Spark The Definitive Guide Big Data Processing Made Simple

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

Combine advanced analytics including Machine Learning, Deep Learning Neural Networks and Natural Language Processing with modern scalable technologies including Apache Spark to derive actionable insights from Big Data in real-time Key Features Make a hands-on start in the fields of Big Data, Distributed Technologies and Machine Learning Learn how to design, develop and interpret the results of common Machine Learning algorithms Uncover hidden patterns in your data in order to derive real actionable insights and business value Book Description Every person and every organization in the world manages data, whether they realize it or not. Data is used to describe the world around us and can be used for almost any purpose, from analyzing consumer habits to fighting disease and serious organized crime. Ultimately, we manage data in order to derive value from it, and many organizations around the world have traditionally invested in technology to help process their data faster and more efficiently. But we now live in an interconnected world driven by mass data creation and consumption where data is no longer rows and columns restricted to a spreadsheet, but an organic and evolving asset in its own right. With this realization comes major challenges for organizations: how do we manage the sheer size of data being created every second (think not only spreadsheets and databases, but also social media posts, images, videos, music, blogs and so on)? And once we can manage all of this data, how do we derive real value from it? The focus of Machine Learning with Apache Spark is to help us answer these questions in a hands-on manner. We introduce the latest scalable technologies to help us manage and process big data. We then introduce advanced analytical algorithms applied to real-world use cases in order to uncover patterns, derive actionable insights, and learn from this big data. What you will learn Understand how Spark fits in the context of the big data ecosystem Understand how to deploy and configure a local development environment using Apache Spark Understand how to design supervised and unsupervised learning models Build models to perform NLP, deep learning, and cognitive services using Spark ML libraries Design real-time machine learning pipelines in Apache Spark Become familiar with advanced techniques for processing a large volume of data by applying machine learning algorithms Who this book is for This book is aimed at Business Analysts, Data Analysts and Data Scientists who wish to make a hands-on start in order to take advantage of modern Big Data technologies combined with Advanced Analytics.

A guide to the historical coal towns of the Big Sandy River Valley that provides brief histories of each town, descriptions of the buildings and structures that remain, and insight into the town's residents.

A handy reference guide for data analysts and data scientists to fetch "Value" out of big data analytics using Spark on Hadoop ClustersAbout This Book* Practical tutorial with real-world examples that explores Spark on Hadoop clusters* This book is based on the latest version of Apache Spark and Hadoop integrated with the most commonly used tools* Learn about all the Spark stack components including the latest topics such as DataFrames, DataSets, and SparkRWho This Book Is ForThough this book is primarily aimed at data analysts and data scientists, it will also help architects, programmers, and practitioners. Knowledge of either Spark or Hadoop would be beneficial. It is assumed that you have basic programming background in Scala, Python, SQL, or R programming with basic Linux experience. Working experience within big data environments is not mandatory.What You Will Learn* Find out about and implement the tools and techniques of big data analytics using Spark on Hadoop clusters* Understand all the Hadoop and Spark ecosystem components and how Spark replaced MapReduce* Get to know all the Spark components: Spark Core, Spark SQL, DataFrames, DataSets, Streaming, MLLib, and Graphx* See batch and real-time data analytics using Spark Core, Spark SQL, and Spark Streaming* Get to grips with data science and machine learning using MLLib, H2O, Hivemall, Graphx, and SparkR* Get an introduction to all the new tools (based on Notebooks, Data Flow, and Spark as a Service) and their integrations with Spark and HadoopIn DetailThis book explains the fundamentals of Apache Spark and Hadoop, and how they are easily integrated together with the most commonly used tools and techniques. All the Spark components-Spark Core, Spark SQL, DataFrames, Data sets, Streaming, MLlib, Graphx, and Hadoop core components-HDFS, MapReduce, and Yarn are explored in greater depth with implementation examples on Spark and Hadoop clusters.The big data analytics industry is moving away from MapReduce to Spark. In this book, the advantages of Spark over MapReduce are explained at great depth so you can reap the benefits of in-memory speeds. The DataFrames API, Data Sources API, and new Data sets API are explained so you can build big data analytical applications.We'll explore real-time data analytics using Spark Streaming with Apache Kafka and HBase to help you build streaming applications. You'll get to know the machine learning techniques using MLLib and SparkR, and Graph Analytics with the GraphX component of Spark.You will also get the opportunity to start working with web-based notebooks such as Jupyter, Apache Zeppelin, and the data flow tool Apache NiFi to analyze and visualize data.

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.

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

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

???PMBOK??(?5?)?????,???PMBOK??(?5?)????,??47????????????????????????,?????.?????????????,??????,??????,???????.

????????(?????????)?????(??????????).????AVL?????,?????,?????,????,?????????,???????????.

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.

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.

This proceedings book gathers papers presented at the 4th International Conference on Advanced Engineering Theory and Applications 2017 (AETA 2017), held on 7–9 December 2017 at Ton Duc Thang University, Ho Chi Minh City, Vietnam. It presents selected papers on 13 topical areas, including robotics, control systems, telecommunications, computer science and more. All selected papers represent interesting ideas and collectively provide a state-of-the-art overview. Readers will find intriguing papers on the design and implementation of control algorithms for aerial and underwater robots, for mechanical systems, efficient protocols for vehicular ad hoc networks, motor control, image and signal processing, energy saving, optimization methods in various fields of electrical engineering, and others. The book also offers a valuable resource for practitioners who want to apply the content discussed to solve real-life problems in their challenging applications. It also addresses common and related subjects in modern electric, electronic and related technologies. As such, it will benefit all scientists and engineers working in the above-mentioned fields of application.

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

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.

?????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 8th International Conference on Big Data Analytics, BDA 2020, which took place during December 15-18, 2020, in Sonepat, India. The 11 full and 3 short papers included in this volume were carefully reviewed and selected from 48 submissions; the book also contains 4 invited and 3 tutorial papers. The contributions were organized in topical sections named as follows: data science systems; data science architectures; big data analytics in healthcare; information interchange of Web data resources; and business analytics.

2008???????????????????? 50????????49?????? 35??????????????? 2012????????????????? ???50?????? ???????????????????????????? ???????????????????????????? ???????? ????????????????? ?????????????? ???????????????????? ???????????????????20 ?????????1??????????1? ??????????????10? ????????2012??????????1? ??????????????2012????????????? ??????????2012????????????

?????????????10??????????!? ?2009?????????????????????????????
?2012????????????????????????? ???????????????
???????????????????????911????? ?2007?????????????????????????????
???????? (big data)???
??
????????????????????! ???
??????????????????????????????????????(????:???????????????????) ??????
??
?????????????????;??
???????????????????? ???
??
??(???????????????:
???8?) ????????????: ?????????????????????????
???
??????????????????????????? ??????????????????
????????????????????????????????????
???????????????????????????????????????
???
?????????????????????????????????????

There’s a lot of information about big data technologies, but splicing these technologies into an end-to-end enterprise data platform is a daunting task not widely covered. With this practical book, you’ll learn how to build big data infrastructure both on-premises and in the cloud and successfully architect a modern data platform. Ideal for enterprise architects, IT managers, application architects, and data engineers, this book shows you how to overcome the many challenges that emerge during Hadoop projects. You’ll explore the vast landscape of tools available in the Hadoop and big data realm in a thorough technical primer before diving into: Infrastructure: Look at all component layers in a modern data platform, from the server to the data center, to establish a solid foundation for data in your enterprise Platform: Understand aspects of deployment, operation, security, high availability, and disaster recovery, along with everything you need to know to integrate your platform with the rest of your enterprise IT Taking Hadoop to the cloud: Learn the important architectural aspects of running a big data platform in the cloud while maintaining enterprise security and high availability

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

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

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.

The Champion's Guide to Inner Excellence Transform performance by changing the way your think, feel and behave. Direct, blunt, and brutally honest, Grover breaks down what it takes to be unstoppable: you keep going when everyone else is giving up, you thrive under pressure, you never let your emotions make you weak. What do your next 1 year look like? If you find yourself shrugging, you need to this book. A “wait and see” approach is no longer an option for you. It’s time to take action. This book will help you solve some of their most pressing challenges, including: - Management development - Performance management - Ethics - Respect - Diversity and inclusion - Employee engagement - Change - Onboarding and , How to adopt a positive mindset How to repair broken relationships How to resolve conflict successfully How to influence others How to minimize stress and gain energy How to be more creative more more... Unleash The Beast!

Utilize this practical and easy-to-follow guide to modernize traditional enterprise data warehouse and business intelligence environments with next-generation big data technologies. Next-Generation Big Data takes a holistic approach, covering the most important aspects of modern enterprise big data. The book covers not only the main technology stack but also the next-generation tools and applications used for big data warehousing, data warehouse optimization, real-time and batch data ingestion and processing, real-time data visualization, big data governance, data wrangling, big data cloud deployments, and distributed in-memory big data computing. Finally, the book has an extensive and detailed coverage of big data case studies from Navistar, Cerner, British Telecom, Shopzilla, Thomson Reuters, and Mastercard. What You’ll Learn Install Apache Kudu, Impala, and Spark to modernize enterprise data warehouse and business intelligence environments, complete with real-world, easy-to-follow examples, and practical advice Integrate HBase, Solr, Oracle, SQL Server, MySQL, Flume, Kafka, HDFS, and Amazon S3 with Apache Kudu, Impala, and Spark Use StreamSets, Talend, Pentaho, and CDAP for real-time and batch data ingestion and processing Utilize Trifacta, Alteryx, and Datameer for data wrangling and interactive data processing Turbocharge Spark with Alluxio, a distributed in-memory storage platform Deploy big data in the cloud using Cloudera Director Perform real-time data visualization and time series analysis using Zoomdata, Apache Kudu, Impala, and Spark Understand enterprise big data topics such as big data governance, metadata management, data lineage, impact analysis, and policy enforcement, and how to use Cloudera Navigator to perform common data governance tasks Implement big data use cases such as big data warehousing, data warehouse optimization, Internet of Things, real-time data ingestion and analytics, complex event processing, and scalable predictive modeling Study real-world big data case studies from innovative companies, including Navistar, Cerner, British Telecom, Shopzilla, Thomson Reuters, and Mastercard Who This Book Is For BI and big data warehouse professionals interested in gaining practical and real-world insight into next-generation big data processing and analytics using Apache Kudu, Impala, and Spark; and those who want to learn more about other advanced enterprise topics

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.

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

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.

A comprehensive end-to-end guide that gives hands-on practice in big data and Artificial Intelligence Key Features Learn to build and run a big data application with sample code Explore examples to implement activities that a big data architect performs Use Machine Learning and AI for structured and unstructured data Book Description The big data architects are the “masters” of data, and hold high value in today’s market. Handling big data, be it of good or bad quality, is not an easy task. The prime job for any big data architect is to build an end-to-end big data solution that integrates data from different sources and analyzes it to find useful, hidden insights. Big Data Architect’s Handbook takes you through developing a complete, end-to-end big data pipeline, which will lay the foundation for you and provide the necessary knowledge required to be an architect in big data. Right from understanding the design considerations to implementing a solid, efficient, and scalable data pipeline, this book walks you through all the essential aspects of big data. It also gives you an overview of how you can leverage the power of various big data tools such as Apache Hadoop and ElasticSearch in order to bring them together and build an efficient big data solution. By the end of this book, you will be able to build your own design system which integrates, maintains, visualizes, and monitors your data. In addition, you will have a smooth design flow in each process, putting insights in action. What you will learn Learn Hadoop Ecosystem and Apache projects Understand, compare NoSQL database and essential software architecture Cloud infrastructure design considerations for big data Explore application scenario of big data tools for daily activities Learn to analyze and visualize results to uncover valuable insights Build and run a big data application with sample code from end to end Apply Machine Learning and AI to perform big data intelligence Practice the daily activities performed by big data architects Who this book is for Big Data Architect’s Handbook is for you if you are an aspiring data professional, developer, or IT enthusiast who aims to be an all-round architect in big data. This book is your one-stop solution to enhance your knowledge and carry out easy to complex activities required to become a big data architect.

Do you want to write historical fiction? Join Meredith Allard, the executive editor of The Copperfield Review, the award-winning literary journal for readers and writers of historical fiction, as she shares tips and tricks for creating believable historical worlds through targeted research and a vivid imagination. Give in to your daydreams. Do the work. Let your creativity loose into the world so you can share your love of history and your passion for the written word with others.

This book presents the current trends, technologies, and challenges in Big Data in the diversified field of engineering and sciences. It covers the applications of Big Data ranging from conventional fields of mechanical engineering, civil engineering to electronics, electrical, and computer science to areas in pharmaceutical and biological sciences. This book consists of contributions from various authors from all sectors of academia and industries, demonstrating the imperative application of Big Data for the decision-making process in sectors where the volume, variety, and velocity of information keep increasing. The book is a useful reference for graduate students, researchers and scientists interested in exploring the potential of Big Data in the application of engineering areas.

C?C++????

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.

Get started with distributed computing using PySpark, a single unified framework to solve end-to-end data analytics at scale Key Features Discover how to convert huge amounts of raw data into meaningful and actionable insights Use Spark's unified analytics engine for end-to-end analytics, from data preparation to predictive analytics Perform data ingestion, cleansing, and integration for ML, data analytics, and data visualization Book Description Apache Spark is a unified data analytics engine designed to process huge volumes of data quickly and efficiently. PySpark is Apache Spark's Python language API, which offers Python developers an easy-to-use scalable data analytics framework. Essential PySpark for Scalable Data Analytics starts by exploring the distributed computing paradigm and provides a high-level overview of Apache Spark. You'll begin your analytics journey with the data engineering process, learning how to perform data ingestion, cleansing, and integration at scale. This book helps you build real-time analytics pipelines that help you gain insights faster. You'll then discover methods for building cloud-based data lakes, and explore Delta Lake, which brings reliability to data lakes. The book also covers Data Lakehouse, an emerging paradigm, which combines the structure and performance of a data warehouse with the scalability of cloud-based data lakes. Later, you'll perform scalable data science and machine learning tasks using PySpark, such as data preparation, feature engineering, and model training and productionization. Finally, you'll learn ways to scale out standard Python ML libraries along with a new pandas API on top of PySpark called Koalas. By the end of this PySpark book, you'll be able to harness the power of PySpark to solve business problems. What you will learn Understand the role of distributed computing in the world of big data Gain an appreciation for Apache Spark as the de facto go-to for big data processing Scale out your data analytics process using Apache Spark Build data pipelines using data lakes, and perform data visualization with PySpark and Spark SQL Leverage the cloud to build truly scalable and real-time data analytics applications Explore the applications of data science and scalable machine learning with PySpark Integrate your clean and curated data with BI and SQL analysis tools Who this book is for This book is for practicing data engineers, data scientists, data analysts, and data enthusiasts who are already using data analytics to explore distributed and scalable data analytics. Basic to intermediate knowledge of the disciplines of data engineering, data science, and SQL analytics is expected. General proficiency in using any programming language, especially Python, and working knowledge of performing data analytics using frameworks such as pandas and SQL will help you to get the most out of this 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'll understand how Kafka works and how it's 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 design Pick 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 Get detailed use-cases

See a Mesos-based big data stack created and the components used. You will use currently available Apache full and incubating systems. The components are introduced by example and you learn how they work together. In the Complete Guide to Open Source Big Data Stack, the author begins by creating a private cloud and then installs and examines Apache Brooklyn. After that, he uses each chapter to introduce one piece of the big data stack—sharing how to source the software and how to install it. You learn by simple example, step by step and chapter by chapter, as a real big data stack is created. The book concentrates on Apache-based systems and shares detailed examples of cloud storage, release management, resource management, processing, queuing, frameworks, data visualization, and more. What You'll Learn Install a private cloud onto the local cluster using Apache cloud stack Source, install, and configure Apache: Brooklyn, Mesos, Kafka, and Zeppelin See how Brooklyn can be used to install Mule ESB on a cluster and Cassandra in the cloud Install and use DCOS for big data processing Use Apache Spark for big data stack data processing Who This Book Is For Developers, architects, IT project managers, database administrators, and others charged with developing or supporting a big data system. It is also for anyone interested in Hadoop or big data, and those experiencing problems with data size.

This book is a step-by-step guide for learning 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. It covers Spark core and its add-on libraries, including Spark SQL, Spark Streaming, GraphX, MLlib, and Spark ML. Big Data Analytics with Spark shows you how to use Spark and leverage its easy-to-use features to increase your productivity. You learn to perform fast data analysis using its in-memory caching and advanced execution engine, employ in-memory computing capabilities for building high-performance machine learning and low-latency interactive analytics applications, and much more. Moreover, the book shows you how to use Spark as a single integrated platform for a variety of data processing tasks, including ETL pipelines, BI, live data stream processing, graph analytics, and machine learning. The book also includes a chapter on Scala, the hottest functional programming language, and the language 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, such as HDFS, Avro, Parquet, Kafka, Cassandra, HBase, Mesos, and so on. It also provides an introduction to machine learning and graph concepts. 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 have is some programming knowledge in any language.

"Authors Javier Luraschi, Kevin Kuo, and Edgar Ruiz show you how to combine R with Spark to analyze data at scale. This book covers relevant data science topics, cluster computing, and issues that will interest even the most advanced users."-- Back cover.

Learn about the fastest growing open source project in the world, and how it revolutionizes big data analyticsAbout This Book* Exclusive guide that covers how to get up and running with fast data processing using Apache Spark* Explore and exploit various possibilities with Apache Spark using real-world use cases in this book* Want to perform efficient data processing at real time? This book will be your one-stop solution.Who This Book Is ForThis guide appeals to Big Data engineers, analysts, architects, software engineers, even technical managers who need to perform efficient data processing on Hadoop at real time. Basic familiarity with Java or Scala will be helpful.The assumption is that readers will be from a mixed background, but would be typically people with background in engineering/data science and want to understand how Spark can help them on their analytics journey.What you will learn* Overview Big Data Analytics and its importance for organizations and data professionals.* Delve into Spark to see how it is different from existing processing platforms* Understand the intricacies of various file formats, and how to process them with Apache Spark.* Realize how to deploy Spark with YARN, MESOS or a Stand-alone cluster manager.* Learn the concepts of Spark SQL, SchemaRDD, Caching, Spark UDFs and working with Hive and Parquet file formats* Understand the architecture of Spark MLLib while discussing some of the off-the-shelf algorithms that come with Spark.* Introduce yourself to SparkR and walk through the details of data munging including selecting, aggregating and grouping data using R studio.* Walk through the importance of Graph computation and the graph processing systems available in the market* Check the real world example of Spark by building a recommendation engine with Spark using collaborative filtering* Use a telco data set, to predict customer churn using RegressionIn DetailSpark juggernaut keeps on rolling and getting more and more momentum each day. The core challenge are they key capabilities in Spark (Spark SQL, Spark Streaming, Spark ML, Spark R, Graph X) etc. Having understood the key capabilities, it is important to understand how Spark can be used, in terms of being installed as a Standalone framework or as a part of existing Hadoop installation and configuring with Yarn and Mesos.The next part of the journey after installation is using key components, APIs, Clustering, machine learning APIs, data pipelines, parallel programming. It is important to understand why each framework component is key, how widely it is being used, its stability and pertinent use cases.Once we understand the individual components, we will take a couple of real life advanced analytics examples like: * Building a Recommendation system* Predicting customer churn The objective of these real life examples is to give the reader confidence of using Spark for real-world problems.

This book is about how to integrate full-stack open source big data architecture and how 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 datasets 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. The 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'll learn How to make big data architecture without using complex Greek letter architectures. How to build a cheap but effective cluster infrastructure. How to make queries, reports, and graphs that business demands. How to manage and exploit unstructured and No-SQL data sources. How use tools to monitor the performance of your architecture. How to integrate all technologies and decide which replace and which reinforce. Who This Book Is For This book is for developers, data architects, and data scientists looking for how to integrate the most successful big data open stack architecture and how to choose the correct technology in every layer.