effective Kafka

Hands-On Deep Learning with Apache Spark

I Heart Logs

Kafka Streams in Action

Distributed Data at Web Scale

Learning Apache Kafka Second Edition

Build a strong and efficient IoT infrastructure at industrial and enterprise level by mastering Industrial IoT network Key FeaturesGain hands-on experience working with industrial architectureExplore the potential of cloud-based Industrial IoT platforms, analytics, and protocolsImprove business models and transform your workforce with Industry 4.0Book Description We live in an era where advanced automation is used to achieve accurate results. To set up an automation environment, you need to first configure a network that can be accessed anywhere and by any device. This book is a practical guide that helps you discover the technologies and use cases for Industrial Internet of Things (IIOT). Hands-On Industrial Internet of Things takes you through the implementation of industrial processes and specialized control devices and protocols. You’ll study the process of identifying and connecting to different industrial data sources gathered from different sensors. Furthermore, you’ll be able to connect these sensors to cloud network, such as AWS IoT, Azure IoT, Google IoT, and OEM IoT platforms, and extract data from the cloud to your devices. As you progress through the chapters, you’ll gain hands-on experience in using open source Node-Red, Kafka, Cassandra, and Python. You will also learn how to develop streaming and batch-based Machine Learning algorithms. By the end of this book, you will have mastered the features of Industry 4.0 and be able to build stronger, faster, and more reliable IoT infrastructure in your Industry. What you will learnExplore industrial processes, devices, and protocolsDesign and implement the I-IoT network flowGather and transfer industrial data in a secure wayGet to grips with popular cloud-based platformsUnderstand diagnostic analytics to answer critical workforce questionsDiscover the Edge device and understand Edge and Fog computingImplement equipment and process management to achieve business-specific goalsWho this book is for If you’re an IoT architect, developer, or stakeholder working with architectural aspects of Industrial Internet of Things, this book is for you.

What is this book about? PHP, Apache, and MySQL are the three key open source technologies that form the basis for most active Web servers. This book takes you step-by-step through understanding each — using it and combining it with the other two on both Linux and Windows servers. This book guides you through creating your own sites using the open source AMP model. You discover how to install PHP, Apache, and MySQL. Then you create PHP Web pages, including database management and security. Finally, you discover how to integrate your work with e-commerce and other technologies. By building different types of Web sites, you progress from setting up simple database tables to tapping the full potential of PHP, Apache, and MySQL. When you’re finished, you will be able to create well-designed, dynamic Web sites using open source tools. What does this book cover? Here's what you will learn from this book: How PHP server-side scripting language works for connecting HTML-based Web pages to a backend database Syntax, functions, and commands for PHP, Apache, and MySQL Methods and techniques for building user-friendly forms How to easily store, update, and access information using MySQL Ways to allow the user to edit a database E-commerce applications using these three technologies How to set up user logins, profiles, and personalizations Proper protocols for error handling Who is this book for? This book is for beginners who are new to PHP and who need to learn quickly how to create Web sites using open source tools. Some basic HTML knowledge is helpful but not essential.

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

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.

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

Learning Apache Cassandra - Second Edition

Mastering Structured Streaming and Spark Streaming

Cassandra: The Definitive Guide

New Designs Using Apache Kafka and Mapr Streams

Lightning-Fast Big Data Analysis

Learning Apache Apex

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.

With Resilient Distributed Datasets, Spark SQL, Structured Streaming and Spark Machine Learning library

Beginning PHP, Apache, MySQL Web Development

Stream Processing with Apache Flink

Real-time apps and microservices with the Kafka Streams API

Hands-On Industrial Internet of Things

Beginning Apache Spark 2