

Spark The Definitive Guide Big Data Processing Made Simple

Summary The Spark distributed data processing platform provides an easy-to-implement tool for ingesting, streaming, and processing data from any source. In *Spark in Action, Second Edition*, you'll learn to take advantage of Spark's core features and incredible processing speed, with applications including real-time computation, delayed evaluation, and machine learning. Spark skills are a hot commodity in enterprises worldwide, and with Spark's powerful and flexible Java APIs, you can reap all the benefits without first learning Scala or Hadoop. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Analyzing enterprise data starts by reading, filtering, and merging files and streams from many sources. The Spark data processing engine handles this varied volume like a champ, delivering speeds 100 times faster than Hadoop systems. Thanks to SQL support, an intuitive interface, and a straightforward multilanguage API, you can use Spark without learning a complex new ecosystem. About the book *Spark in Action, Second Edition*, teaches you to create end-to-end analytics applications. In this entirely new book, you'll learn from interesting Java-based examples, including a complete data pipeline for processing NASA satellite data. And you'll discover Java, Python, and Scala code samples hosted on GitHub that you can explore and adapt, plus appendixes that give you a cheat sheet for installing tools and understanding Spark-specific terms. What's inside Writing Spark applications in Java Spark application architecture Ingestion through files, databases,

Access PDF Spark The Definitive Guide Big Data Processing Made Simple

streaming, and Elasticsearch Querying distributed datasets with Spark SQL About the reader This book does not assume previous experience with Spark, Scala, or Hadoop. About the author Jean-Georges Perrin is an experienced data and software architect. He is France's first IBM Champion and has been honored for 12 consecutive years. Table of Contents PART 1 - THE THEORY CRIPPLED BY AWESOME EXAMPLES 1 So, what is Spark, anyway? 2 Architecture and flow 3 The majestic role of the dataframe 4 Fundamentally lazy 5 Building a simple app for deployment 6 Deploying your simple app PART 2 - INGESTION 7 Ingestion from files 8 Ingestion from databases 9 Advanced ingestion: finding data sources and building your own 10 Ingestion through structured streaming PART 3 - TRANSFORMING YOUR DATA 11 Working with SQL 12 Transforming your data 13 Transforming entire documents 14 Extending transformations with user-defined functions 15 Aggregating your data PART 4 - GOING FURTHER 16 Cache and checkpoint: Enhancing Spark's performances 17 Exporting data and building full data pipelines 18 Exploring deployment Combine the power of Apache Spark and Python to build effective big data applications Key Features Perform effective data processing, machine learning, and analytics using PySpark Overcome challenges in developing and deploying Spark solutions using Python Explore recipes for efficiently combining Python and Apache Spark to process data Book Description Apache Spark is an open source framework for efficient cluster computing with a strong interface for data parallelism and fault tolerance. The PySpark Cookbook presents effective and time-saving recipes for leveraging the power of Python and putting it to use in the Spark ecosystem. You'll start by learning the Apache Spark architecture and how to set up a Python environment for Spark. You'll then get familiar with the modules available in PySpark and start

Access PDF Spark The Definitive Guide Big Data Processing Made Simple

using them effortlessly. In addition to this, you'll discover how to abstract data with RDDs and DataFrames, and understand the streaming capabilities of PySpark. You'll then move on to using ML and MLlib in order to solve any problems related to the machine learning capabilities of PySpark and use GraphFrames to solve graph-processing problems. Finally, you will explore how to deploy your applications to the cloud using the spark-submit command. By the end of this book, you will be able to use the Python API for Apache Spark to solve any problems associated with building data-intensive applications. What you will learn

- Configure a local instance of PySpark in a virtual environment
- Install and configure Jupyter in local and multi-node environments
- Create DataFrames from JSON and a dictionary using pyspark.sql
- Explore regression and clustering models available in the ML module
- Use DataFrames to transform data used for modeling
- Connect to PubNub and perform aggregations on streams

Who this book is for The PySpark Cookbook is for you if you are a Python developer looking for hands-on recipes for using the Apache Spark 2.x ecosystem in the best possible way. A thorough understanding of Python (and some familiarity with Spark) will help you get the best out of the book.

Summary Spark in Action teaches you the theory and skills you need to effectively handle batch and streaming data using Spark. Fully updated for Spark 2.0. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Big data systems distribute datasets across clusters of machines, making it a challenge to efficiently query, stream, and interpret them. Spark can help. It is a processing system designed specifically for distributed data. It provides easy-to-use interfaces, along with the performance you need for production-quality analytics and machine learning. Spark 2 also

Access PDF Spark The Definitive Guide Big Data Processing Made Simple

adds improved programming APIs, better performance, and countless other upgrades. About the Book Spark in Action teaches you the theory and skills you need to effectively handle batch and streaming data using Spark. You'll get comfortable with the Spark CLI as you work through a few introductory examples. Then, you'll start programming Spark using its core APIs. Along the way, you'll work with structured data using Spark SQL, process near-real-time streaming data, apply machine learning algorithms, and munge graph data using Spark GraphX. For a zero-effort startup, you can download the preconfigured virtual machine ready for you to try the book's code. What's Inside Updated for Spark 2.0 Real-life case studies Spark DevOps with Docker Examples in Scala, and online in Java and Python About the Reader Written for experienced programmers with some background in big data or machine learning. About the Authors Petar Ze?evi? and Marko Bona?i are seasoned developers heavily involved in the Spark community. Table of Contents PART 1 - FIRST STEPS Introduction to Apache Spark Spark fundamentals Writing Spark applications The Spark API in depth PART 2 - MEET THE SPARK FAMILY Sparkling queries with Spark SQL Ingesting data with Spark Streaming Getting smart with MLlib ML: classification and clustering Connecting the dots with GraphX PART 3 - SPARK OPS Running Spark Running on a Spark standalone cluster Running on YARN and Mesos PART 4 - BRINGING IT TOGETHER Case study: real-time dashboard Deep learning on Spark with H2O

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.

Access PDF Spark The Definitive Guide Big Data Processing Made Simple

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

Apache Spark is amazing when everything clicks. But if you haven't seen the performance improvements you expected, or still don't feel confident enough to use Spark in production, this practical book is for you. Authors Holden Karau and Rachel Warren demonstrate performance optimizations to help your Spark queries run faster and handle larger data sizes, while using fewer resources. Ideal for software engineers, data engineers, developers, and system administrators working with large-scale data applications, this book describes techniques that can reduce data infrastructure costs and developer hours. Not only will you gain a more comprehensive understanding of Spark, you'll also learn how to make it sing. With this book, you'll explore: How Spark SQL's new interfaces improve performance over SQL's RDD data structure The choice between data joins in Core Spark and Spark SQL Techniques for getting the most out of standard RDD transformations How to work around performance issues in Spark's key/value pair paradigm Writing high-performance Spark code

without Scala or the JVM How to test for functionality and performance when applying suggested improvements Using Spark MLlib and Spark ML machine learning libraries Spark's Streaming components and external community packages

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.

What could you do with data if scalability wasn't a problem? With this hands-on guide, you'll learn how Apache Cassandra handles hundreds of terabytes of data while remaining highly available across multiple data centers -- capabilities that have attracted Facebook, Twitter, and other data-intensive companies. Cassandra: The Definitive Guide provides the technical details and practical examples you need to assess this database management system and put it to work in a production environment. Author Eben Hewitt demonstrates the advantages of Cassandra's nonrelational design, and pays special attention to data modeling. If you're a developer, DBA, application architect, or manager looking to solve a database scaling issue or future-proof your application, this guide shows you how to harness Cassandra's speed and flexibility. Understand the tenets of Cassandra's column-oriented structure Learn how to write, update, and read Cassandra data Discover how to add or remove nodes from the cluster as your application requires Examine a working application that translates from a relational model to Cassandra's data model Use examples for writing clients in Java, Python, and C# Use the JMX interface to monitor a cluster's usage, memory patterns, and more Tune memory settings, data storage, and caching for better performance

Access PDF Spark The Definitive Guide Big Data Processing Made Simple

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

Access PDF Spark The Definitive Guide Big Data Processing Made Simple

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

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

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

Advanced analytics on your Big Data with latest Apache Spark 2.x About This Book An advanced guide with a combination of instructions and practical examples to extend the most up-to-date Spark functionalities. Extend your data processing capabilities to process huge chunk of data in minimum time using advanced concepts in Spark. Master the art of real-time processing with the help of Apache Spark 2.x Who This Book Is For If you are a developer with

some experience with Spark and want to strengthen your knowledge of how to get around in the world of Spark, then this book is ideal for you. Basic knowledge of Linux, Hadoop and Spark is assumed. Reasonable knowledge of Scala is expected. What You Will Learn Examine Advanced Machine Learning and DeepLearning with MLlib, SparkML, SystemML, H2O and DeepLearning4J Study highly optimised unified batch and real-time data processing using SparkSQL and Structured Streaming Evaluate large-scale Graph Processing and Analysis using GraphX and GraphFrames Apply Apache Spark in Elastic deployments using Jupyter and Zeppelin Notebooks, Docker, Kubernetes and the IBM Cloud Understand internal details of cost based optimizers used in Catalyst, SystemML and GraphFrames Learn how specific parameter settings affect overall performance of an Apache Spark cluster Leverage Scala, R and python for your data science projects In Detail Apache Spark is an in-memory cluster-based parallel processing system that provides a wide range of functionalities such as graph processing, machine learning, stream processing, and SQL. This book aims to take your knowledge of Spark to the next level by teaching you how to expand Spark's functionality and implement your data flows and machine/deep learning programs on top of the platform. The book commences with an overview of the Spark ecosystem. It will introduce you to Project Tungsten and Catalyst, two of the major advancements of Apache Spark 2.x. You will understand how memory management and binary processing, cache-aware computation, and code generation are used to speed things up dramatically. The book extends to show how to incorporate H2O, SystemML, and Deeplearning4j for machine learning, and Jupyter Notebooks and Kubernetes/Docker for cloud-based Spark. During the course of the book, you will learn about the latest enhancements to Apache Spark 2.x, such as interactive querying of live data

Access PDF Spark The Definitive Guide Big Data Processing Made Simple

and unifying DataFrames and Datasets. You will also learn about the updates on the APIs and how DataFrames and Datasets affect SQL, machine learning, graph processing, and streaming. You will learn to use Spark as a big data operating system, understand how to implement advanced analytics on the new APIs, and explore how easy it is to use Spark in day-to-day tasks. **Style and approach** This book is an extensive guide to Apache Spark modules and tools and shows how Spark's functionality can be extended for real-time processing and storage with worked examples.

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.

Work with Apache Spark using Scala to deploy and set up single-node, multi-node, and high-availability clusters. This book discusses various components of Spark such as Spark Core, DataFrames, Datasets and SQL, Spark Streaming, Spark MLib, and R on Spark with the help of practical code snippets for each topic. Practical Apache Spark also covers the integration of Apache Spark with Kafka with examples. You'll follow a learn-to-do-by-yourself approach to

Access PDF Spark The Definitive Guide Big Data Processing Made Simple

learning – learn the concepts, practice the code snippets in Scala, and complete the assignments given to get an overall exposure. On completion, you'll have knowledge of the functional programming aspects of Scala, and hands-on expertise in various Spark components. You'll also become familiar with machine learning algorithms with real-time usage. What You Will Learn Discover the functional programming features of Scala Understand the complete architecture of Spark and its components Integrate Apache Spark with Hive and Kafka Use Spark SQL, DataFrames, and Datasets to process data using traditional SQL queries Work with different machine learning concepts and libraries using Spark's MLlib packages Who This Book Is For Developers and professionals who deal with batch and stream data processing.

No need to spend hours ploughing through endless data – let Spark, one of the fastest big data processing engines available, do the hard work for you. Key Features Get up and running with Apache Spark and Python Integrate Spark with AWS for real-time analytics Apply processed data streams to machine learning APIs of Apache Spark Book Description Processing big data in real time is challenging due to scalability, information consistency, and fault-tolerance. This book teaches you how to use Spark to make your overall analytical workflow faster and more efficient. You'll explore all core concepts and tools within the Spark ecosystem, such as Spark Streaming, the Spark Streaming API, machine learning extension, and structured streaming. You'll begin by learning data processing fundamentals using Resilient Distributed Datasets (RDDs), SQL, Datasets, and Dataframes APIs. After grasping these fundamentals, you'll move on to using Spark Streaming APIs to consume data in real time from TCP sockets, and integrate Amazon Web Services (AWS) for stream consumption. By the end of this book,

you'll not only have understood how to use machine learning extensions and structured streams but you'll also be able to apply Spark in your own upcoming big data projects. What you will learn Write your own Python programs that can interact with Spark Implement data stream consumption using Apache Spark Recognize common operations in Spark to process known data streams Integrate Spark streaming with Amazon Web Services (AWS) Create a collaborative filtering model with the movielens dataset Apply processed data streams to Spark machine learning APIs Who this book is for Data Processing with Apache Spark is for you if you are a software engineer, architect, or IT professional who wants to explore distributed systems and big data analytics. Although you don't need any knowledge of Spark, prior experience of working with Python is recommended.

In this practical book, four Cloudera data scientists present a set of self-contained patterns for performing large-scale data analysis with Spark. The authors bring Spark, statistical methods, and real-world data sets together to teach you how to approach analytics problems by example. You'll start with an introduction to Spark and its ecosystem, and then dive into patterns that apply common techniques—classification, collaborative filtering, and anomaly detection among others—to fields such as genomics, security, and finance. If you have an entry-level understanding of machine learning and statistics, and you program in Java, Python, or Scala, you'll find these patterns useful for working on your own data applications. Patterns include: Recommending music and the Audioscrobbler data set Predicting forest cover with decision trees Anomaly detection in network traffic with K-means clustering Understanding Wikipedia with Latent Semantic Analysis Analyzing co-occurrence networks with GraphX Geospatial and temporal data analysis on the New York City Taxi Trips data Estimating

financial risk through Monte Carlo simulation Analyzing genomics data and the BDG project
Analyzing neuroimaging data with PySpark and Thunder

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

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 The cult classic that can still change your life...Let the dice decide! This is the philosophy that changes the life of bored psychiatrist Luke Rhinehart—and in some ways changes the world as well. Because once you hand over your life to the dice, anything can happen. Entertaining, humorous, scary, shocking, subversive, *The Dice Man* is one of the cult bestsellers of our time. 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

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

Frank Kane's hands-on Spark training course, based on his bestselling Taming Big Data with Apache Spark and Python video, now available in a book.

Understand and analyze large data sets using Spark on a single system or on a cluster. About This Book Understand how Spark can be distributed across computing clusters Develop and run Spark jobs efficiently using Python A hands-on tutorial by Frank Kane with over 15 real-world examples teaching you Big Data processing with Spark Who This Book Is For If you are a data scientist or data analyst who wants to learn Big Data processing using Apache Spark and Python, this book is for you. If you have some programming experience in

Python, and want to learn how to process large amounts of data using Apache Spark, Frank Kane's *Taming Big Data with Apache Spark and Python* will also help you. What You Will Learn Find out how you can identify Big Data problems as Spark problems Install and run Apache Spark on your computer or on a cluster Analyze large data sets across many CPUs using Spark's Resilient Distributed Datasets Implement machine learning on Spark using the MLlib library Process continuous streams of data in real time using the Spark streaming module Perform complex network analysis using Spark's GraphX library Use Amazon's Elastic MapReduce service to run your Spark jobs on a cluster In Detail Frank Kane's *Taming Big Data with Apache Spark and Python* is your companion to learning Apache Spark in a hands-on manner. Frank will start you off by teaching you how to set up Spark on a single system or on a cluster, and you'll soon move on to analyzing large data sets using Spark RDD, and developing and running effective Spark jobs quickly using Python. Apache Spark has emerged as the next big thing in the Big Data domain – quickly rising from an ascending technology to an established superstar in just a matter of years. Spark allows you to quickly extract actionable insights from large amounts of data, on a real-time basis, making it an essential tool in many modern businesses. Frank has packed this book with over 15 interactive, fun-filled examples relevant to the

real world, and he will empower you to understand the Spark ecosystem and implement production-grade real-time Spark projects with ease. Style and approach Frank Kane's *Taming Big Data with Apache Spark and Python* is a hands-on tutorial with over 15 real-world examples carefully explained by Frank in a step-by-step manner. The examples vary in complexity, and you can move through them at your own pace.

The book describes the emergence of big data technologies and the role of Spark in the entire big data stack. It compares Spark and Hadoop and identifies the shortcomings of Hadoop that have been overcome by Spark. The book mainly focuses on the in-depth architecture of Spark and our understanding of Spark RDDs and how RDD complements big data's immutable nature, and solves it with lazy evaluation, cacheable and type inference. It also addresses advanced topics in Spark, starting with the basics of Scala and the core Spark framework, and exploring Spark data frames, machine learning using Mllib, graph analytics using Graph X and real-time processing with Apache Kafka, AWS Kinesis, and Azure Event Hub. It then goes on to investigate Spark using PySpark and R. Focusing on the current big data stack, the book examines the interaction with current big data tools, with Spark being the core processing layer for all types of data. The book is intended for data engineers and scientists

working on massive datasets and big data technologies in the cloud. In addition to industry professionals, it is helpful for aspiring data processing professionals and students working in big data processing and cloud computing environments. Discover how Apache Hadoop can unleash the power of your data. This comprehensive resource shows you how to build and maintain reliable, scalable, distributed systems with the Hadoop framework -- an open source implementation of MapReduce, the algorithm on which Google built its empire. Programmers will find details for analyzing datasets of any size, and administrators will learn how to set up and run Hadoop clusters. This revised edition covers recent changes to Hadoop, including new features such as Hive, Sqoop, and Avro. It also provides illuminating case studies that illustrate how Hadoop is used to solve specific problems. Looking to get the most out of your data? This is your book. Use the Hadoop Distributed File System (HDFS) for storing large datasets, then run distributed computations over those datasets with MapReduce Become familiar with Hadoop's data and I/O building blocks for compression, data integrity, serialization, and persistence Discover common pitfalls and advanced features for writing real-world MapReduce programs Design, build, and administer a dedicated Hadoop cluster, or run Hadoop in the cloud Use Pig, a high-level query language for large-scale data processing

Analyze datasets with Hive, Hadoop's data warehousing system Take advantage of HBase, Hadoop's database for structured and semi-structured data Learn ZooKeeper, a toolkit of coordination primitives for building distributed systems "Now you have the opportunity to learn about Hadoop from a master -- not only of the technology, but also of common sense and plain talk." --Doug Cutting, Cloudera

Harness the power of Scala to program Spark and analyze tonnes of data in the blink of an eye! About This Book Learn Scala's sophisticated type system that combines Functional Programming and object-oriented concepts Work on a wide array of applications, from simple batch jobs to stream processing and machine learning Explore the most common as well as some complex use-cases to perform large-scale data analysis with Spark Who This Book Is For Anyone who wishes to learn how to perform data analysis by harnessing the power of Spark will find this book extremely useful. No knowledge of Spark or Scala is assumed, although prior programming experience (especially with other JVM languages) will be useful to pick up concepts quicker. What You Will Learn Understand object-oriented & functional programming concepts of Scala In-depth understanding of Scala collection APIs Work with RDD and DataFrame to learn Spark's core abstractions Analysing structured and unstructured data using

SparkSQL and GraphX Scalable and fault-tolerant streaming application development using Spark structured streaming Learn machine-learning best practices for classification, regression, dimensionality reduction, and recommendation system to build predictive models with widely used algorithms in Spark MLlib & ML Build clustering models to cluster a vast amount of data Understand tuning, debugging, and monitoring Spark applications Deploy Spark applications on real clusters in Standalone, Mesos, and YARN In Detail Scala has been observing wide adoption over the past few years, especially in the field of data science and analytics. Spark, built on Scala, has gained a lot of recognition and is being used widely in productions. Thus, if you want to leverage the power of Scala and Spark to make sense of big data, this book is for you. The first part introduces you to Scala, helping you understand the object-oriented and functional programming concepts needed for Spark application development. It then moves on to Spark to cover the basic abstractions using RDD and DataFrame. This will help you develop scalable and fault-tolerant streaming applications by analyzing structured and unstructured data using SparkSQL, GraphX, and Spark structured streaming. Finally, the book moves on to some advanced topics, such as monitoring, configuration, debugging, testing, and deployment. You will also learn how to develop Spark applications using SparkR

and PySpark APIs, interactive data analytics using Zeppelin, and in-memory data processing with Alluxio. By the end of this book, you will have a thorough understanding of Spark, and you will be able to perform full-stack data analytics with a feel that no amount of data is too big. Style and approach Filled with practical examples and use cases, this book will not only help you get up and running with Spark, but will also take you farther down the road to becoming a data scientist.

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,

Access PDF Spark The Definitive Guide Big Data Processing Made Simple

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

Ready to unlock the power of your data? With 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. You'll find illuminating case studies that demonstrate how Hadoop is used to solve specific problems. This third edition covers recent changes to Hadoop, including material on the new MapReduce API, as well as MapReduce 2 and its more flexible execution model (YARN). Store large datasets with the Hadoop Distributed File System (HDFS) Run distributed computations with MapReduce Use Hadoop's data and I/O building blocks for compression, data integrity, serialization (including Avro), and persistence Discover common pitfalls and advanced features for writing real-world MapReduce programs Design, build, and administer a dedicated Hadoop cluster—or run Hadoop in the cloud Load data from relational databases into HDFS, using Sqoop Perform large-scale data processing with the Pig query language Analyze datasets with Hive, Hadoop's data warehousing system Take advantage of HBase for structured and semi-structured data, and ZooKeeper for building distributed systems

Analyze vast amounts of data in record time using Apache Spark with Databricks in the Cloud. Learn the fundamentals, and more, of running analytics on large clusters in Azure and AWS, using Apache 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 Apache 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 Apache 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 Apache 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 Apache Spark and Azure Databricks. The book is also recommended for people who want to get started in the analytics field, as it provides a strong foundation.

Apache Spark is a fast, scalable, and flexible open source distributed processing engine for big

data systems and is one of the most active open source big data projects to date. In just 24 lessons of one hour or less, Sams Teach Yourself Apache Spark in 24 Hours helps you build practical Big Data solutions that leverage Spark's amazing speed, scalability, simplicity, and versatility. This book's straightforward, step-by-step approach shows you how to deploy, program, optimize, manage, integrate, and extend Spark—now, and for years to come. You'll discover how to create powerful solutions encompassing cloud computing, real-time stream processing, machine learning, and more. Every lesson builds on what you've already learned, giving you a rock-solid foundation for real-world success. Whether you are a data analyst, data engineer, data scientist, or data steward, learning Spark will help you to advance your career or embark on a new career in the booming area of Big Data. Learn how to

- Discover what Apache Spark does and how it fits into the Big Data landscape
- Deploy and run Spark locally or in the cloud
- Interact with Spark from the shell
- Make the most of the Spark Cluster Architecture
- Develop Spark applications with Scala and functional Python
- Program with the Spark API, including transformations and actions
- Apply practical data engineering/analysis approaches designed for Spark
- Use Resilient Distributed Datasets (RDDs) for caching, persistence, and output
- Optimize Spark solution performance
- Use Spark with SQL (via Spark SQL) and with NoSQL (via Cassandra)
- Leverage cutting-edge functional programming techniques
- Extend Spark with streaming, R, and Sparkling Water
- Start building Spark-based machine learning and graph-processing applications
- Explore advanced messaging technologies, including Kafka
- Preview and prepare for Spark's next generation of innovations

Instructions walk you through common questions, issues, and tasks; Q-and-As, Quizzes, and Exercises build and test your knowledge; "Did You Know?" tips offer insider

Access PDF Spark The Definitive Guide Big Data Processing Made Simple

advice and shortcuts; and "Watch Out!" alerts help you avoid pitfalls. By the time you're finished, you'll be comfortable using Apache Spark to solve a wide spectrum of Big Data problems.

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

Every enterprise application creates data, whether it's log messages, metrics, user activity, outgoing messages, or something else. And how to move all of this data becomes nearly as important as the data itself. If you're an application architect, developer, or production engineer new to Apache Kafka, this practical guide shows you how to use this open source streaming platform to handle real-time data feeds. Engineers from Confluent and LinkedIn who are responsible for developing Kafka explain how to deploy production Kafka clusters, write reliable event-driven microservices, and build scalable stream-processing applications with this platform. Through detailed examples, you'll learn Kafka's design principles, reliability guarantees, key APIs, and architecture details, including the replication protocol, the controller, and the storage layer. Understand publish-subscribe messaging and how it fits in the big data ecosystem. Explore Kafka producers and consumers for writing and reading messages Understand Kafka patterns and use-case requirements to ensure reliable data delivery Get best practices for building data pipelines and applications with Kafka Manage Kafka in production, and learn to perform monitoring, tuning, and maintenance tasks Learn the most critical metrics among Kafka's operational measurements Explore how Kafka's stream delivery capabilities make it a perfect source for stream processing systems

Solve Data Analytics Problems with Spark, PySpark, and Related Open Source Tools Spark is at the heart of today's Big Data revolution, helping data professionals supercharge efficiency

and performance in a wide range of data processing and analytics tasks. In this guide, Big Data expert Jeffrey Aven covers all you need to know to leverage Spark, together with its extensions, subprojects, and wider ecosystem. Aven combines a language-agnostic introduction to foundational Spark concepts with extensive programming examples utilizing the popular and intuitive PySpark development environment. This guide's focus on Python makes it widely accessible to large audiences of data professionals, analysts, and developers—even those with little Hadoop or Spark experience. Aven's broad coverage ranges from basic to advanced Spark programming, and Spark SQL to machine learning. You'll learn how to efficiently manage all forms of data with Spark: streaming, structured, semi-structured, and unstructured. Throughout, concise topic overviews quickly get you up to speed, and extensive hands-on exercises prepare you to solve real problems. Coverage includes:

- Understand Spark's evolving role in the Big Data and Hadoop ecosystems
- Create Spark clusters using various deployment modes
- Control and optimize the operation of Spark clusters and applications
- Master Spark Core RDD API programming techniques
- Extend, accelerate, and optimize Spark routines with advanced API platform constructs, including shared variables, RDD storage, and partitioning
- Efficiently integrate Spark with both SQL and nonrelational data stores
- Perform stream processing and messaging with Spark Streaming and Apache Kafka
- Implement predictive modeling with SparkR and Spark MLlib

Build data-intensive applications locally and deploy at scale using the combined powers of Python and Spark 2.0

About This Book Learn why and how you can efficiently use Python to process data and build machine learning models in Apache Spark 2.0

Develop and deploy efficient, scalable real-time Spark solutions Take your understanding of using Spark with Python to the next level with this jump start guide Who This Book Is For If you are a Python developer who wants to learn about the Apache Spark 2.0 ecosystem, this book is for you. A firm understanding of Python is expected to get the best out of the book. Familiarity with Spark would be useful, but is not mandatory. What You Will Learn Learn about Apache Spark and the Spark 2.0 architecture Build and interact with Spark DataFrames using Spark SQL Learn how to solve graph and deep learning problems using GraphFrames and TensorFrames respectively Read, transform, and understand data and use it to train machine learning models Build machine learning models with MLlib and ML Learn how to submit your applications programmatically using spark-submit Deploy locally built applications to a cluster In Detail Apache Spark is an open source framework for efficient cluster computing with a strong interface for data parallelism and fault tolerance. This book will show you how to leverage the power of Python and put it to use in the Spark ecosystem. You will start by getting a firm understanding of the Spark 2.0 architecture and how to set up a Python environment for Spark. You will get familiar with the modules available in PySpark. You will learn how to abstract data with RDDs and DataFrames and understand the streaming capabilities of PySpark. Also, you will get a thorough overview of machine learning capabilities of PySpark using ML and MLlib, graph processing using GraphFrames, and polyglot persistence using Blaze. Finally,

you will learn how to deploy your applications to the cloud using the spark-submit command. By the end of this book, you will have established a firm understanding of the Spark Python API and how it can be used to build data-intensive applications. Style and approach This book takes a very comprehensive, step-by-step approach so you understand how the Spark ecosystem can be used with Python to develop efficient, scalable solutions. Every chapter is standalone and written in a very easy-to-understand manner, with a focus on both the hows and the whys of each concept.

“This beautiful book makes one of America’s favorite culinary pastimes accessible to anyone. The recipes are easy to follow yet refined.” —Marcus Samuelsson Now home cooks will discover their ideal grilling companion and coach in *The Grilling Book*.

Offering more than 350 foolproof recipes, dozens of luscious full-color photographs, crystal clear illustrations, and plenty of plainspoken, here’s-how-to-do-it guidelines, *The Grilling Book* welcomes you to everything that is sensational (and sensationally simple) about grilling. Here are thick, juicy steaks that need no more than salt and freshly cracked pepper to create an unforgettable meal; baby back ribs rendered succulent and fall-apart tender with flavor-rich rubs and a shellacking of barbecue sauce (plus the secret cheat method that makes them ridiculously easy to prepare); moist fish, seasoned and grilled quickly over a hot fire; irresistible grilled flatbreads, crispy on the outside and chewy on the inside, creating the perfect canvas for every topping you crave. Not to mention the salads, slaws, sides, and drinks that complete the perfect

grilled meal. Whether you've been grilling for years or have never bought a bag of charcoal in your life, *The Grilling Book* is the only book you'll need in your backyard this summer—or any summer. “Indispensable. Chock-full of recipes tested and approved by the trusted editors of *Bon Appétit*, it's sure to make you a master of all things grilling.” —Tom Colicchio “*Bon Appétit*'s new bible on live-fire cooking is one you'll want to keep close at hand every summer. If you're new to the grill, the easy-to-follow recipes, protein-specific technique tips, and time-saving tricks will help you get started.” —Matt Lee and Ted Lee

We all have a bit of creativity inside us, waiting to burn brighter. Sometimes all we need is encouragement. At other times, we need skills or know-how. And still at other times, what we need most is the chance to explore our interests, make mistakes, and have fun. This fill-in-the-blank, do-it-yourself guidebook will give you all these things and more as you discover what it means to be creative.

Develop applications for the big data landscape with Spark and Hadoop. This book also explains the role of Spark in developing scalable machine learning and analytics applications with Cloud technologies. *Beginning Apache Spark 2* gives you an introduction to Apache Spark and shows you how to work with it. Along the way, you'll discover resilient distributed datasets (RDDs); use Spark SQL for structured data; and learn stream processing and build real-time applications with Spark Structured Streaming. Furthermore, you'll learn the fundamentals of Spark ML for machine

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

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

