Apache Spark The Definitive

Spark: The Definitive GuideBig Data Processing Made Simple"O'Reilly Media, Inc."

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 third edition—updated for Cassandra 4.0—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 nonrelational 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

This open access book was prepared as a Final Publication of the COST Action IC1406 "High-Performance Modelling and Simulation for Big Data Applications (cHiPSet)" project. Long considered important pillars of the scientific method, Modelling and Simulation have evolved from traditional discrete

numerical methods to complex data-intensive continuous analytical optimisations. Resolution, scale, and accuracy have become essential to predict and analyse natural and complex systems in science and engineering. When their level of abstraction raises to have a better discernment of the domain at hand, their representation gets increasingly demanding for computational and data resources. On the other hand, High Performance Computing typically entails the effective use of parallel and distributed processing units coupled with efficient storage, communication and visualisation systems to underpin complex data-intensive applications in distinct scientific and technical domains. It is then arguably required to have a seamless interaction of High Performance Computing with Modelling and Simulation in order to store, compute, analyse, and visualise large data sets in science and engineering. Funded by the European Commission, cHiPSet has provided a dynamic trans-European forum for their members and distinguished guests to openly discuss novel perspectives and topics of interests for these two communities. This cHiPSet compendium presents a set of selected case studies related to healthcare, biological data, computational advertising, multimedia, finance, bioinformatics, and telecommunications.

This thoroughly revised guide demonstrates how the flexibility of the command line can help you become a more efficient and productive data scientist. You'll learn how to combine small yet powerful command-line tools to quickly obtain, scrub, explore, and model your data. To get you started, author Jeroen Janssens provides a

Docker image packed with over 80 tools--useful whether you work with Windows, macOS, or Linux. You'll quickly discover why the command line is an agile, scalable, and extensible technology. Even if you're comfortable processing data with Python or R, you'll learn how to greatly improve your data science workflow by leveraging the command line's power. This book is ideal for data scientists, analysts, and engineers; software and machine learning engineers; and system administrators. Obtain data from websites, APIs, databases, and spreadsheets Perform scrub operations on text, CSV, HTM, XML, and JSON files Explore data, compute descriptive statistics, and create visualizations Manage your data science workflow Create reusable command-line tools from one-liners and existing Python or R code Parallelize and distribute data-intensive pipelines Model data with dimensionality reduction, clustering, regression, and classification algorithms

Work with petabyte-scale datasets while building a collaborative, agile workplace in the process. This practical book is the canonical reference to Google BigQuery, the query engine that lets you conduct interactive analysis of large datasets. BigQuery enables enterprises to efficiently store, query, ingest, and learn from their data in a convenient framework. With this book, you'll examine how to analyze data at scale to derive insights from large datasets efficiently. Valliappa Lakshmanan, tech lead for Google Cloud Platform, and Jordan Tigani, engineering director for the BigQuery team, provide best practices for modern data warehousing within an autoscaled, serverless public

cloud. Whether you want to explore parts of BigQuery you're not familiar with or prefer to focus on specific tasks, this reference is indispensable.

?CPython?????,?C????,????????Python???????????Python???????,??????????Python????Python???????.

C?C++????

2008??????????????????? 50????????49??????

35?????????????? 2012?????????????? ???50??????

??????????????????????????

??????????????????????????? ????????

???????????????? ?????????????

??????????????????? ??????????????????20

????????1???????????1? ?????????????10?

????????2012?????????1?

????????????2012???????????

??????????2012???????????

????????????10?????????!?

?2009???????????????????????????

?2012????????????????????????? ??????????????

??????????????????????911?????

?2007????????????????????????????? ???????? (big data)??

??

??

! ??

??????????????????????????????????????(????:?????

?????????????) ????????????????????????????????

??

??????????;????????????????????????????????????

????????????????????????????????? ?????????????

??

??
??
????????(??????????????: ???8?) ???????????:
??????????????????????
??
??????????????????????? ????????????????
??????????????????????????????????
??????????????????????????????????? ?????????
??????????????????????????????????????
???????????????????????????????????

This book includes high-quality, peer-reviewed papers from the International Conference on Recent Advancement in Computer, Communication and Computational Sciences (RACCCS-2017), held at Aryabhatta College of Engineering & Research Center, Ajmer, India on September 2–3, 2017, presenting the latest developments and technical solutions in computational sciences. Data science, data- and knowledge engineering require networking and communication as a backbone and have a wide scope of implementation in engineering sciences. Keeping this ideology in mind, the book offers insights that reflect the advances in these fields from upcoming researchers and leading academicians across the globe. Covering a variety of topics, such as intelligent hardware and software design, advanced communications, intelligent computing technologies, advanced software engineering, the web and informatics, and intelligent image processing, it helps those in the computer industry and academia use the advances of next-generation communication and computational technology to shape real-world applications.

Discover how graph algorithms can help you leverage the relationships within your data to develop more intelligent solutions and enhance your machine learning models. You'll

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

With Early Release ebooks, you get books in their earliest form—the author's raw and unedited content as he or she writes—so you can take advantage of these technologies long before the official release of these titles. You'll also receive updates when significant changes are made, new chapters are available, and the final ebook bundle is released. Learn how to use, deploy, and maintain Apache Spark with this comprehensive guide, written by the creators of this 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 Spark's Structured Streaming and MLlib for machine learning tasks Explore the wider Spark ecosystem, including SparkR and Graph Analysis Examine Spark deployment, including coverage of Spark in the Cloud

This volume constitutes the refereed proceedings of the three workshops held at the 31st International Conference on Database and Expert Systems Applications, DEXA 2020, held in September 2020: The 11th International Workshop on Biological Knowledge Discovery from Data, BIOKDD 2020, the 4th International Workshop on Cyber-Security and Functional Safety in Cyber-Physical Systems, IWCFS 2020, the 2nd International Workshop on Machine Learning and Knowledge Graphs, MLKgraphs2019. Due to the COVID-19 pandemic the conference and workshop were held virtually. The 10 papers were thoroughly reviewed and selected from 15 submissions, and discuss a range of topics including: knowledge discovery, biological data, cyber security, cyber-physical system, machine learning, knowledge graphs, information retriever, data base, and artificial intelligent.

Access real-world documentation and examples for the Spark platform for building large-scale, enterprise-grade machine learning applications. The past decade has seen an

astonishing series of advances in machine learning. These breakthroughs are disrupting our everyday life and making an impact across every industry. Next-Generation Machine Learning with Spark provides a gentle introduction to Spark and Spark MLlib and advances to more powerful, third-party machine learning algorithms and libraries beyond what is available in the standard Spark MLlib library. By the end of this book, you will be able to apply your knowledge to real-world use cases through dozens of practical examples and insightful explanations. What You Will Learn Be introduced to machine learning, Spark, and Spark MLlib 2.4.x Achieve lightning-fast gradient boosting on Spark with the XGBoost4J-Spark and LightGBM libraries Detect anomalies with the Isolation Forest algorithm for Spark Use the Spark NLP and Stanford CoreNLP libraries that support multiple languages Optimize your ML workload with the Alluxio in-memory data accelerator for Spark Use GraphX and GraphFrames for Graph Analysis Perform image recognition using convolutional neural networks Utilize the Keras framework and distributed deep learning libraries with Spark Who This Book Is For Data scientists and machine learning engineers who want to take their knowledge to the next level and use Spark and more powerful, next-generation algorithms and libraries beyond what is available in the standard Spark MLlib library; also serves as a primer for aspiring data scientists and engineers who need an introduction to machine learning, Spark, and Spark MLlib.

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

Virtual, hands-on learning labs allow you to apply your technical skills in realistic environments. So Sybex has bundled AWS labs from XtremeLabs with our popular AWS Certified Data Analytics Study Guide to give you the same experience working in these labs as you prepare for the Certified Data Analytics Exam that you would face in a real-life application. These labs in addition to the book are a proven way to prepare for the certification and for work as an AWS Data Analyst. AWS Certified Data Analytics Study Guide: Specialty (DAS-C01) Exam is intended for individuals who perform in a data analytics-focused role. This UPDATED exam validates an examinee's comprehensive understanding of using AWS services to design, build, secure, and maintain analytics solutions that provide insight from data. It assesses an examinee's ability to define AWS data analytics services and understand how they integrate with each other; and explain how AWS data analytics services fit in the data lifecycle of collection, storage, processing, and visualization. The book focuses on the following domains: • Collection • Storage and Data Management • Processing • Analysis and Visualization • Data Security This is your opportunity to take the next step in your career by expanding and validating your

skills on the AWS cloud. AWS is the frontrunner in cloud computing products and services, and the AWS Certified Data Analytics Study Guide: Specialty exam will get you fully prepared through expert content, and real-world knowledge, key exam essentials, chapter review questions, and much more. Written by an AWS subject-matter expert, this study guide covers exam concepts, and provides key review on exam topics. Readers will also have access to Sybex's superior online interactive learning environment and test bank, including chapter tests, practice exams, a glossary of key terms, and electronic flashcards. And included with this version of the book, XtremeLabs virtual labs that run from your browser. The registration code is included with the book and gives you 6 months of unlimited access to XtremeLabs AWS Certified Data Analytics Labs with 3 unique lab modules based on the book.

Learn how to take full advantage of Apache Kafka, the distributed, publish-subscribe queue for handling real-time data feeds. With this comprehensive book, you will understand how Kafka works and how it is designed. Authors Neha Narkhede, Gwen Shapira, and Todd Palino show you how to deploy production Kafka clusters; secure, tune, and monitor them; write rock-solid applications that use Kafka; and build scalable stream-processing applications. Learn how Kafka compares to other queues, and where it fits in the big data ecosystem. Dive into Kafka's internal designPick up best practices for developing applications that use Kafka. Understand the best way to deploy Kafka in production monitoring, tuning, and maintenance tasks. Learn how to secure a Kafka cluster.

This book focuses on the core areas of computing and their applications in the real world. Presenting papers from the Computing Conference 2020 covers a diverse range of research areas, describing various detailed techniques that

have been developed and implemented. The Computing Conference 2020, which provided a venue for academic and industry practitioners to share new ideas and development experiences, attracted a total of 514 submissions from pioneering academic researchers, scientists, industrial engineers and students from around the globe. Following a double-blind, peer-review process, 160 papers (including 15 poster papers) were selected to be included in these proceedings. Featuring state-of-the-art intelligent methods and techniques for solving real-world problems, the book is a valuable resource and will inspire further research and technological improvements in this important area.

"[This] book provides a most comprehensive view of an Enterprise IoT stack, detailed IoT use cases on manufacturing, automotive and home automation and how to implement IoT applications using Microsoft, IBM, Amazon and GE Preix IoT ... and various open source technologies like Apache Kakfa [i.e. Kafka, and] Apache Spark"--Page 4 of cover.

Analysis and machine learning models are only as good as the data they're built on. Querying processed data and getting insights from it requires a robust data pipeline--and an effective storage solution that ensures data quality, data integrity, and performance. This guide introduces you to Delta Lake, an open-source format that enables building a lakehouse architecture on top of existing storage systems such as S3, ADLS, GCS, and HDFS. Delta Lake enhances Apache Spark and makes it easy to store and manage massive amounts of complex data by supporting data integrity, data quality, and performance. Data engineers, data scientists, and data practitioners will learn how to build reliable data lakes and data pipelines at scale using Delta Lake. Understand key data reliability challenges and how to tackle them Learn how to use Delta Lake to realize data

reliability improvements Concurrently run streaming and batch jobs against your data lake Execute update, delete, and merge commands against your data lake Use time travel to roll back and examine previous versions of your data Learn best practices to build effective, high-quality end-to-end data pipelines for real world use cases Integrate with other data technologies like Presto, Athena, Redshift and other BI tools Learn how thousands of companies are processing exabytes of data per month with their lakehouse architecture using Delta Lake.

This three-volume set LNCS 12452, 12453, and 12454 constitutes the proceedings of the 20th International Conference on Algorithms and Architectures for Parallel Processing, ICA3PP 2020, in New York City, NY, USA, in October 2020. The total of 142 full papers and 5 short papers included in this proceedings volumes was carefully reviewed and selected from 495 submissions. ICA3PP is covering the many dimensions of parallel algorithms and architectures, encompassing fundamental theoretical approaches, practical experimental projects, and commercial components and systems. As applications of computing systems have permeated in every aspects of daily life, the power of computing system has become increasingly critical. This conference provides a forum for academics and practitioners from countries around the world to exchange ideas for improving the efficiency, performance, reliability, security and interoperability of computing systems and applications. ICA3PP 2020 focus on two broad areas of parallel and distributed computing, i.e. architectures, algorithms and networks, and systems and applications.--

??????????????? ??????????????? ???????????????? ??? ??????? ??? ?Wired????????? ??? ???? ?????????????????????????? ???????????? —— ????Lawrence Lessig???????????????? ??????????????????????????

??????????????????????? ??????????????????? ??????????????????????? ??????????????????????????? ????????????????????????????????? ????????????????????????????????? ??????????????????????????? ????????? ??????????????????????????? ?????????????……

Get ready to unlock the power of your data. With the fourth edition of this comprehensive guide, you'll learn how to build and maintain reliable, scalable, distributed systems with Apache Hadoop. This book is ideal for programmers looking to analyze datasets of any size, and for administrators who want to set up and run Hadoop clusters. Using Hadoop 2 exclusively, author Tom White presents new chapters on YARN and several Hadoop-related projects such as Parquet, Flume, Crunch, and Spark. You'll learn about recent changes to Hadoop, and explore new case studies on Hadoop's role in healthcare systems and genomics data processing. Learn fundamental components such as MapReduce, HDFS, and YARN Explore MapReduce in depth, including steps for developing applications with it Set up and maintain a Hadoop cluster running HDFS and MapReduce on YARN Learn two data formats: Avro for data serialization and Parquet for nested data Use data ingestion tools such as Flume (for streaming data) and Sqoop (for bulk data transfer) Understand how high-level data processing tools like Pig, Hive, Crunch, and Spark work with Hadoop Learn the HBase distributed database and the ZooKeeper distributed configuration service

?????Go???????????????????Go???????????????????Go?????????JavaScript?Ruby?Python?Java?C++??????????? ??? ??????Go????????????????????????????????????? ???

??????Go?? ??????????????????????????????????????Go????????????? ? ???????????Go?? ?????????????????????????? ?????????????????????????? goroutine?channel??????????Go?????????????????????????? ?????????????????????????????????? ???????Go???????????reflection??????????????unsafe??????????????????cgo???? Go?C?????? ???????????Go???????????????????????????? ?????????????????????????????http://gopl.io/??????go get???????????? #???? GOTOP Information Inc.

This book constitutes the proceedings of the 20th IFIP International Conference on Distributed Applications and Interoperable Systems, DAIS 2020, which was supposed to be held in Valletta, Malta, in June 2020, as part of the 15th International Federated Conference on Distributed Computing Techniques, DisCoTec 2020. The conference was held virtually due to the COVID-19 pandemic. The 10 full papers presented together with 1 short paper and 1 invited paper were carefully reviewed and selected from 17 submissions. The papers addressed challenges in multiple application areas, such as privacy and security, cloud and systems, fault-tolerance and reproducibility, machine learning for systems, and distributed algorithms.

If you're training a machine learning model but aren't sure how to put it into production, this book will get you there. Kubeflow provides a collection of cloud native tools for different stages of a model's lifecycle, from data exploration, feature preparation, and model training to model serving. This guide helps data scientists build production-grade machine learning implementations with Kubeflow and shows data engineers how to make models scalable and

reliable. Using examples throughout the book, authors Holden Karau, Trevor Grant, Ilan Filonenko, Richard Liu, and Boris Lublinsky explain how to use Kubeflow to train and serve your machine learning models on top of Kubernetes in the cloud or in a development environment on-premises. Understand Kubeflow's design, core components, and the problems it solves Understand the differences between Kubeflow on different cluster types Train models using Kubeflow with popular tools including Scikit-learn, TensorFlow, and Apache Spark Keep your model up to date with Kubeflow Pipelines Understand how to capture model training metadata Explore how to extend Kubeflow with additional open source tools Use hyperparameter tuning for training Learn how to serve your model in production
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

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

This book presents a focus on proteins and their structures. The text describes various scalable solutions for protein structure similarity searching, carried out at main representation levels and for prediction of 3D structures of proteins. Emphasis is placed on techniques that can be used to accelerate similarity searches and protein structure modeling processes. The content of the book is divided into four parts. The first part provides background information on proteins and their representation levels, including a formal model of a 3D protein structure used in computational processes, and a brief overview of the technologies used in the solutions presented in the book. The second part of the book discusses Cloud services that are utilized in the development of scalable and reliable cloud applications for 3D protein structure similarity searching and protein structure prediction. The third part of the book shows the utilization of scalable Big

Data computational frameworks, like Hadoop and Spark, in massive 3D protein structure alignments and identification of intrinsically disordered regions in protein structures. The fourth part of the book focuses on finding 3D protein structure similarities, accelerated with the use of GPUs and the use of multithreading and relational databases for efficient approximate searching on protein secondary structures. The book introduces advanced techniques and computational architectures that benefit from recent achievements in the field of computing and parallelism. Recent developments in computer science have allowed algorithms previously considered too time-consuming to now be efficiently used for applications in bioinformatics and the life sciences. Given its depth of coverage, the book will be of interest to researchers and software developers working in the fields of structural bioinformatics and biomedical databases.

Data Processing is one of the core functionalities of distributed and cloud computing. There is a high demand on low latency and high performance computing as well as the support of abstract processing methods such as SQL querying, analytic frameworks or graph processing by data processing engines. The Definitive Guide to Apache Flink by Papp starts with the history of Big Data processing with Hadoop and explains the shortcomings of Map Reduce. It shows how YARN and Hadoop 2.x

changed the game and how new technologies started to compete to become the successor of Map Reduce. After some detailed information on Tez and Spark and how they try to solve shortcomings of Map Reduce, this book deals with some architectural patterns for creating a solid data processing engine, such as advanced pipelining methods or in-memory caching. It shows how Flink is using these concepts. Flink programming will be introduced in a hands-on approach. It starts with how to create a ten minutes build and how to run the first "Word Count" with Flink. Then it continues with more advanced topics such as programming more complex programs. All samples are programmed with Java or Scala. It shows that Apache Flink has the potential to become one of the key technologies for distributed computing. It aims to replace many small technologies with a more powerful one that covers many aspects of Hadoop programming.

Apache Spark is a flexible in-memory framework that allows processing of both batch and real-time data. Its unified engine has made it quite popular for big data use cases. This book will help you to quickly get started with Apache Spark 2.0 and write efficient big data applications for a variety of use cases.

Perform fast interactive analytics against different data sources using the Trino high-performance distributed SQL query engine. With this practical guide, you'll learn how to conduct analytics on data

where it lives, whether it's Hive, Cassandra, a relational database, or a proprietary data store. Analysts, software engineers, and production engineers will learn how to manage, use, and even develop with Trino. Initially developed by Facebook, open source Trino is now used by Netflix, Airbnb, LinkedIn, Twitter, Uber, and many other companies. Matt Fuller, Manfred Moser, and Martin Traverso show you how a single Trino query can combine data from multiple sources to allow for analytics across your entire organization. Get started: Explore Trino's use cases and learn about tools that will help you connect to Trino and query data Go deeper: Learn Trino's internal workings, including how to connect to and query data sources with support for SQL statements, operators, functions, and more Put Trino in production: Secure Trino, monitor workloads, tune queries, and connect more applications; learn how other organizations apply Trino

Work with all aspects of batch processing in a modern Java environment using a selection of Spring frameworks. This book provides up-to-date examples using the latest configuration techniques based on Java configuration and Spring Boot. The Definitive Guide to Spring Batch takes you from the “Hello, World!” of batch processing to complex scenarios demonstrating cloud native techniques for developing batch applications to be run on modern

platforms. Finally this book demonstrates how you can use areas of the Spring portfolio beyond just Spring Batch 4 to collaboratively develop mission-critical batch processes. You'll see how a new class of use cases and platforms has evolved to have an impact on batch-processing. Data science and big data have become prominent in modern IT and the use of batch processing to orchestrate workloads has become commonplace. The Definitive Guide to Spring Batch covers how running finite tasks on cloud infrastructure in a standardized way has changed where batch applications are run. Additionally, you'll discover how Spring Batch 4 takes advantage of Java 9, Spring Framework 5, and the new Spring Boot 2 micro-framework. After reading this book, you'll be able to use Spring Boot to simplify the development of your own Spring projects, as well as take advantage of Spring Cloud Task and Spring Cloud Data Flow for added cloud native functionality. Includes a foreword by Dave Syer, Spring Batch project founder. What You'll Learn Discover what is new in Spring Batch 4 Carry out finite batch processing in the cloud using the Spring Batch project Understand the newest configuration techniques based on Java configuration and Spring Boot using practical examples Master batch processing in complex scenarios including in the cloud Develop batch applications to be run on modern platforms Use

areas of the Spring portfolio beyond Spring Batch to develop mission-critical batch processes Who This Book Is For Experienced Java and Spring coders new to the Spring Batch platform. This definitive book will be useful in allowing even experienced Spring Batch users and developers to maximize the Spring Batch tool.

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

Best practices for building data pipelines and applications with Kafka How to perform monitoring, tuning, and maintenance tasks with Kafka in production The most critical metrics among Kafka's operational measurements Kafka's delivery capabilities for stream processing systems
This book presents the proceedings of the International Conference on Cyber-Physical Systems and Control (CPS&C'2019), held in Peter the Great St. Petersburg Polytechnic University, which is celebrating its 120th anniversary in 2019. The CPS&C'2019 was dedicated to the 35th anniversary of the partnership between Peter the Great St. Petersburg Polytechnic University and Leibniz University of Hannover. Cyber-physical systems (CPSs) are a new generation of control systems and techniques that help promote prospective interdisciplinary research. A wide range of theories and methodologies are currently being investigated and developed in this area to tackle various complex and challenging problems. Accordingly, CPSs represent a scientific and engineering discipline that is set to make an impact on future systems of industrial and social scale that are characterized by the deep integration of real-time processing, sensing, and actuation into logical and physical heterogeneous domains. The CPS&C'2019 brought together researchers and practitioners from all over the world and to discuss cross-cutting fundamental

scientific and engineering principles that underline the integration of cyber and physical elements across all application fields. The participants represented research institutions and universities from Austria, Belgium, Bulgaria, China, Finland, Germany, the Netherlands, Russia, Syria, Ukraine, the USA, and Vietnam. These proceedings include 75 papers arranged into five sections, namely keynote papers, fundamentals, applications, technologies, and education and social aspects. Understand the complexities of modern-day data engineering platforms and explore strategies to deal with them with the help of use case scenarios led by an industry expert in big data Key Features Become well-versed with the core concepts of Apache Spark and Delta Lake for building data platforms Learn how to ingest, process, and analyze data that can be later used for training machine learning models Understand how to operationalize data models in production using curated data Book Description In the world of ever-changing data and schemas, it is important to build data pipelines that can auto-adjust to changes. This book will help you build scalable data platforms that managers, data scientists, and data analysts can rely on. Starting with an introduction to data engineering, along with its key concepts and architectures, this book will show you how to use Microsoft Azure Cloud services effectively for data engineering. You'll cover data

lake design patterns and the different stages through which the data needs to flow in a typical data lake. Once you've explored the main features of Delta Lake to build data lakes with fast performance and governance in mind, you'll advance to implementing the lambda architecture using Delta Lake. Packed with practical examples and code snippets, this book takes you through real-world examples based on production scenarios faced by the author in his 10 years of experience working with big data. Finally, you'll cover data lake deployment strategies that play an important role in provisioning the cloud resources and deploying the data pipelines in a repeatable and continuous way. By the end of this data engineering book, you'll know how to effectively deal with ever-changing data and create scalable data pipelines to streamline data science, ML, and artificial intelligence (AI) tasks. What you will learn Discover the challenges you may face in the data engineering world Add ACID transactions to Apache Spark using Delta Lake Understand effective design strategies to build enterprise-grade data lakes Explore architectural and design patterns for building efficient data ingestion pipelines Orchestrate a data pipeline for preprocessing data using Apache Spark and Delta Lake APIs Automate deployment and monitoring of data pipelines in production Get to grips with securing, monitoring, and managing data pipelines models efficiently Who this book is for This

book is for aspiring data engineers and data analysts who are new to the world of data engineering and are looking for a practical guide to building scalable data platforms. If you already work with PySpark and want to use Delta Lake for data engineering, you'll find this book useful. Basic knowledge of Python, Spark, and SQL is expected.

Big Data Analytics with Spark is a step-by-step guide for learning Spark, which is an open-source fast and general-purpose cluster computing framework for large-scale data analysis. You will learn how to use Spark for different types of big data analytics projects, including batch, interactive, graph, and stream data analysis as well as machine learning. In addition, this book will help you become a much sought-after Spark expert. Spark is one of the hottest Big Data technologies. The amount of data generated today by devices, applications and users is exploding. Therefore, there is a critical need for tools that can analyze large-scale data and unlock value from it. Spark is a powerful technology that meets that need. You can, for example, use Spark to perform low latency computations through the use of efficient caching and iterative algorithms; leverage the features of its shell for easy and interactive Data analysis; employ its fast batch processing and low latency features to process your real time data streams and so on. As a result, adoption of Spark is rapidly growing and is replacing Hadoop MapReduce

as the technology of choice for big data analytics. This book provides an introduction to Spark and related big-data technologies. It covers Spark core and its add-on libraries, including Spark SQL, Spark Streaming, GraphX, and MLlib. Big Data Analytics with Spark is therefore written for busy professionals who prefer learning a new technology from a consolidated source instead of spending countless hours on the Internet trying to pick bits and pieces from different sources. The book also provides a chapter on Scala, the hottest functional programming language, and the program that underlies Spark. You'll learn the basics of functional programming in Scala, so that you can write Spark applications in it. What's more, Big Data Analytics with Spark provides an introduction to other big data technologies that are commonly used along with Spark, like Hive, Avro, Kafka and so on. So the book is self-sufficient; all the technologies that you need to know to use Spark are covered. The only thing that you are expected to know is programming in any language. There is a critical shortage of people with big data expertise, so companies are willing to pay top dollar for people with skills in areas like Spark and Scala. So reading this book and absorbing its principles will provide a boost—possibly a big boost—to your career. ????? ??????1????????1? ?????????????Top 3??????????? ?????????????????????????????? ———————————————— ????????????????????

??????????????????????????? ???????????
????????????????????????????????? ??????????????
??
??????????????????????
????????????????????????????????? ??????????????
??
??????????????????????????
????????????????????????????????? ??????????????
??
??????????????????????????? ????????????????????
??
??
????????????????????????????????? ???????????????????
??
???????? ??
??????????????????????????????? ????????????????????
????????????????????????????????????……?????? ????
??
????????????????? ????????????????????????????????
??????????????????????? ????????????????????????????
????????????????????????????????? ?2010?5?6??????
??
????? ??
??——????????????? ????????????????????????????????
????????????????????????37???? ???????????????????
??
??? ????
??……
…… ??

?????????——?????????????????????????????????
???? ?????? ?????????????????????????????????
????????????????????? ???? | ????????Brian Christian ????????Most Human Human???????????
???
???
???????????? ???????Tom Griffiths ????????????
???
???
? ???? | ??? ??????????????????????Discovery
??????????????????????BBC???????????????????
??????????????????1905????????????????????????
???
????

If you’re like most R users, you have deep knowledge and love for statistics. But as your organization continues to collect huge amounts of data, adding tools such as Apache Spark makes a lot of sense. With this practical book, data scientists and professionals working with large-scale data applications will learn how to use Spark from R to tackle big data and big compute problems. Authors Javier Luraschi, Kevin Kuo, and Edgar Ruiz show you how to use R with Spark to solve different data analysis problems. This book covers relevant data science topics, cluster computing, and issues that should interest even the most advanced users. Analyze, explore, transform, and visualize data in Apache Spark with R Create statistical models to

extract information and predict outcomes; automate the process in production-ready workflows Perform analysis and modeling across many machines using distributed computing techniques Use large-scale data from multiple sources and different formats with ease from within Spark Learn about alternative modeling frameworks for graph processing, geospatial analysis, and genomics at scale Dive into advanced topics including custom transformations, real-time data processing, and creating custom Spark extensions

Analyze vast amounts of data in record time using Spark with Databricks in the Cloud. Learn the basic fundamentals, and more, of running analytics on large clusters in Azure and AWS, using Spark with Databricks on top. Discover how to squeeze the most value out of your data at a mere fraction of what classical analytics solutions cost, while at the same time getting the results you need, incrementally faster. This book explains how the confluence of these pivotal technologies gives you enormous power, and cheaply, when it comes to huge datasets. You will begin by learning how cloud infrastructure makes it possible to scale your code to large amounts of processing units, without having to pay for the machinery in advance. From there you will learn how Spark, an open source framework, can enable all those CPUs for data analytics use. Finally, you will see how services such as Databricks

provide the power of Spark, without you having to know anything about configuring hardware or software. By removing the need for expensive experts and hardware, your resources can instead be allocated to actually finding business value in the data. This book guides you through some advanced topics such as analytics in the cloud, data lakes, data ingestion, architecture, machine learning, and tools including Apache Spark, Apache Hadoop, Apache Hive, Python, and SQL. Valuable exercises help reinforce what you have learned. What You Will Learn Discover the value of big data analytics that leverage the power of the cloud Get started with Databricks using SQL and Python in either Microsoft Azure or AWS Understand the underlying technology, and how the cloud and Spark fit into the bigger picture See how these tools are used in the real world Run basic analytics, including machine learning, on billions of rows at a fraction of a cost or free Who This Book Is For Data engineers, data scientists, and cloud architects who want or need to run advanced analytics in the cloud. It is assumed that the reader has data experience, but perhaps minimal exposure to Spark and Databricks. The book is also recommended for people who want to get started in the analytics field, as it provides a strong foundation.