Spark The Definitive Guide

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.
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
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 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. Mastering Structured Streaming and Spark Streaming

Patterns for Learning from Data at Scale
Best Practices for Scaling and Optimizing Apache Spark
Beginning Apache Spark Using Azure Databricks
Data Engineering with Apache Spark, Delta Lake, and Lakehouse
Hadoop: The Definitive Guide

Perfect for readers of Song for a Whale and Counting by 7s, a neurodivergent girl campaigns for a memorial when she learns that her small

Scottish town used to burn witches simply because they were different. "A must-read for students and adults alike." -School Library Journal, Starred Review Ever since Ms. Murphy told us about the witch trials that happened centuries ago right here in Juniper, I can't stop thinking about them. Those people weren't magic. They were like me. Different like me. I'm autistic. I see things that others do not. I hear sounds that they can ignore. And sometimes I feel things all at once. I think about the witches, with no one to speak for them. Not everyone in

our small town understands. But if I keep trying, maybe someone will. I won't let the witches be forgotten. Because there is more to their story. Just like there is more to mine. Award-winning and neurodivergent author Elle McNicoll delivers an insightful and stirring debut about the European witch trials and a girl who refuses to relent in the fight for what she knows is right. 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 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 Complete Guide to Data Science with Hadoop—For Technical Professionals, Businesspeople, and Students Demand is soaring for professionals who can solve real data science problems with Hadoop and Spark. Practical Data Science with Hadoop® and Spark is your complete guide to doing just that. Drawing on immense experience with Hadoop and big data, three leading experts bring together everything you need: high-level concepts, deep-dive techniques, real-world use cases, practical applications, and hands-on

tutorials. The authors introduce the essentials of data science and the modern Hadoop ecosystem, explaining how Hadoop and Spark have evolved into an effective platform for solving data science problems at scale. In addition to comprehensive application coverage, the authors also provide useful guidance on the important steps of data ingestion, data munging, and visualization. Once the groundwork is in place, the authors focus on specific applications, including machine learning, predictive modeling for sentiment analysis, clustering for document

analysis, anomaly detection, and natural language processing (NLP). This guide provides a strong technical foundation for those who want to do practical data science, and also presents business-driven guidance on how to apply Hadoop and Spark to optimize ROI of data science initiatives. Learn What data science is, how it has evolved, and how to plan a data science career How data volume, variety, and velocity shape data science use cases Hadoop and its ecosystem, including HDFS, MapReduce, YARN, and Spark Data importation with Hive and

Spark Data quality, preprocessing, preparation, and modeling Visualization: surfacing insights from huge data sets Machine learning: classification, regression, clustering, and anomaly detection Algorithms and Hadoop tools for predictive modeling Cluster analysis and similarity functions Large-scale anomaly detection NLP: applying data science to human language

Stream Processing with Apache Spark

Deep Learning

Learning Spark

Data Warehousing, Analytics, and Machine Learning at Scale
Covers Apache Spark 3 with Examples in Java, Python, and Scala
Learning PySpark

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

Although interest in machine learning has reached a high point, lofty expectations often scuttle projects before they get very far. How can machine learning—especially deep neural networks—make a real difference in your organization? This hands-on guide not only provides the most practical information available on the subject, but also helps you get started building efficient deep learning networks. Authors Adam Gibson and Josh Patterson provide theory on deep learning before introducing their open-

source Deeplearning4j (DL4J) library for developing production-class workflows. Through real-world examples, you'll learn methods and strategies for training deep network architectures and running deep learning workflows on Spark and Hadoop with DL4J. Dive into machine learning concepts in general, as well as deep learning in particular Understand how deep networks evolved from neural network fundamentals Explore the major deep network architectures, including Convolutional and Recurrent Learn how to

map specific deep networks to the right problem Walk through the fundamentals of tuning general neural networks and specific deep network architectures Use vectorization techniques for different data types with DataVec, DL4J's workflow tool Learn how to use DL4J natively on Spark and Hadoop

Understand the complexities of modern-day data engineering platforms and explore strategies to deal with them with the help of use case scenarios led by an industry expert in big data Key Features Become

well-versed with the core concepts of Apache Spark and Delta Lake for building data platforms Learn how to ingest, process, and analyze data that can be later used for training machine learning models Understand how to operationalize data models in production using curated data Book Description In the world of ever-changing data and schemas, it is important to build data pipelines that can auto-adjust to changes. This book will help you build scalable data platforms that managers, data scientists, and data

analysts can rely on. Starting with an introduction to data engineering, along with its key concepts and architectures, this book will show you how to use Microsoft Azure Cloud services effectively for data engineering. You'll cover data lake design patterns and the different stages through which the data needs to flow in a typical data lake. Once you've explored the main features of Delta Lake to build data lakes with fast performance and governance in mind, you'll advance to implementing the lambda architecture using

Delta Lake. Packed with practical examples and code snippets, this book takes you through real-world examples based on production scenarios faced by the author in his 10 years of experience working with big data. Finally, you'll cover data lake deployment strategies that play an important role in provisioning the cloud resources and deploying the data pipelines in a repeatable and continuous way. By the end of this data engineering book, you'll know how to effectively deal with ever-changing data and create scalable data

pipelines to streamline data science, ML, and artificial intelligence (AI) tasks. What you will learn Discover the challenges you may face in the data engineering world Add ACID transactions to Apache Spark using Delta Lake Understand effective design strategies to build enterprise-grade data lakes Explore architectural and design patterns for building efficient data ingestion pipelines Orchestrate a data pipeline for preprocessing data using Apache Spark and Delta Lake APIs Automate deployment and

monitoring of data pipelines in production Get to grips with securing, monitoring, and managing data pipelines models efficiently Who this book is for This book is for aspiring data engineers and data analysts who are new to the world of data engineering and are looking for a practical guide to building scalable data platforms. If you already work with PySpark and want to use Delta Lake for data engineering, you'll find this book useful. Basic knowledge of Python, Spark, and SQL is expected.

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 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.
With DataFrame, Spark SQL, Structured Streaming, and Spark Machine Learning Library
A Kind of Spark
Big Data Analytics with Spark
Beginning Apache Spark 3
PySpark Cookbook
Mastering Spark with R

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

Learn how easy it is to apply sophisticated statistical and machine learning methods to real-world problems when you build on top of the Google Cloud Platform (GCP). This hands-on guide shows developers entering the data science field how to implement an end-to-end

data pipeline, using statistical and machine learning methods and tools on GCP. Through the course of the book, you'll work through a sample business decision by employing a variety of data science approaches. Follow along by implementing these statistical and machine learning solutions in your own project on GCP, and discover how this platform provides a transformative and more collaborative way of doing data science. You'll learn how to: Automate and schedule data ingest, using an App Engine application Create and populate a dashboard in Google Data Studio Build a real-time analysis pipeline to carry out streaming

analytics Conduct interactive data exploration with Google BigQuery Create a Bayesian model on a Cloud Dataproc cluster Build a logistic regression machine-learning model with Spark Compute time-aggregate features with a Cloud Dataflow pipeline Create a high-performing prediction model with TensorFlow Use your deployed model as a microservice you can access from both batch and real-time pipelines Take a journey toward discovering, learning, and using Apache Spark 3.0. In this book, you will gain expertise on the powerful and efficient distributed data processing engine inside of Apache Spark; its user-friendly, comprehensive,

and flexible programming model for processing data in batch and streaming; and the scalable machine learning algorithms and practical utilities to build machine learning applications. Beginning Apache Spark 3 begins by explaining different ways of interacting with Apache Spark, such as Spark Concepts and Architecture, and Spark Unified Stack. Next, it offers an overview of Spark SQL before moving on to its advanced features. It covers tips and techniques for dealing with performance issues, followed by an overview of the structured streaming processing engine. It concludes with a demonstration of how to develop machine learning applications

using Spark MLlib and how to manage the machine learning development lifecycle. This book is packed with practical examples and code snippets to help you master concepts and features immediately after they are covered in each section. After reading this book, you will have the knowledge required to build your own big data pipelines, applications, and machine learning applications. What You Will Learn Master the Spark unified data analytics engine and its various components Work in tandem to provide a scalable, fault tolerant and performant data processing engine Leverage the user-friendly and flexible programming model

to perform simple to complex data analytics using dataframe and Spark SQL Develop machine learning applications using Spark MLlib Manage the machine learning development lifecycle using MLflow Who This Book Is For Data scientists, data engineers and software developers.

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 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 Zečević and Marko Bonaći 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
Google BigQuery: The Definitive Guide
Spark
Cassandra: The Definitive Guide
Data Science on the Google Cloud Platform
The Definitive Guide

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 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.
Over insightful 90 recipes to get lightning-fast analytics with Apache Spark About This Book Use Apache Spark for data processing with these hands-on recipes Implement end-to-end, large-scale data analysis better than ever before Work with powerful libraries such as MLLib, SciPy, NumPy, and Pandas to gain insights from your data Who This Book Is For This book is for novice and intermediate level data science professionals and data analysts who want to solve data science

problems with a distributed computing framework. Basic experience with data science implementation tasks is expected. Data science professionals looking to skill up and gain an edge in the field will find this book helpful. What You Will Learn Explore the topics of data mining, text mining, Natural Language Processing, information retrieval, and machine learning. Solve real-world analytical problems with large data sets. Address data science challenges with analytical tools on a distributed system like Spark (apt for iterative algorithms), which offers in-memory processing and more

flexibility for data analysis at scale. Get hands-on experience with algorithms like Classification, regression, and recommendation on real datasets using Spark MLLib package. Learn about numerical and scientific computing using NumPy and SciPy on Spark. Use Predictive Model Markup Language (PMML) in Spark for statistical data mining models. In Detail Spark has emerged as the most promising big data analytics engine for data science professionals. The true power and value of Apache Spark lies in its ability to execute data science tasks with speed and accuracy. Spark's selling point is that it

combines ETL, batch analytics, real-time stream analysis, machine learning, graph processing, and visualizations. It lets you tackle the complexities that come with raw unstructured data sets with ease. This guide will get you comfortable and confident performing data science tasks with Spark. You will learn about implementations including distributed deep learning, numerical computing, and scalable machine learning. You will be shown effective solutions to problematic concepts in data science using Spark's data science libraries such as MLLib, Pandas, NumPy, SciPy, and more. These simple

and efficient recipes will show you how to implement algorithms and optimize your work. Style and approach This book contains a comprehensive range of recipes designed to help you learn the fundamentals and tackle the difficulties of data science. This book outlines practical steps to produce powerful insights into Big Data through a recipe-based approach.

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

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
Practical Apache Spark
Spark: The Definitive Guide
Spark in Action
Apache Spark for Data Science Cookbook
Storage and Analysis at Internet Scale
Distributed Data at Web Scale

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

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.

If your organization is looking for a storage solution to accommodate a virtually endless amount of data, this book will show you how Apache HBase can fulfill your needs. As the open source implementation of Google's BigTable architecture, HBase scales to billions of rows and millions of columns, while ensuring that write and read performance remain constant.HBase: The Definitive Guideprovides the details you require, whether you simply want to evaluate this high-performance, non-relational database, or put it into practice right away. HBase's

adoption rate is beginning to climb, and several IT executives are asking pointed questions about this high-capacity database. This is the only book available to give you meaningful answers. Learn how to distribute large datasets across an inexpensive cluster of commodity servers Develop HBase clients in many programming languages, including Java, Python, and Ruby Get details on HBase's primary storage system, HDFS—Hadoop's distributed and replicated filesystem Learn how HBase's native interface to Hadoop's MapReduce framework enables easy development and execution of batch jobs that can scan entire tables Discover the integration between

HBase and other facets of the Apache Hadoop project 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 cloudGet started with Databricks using SQL and Python in either Microsoft Azure or AWSUnderstand 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.

Real-Time Data and Stream Processing at Scale

Using the Scala API

Big Data Processing with Apache Spark

Advanced Analytics with Spark

The Life, Times, and Music of Merle Haggard

Over 60 recipes for implementing big data processing and

analytics using Apache Spark and Python
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 FeaturesGet up and running with Apache Spark and PythonIntegrate Spark with AWS for real-time analyticsApply processed data streams to machine learning APIs of Apache SparkBook 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 learnWrite your own Python programs that can interact with SparkImplement data stream consumption using Apache SparkRecognize common operations in Spark to process known data streamsIntegrate Spark streaming with Amazon Web Services (AWS)Create a

collaborative filtering model with the movielens datasetApply processed data streams to Spark machine learning APIsWho 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.

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 in all domains is getting bigger. How can you work with it efficiently? Recently updated for Spark 1.3, this book introduces Apache Spark, the open source cluster computing system that makes data analytics fast to write and fast to run. With Spark, you can tackle big datasets

quickly through simple APIs in Python, Java, and Scala. This edition includes new information on Spark SQL, Spark Streaming, setup, and Maven coordinates. Written by the developers of Spark, this book will have data scientists and engineers up and running in no time. You'll learn how to express parallel jobs with just a few lines of code, and cover applications from simple batch jobs to stream processing and machine learning. Quickly dive into Spark capabilities such as distributed datasets, in-memory caching, and the interactive shell Leverage Spark's powerful built-in libraries, including Spark SQL, Spark Streaming, and MLlib Use one programming paradigm instead of mixing and matching tools like Hive, Hadoop, Mahout, and Storm Learn how to deploy interactive, batch, and streaming applications

Connect to data sources including HDFS, Hive, JSON, and S3
Master advanced topics like data partitioning and shared variables

In a future United States under the power of a charismatic leader, everyone gets the Mark at age thirteen. The Mark lets citizen shop, go to school, and even get medical care—but without it, you're on your own. Few refuse to get the Mark. Those who do . . . disappear. Parents are looking for fiction that makes Christianity exciting for kids. This series is an alternative to the Hunger Games series and other dark dystopian fiction. It’s packed with action and intrigue, but the message is written from a Christian worldview. Logan Langly went to get his Mark but backed out at the last minute. Ever since, he’s been on the run from

government agents and on a quest to find his sister Lily, who disappeared when she went to get her Mark five years ago. His journey leads him to befriend the Dust, a network of Markless who oppose the iron-grip rule of the government. On the way to the capital to find Lily, the Dust receive some startling information from the Markless community, warning that humanity is now entering the End of Days. Spark introduces nine-year-old Ali, a beggar living in the Dark Lands city of al-Balat. Ali meets a stranger who gives her his tablet, a portal to a tech world that Ali never knew existed. But one day, the tablet begins to communicate back to her—and takes her on a journey that will cross her path with exiled Logan Langly, Chancellor Cylis, and the fierce battle for power that spans reality and

the virtual world. Meets national education standards.

Data Analytics with Spark Using Python

A Practitioner's Approach

An Architecture for Fast and General Data Processing on Large Clusters

Designing and Building Effective Analytics at Scale

Efficiently tackle large datasets and big data analysis with Spark and Python

A Practitioner's Guide to Using Spark for Large Scale Data Analysis

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

Imagine what you could do if scalability wasn't a problem. With this hands-on guide, you'll learn how the Cassandra

database management system handles hundreds of terabytes of data while remaining highly available across multiple data centers. This expanded second edition—updated for Cassandra 3.0—provides the technical details and practical examples you need to put this database to work in a production environment. Authors Jeff Carpenter and Eben Hewitt demonstrate the advantages of Cassandra's non-relational design, with special attention to data modeling. If you're a developer, DBA, or application architect looking to solve a database scaling issue or future-proof your application, this guide helps you harness Cassandra's speed and flexibility. Understand Cassandra's distributed and decentralized structure Use the Cassandra Query Language (CQL) and cqlsh—the CQL shell Create a working data model

and compare it with an equivalent relational model Develop sample applications using client drivers for languages including Java, Python, and Node.js Explore cluster topology and learn how nodes exchange data Maintain a high level of performance in your cluster Deploy Cassandra on site, in the Cloud, or with Docker Integrate Cassandra with Spark, Hadoop, Elasticsearch, Solr, and Lucene

The definitive biography of country legend Merle Haggard by the New York Times bestselling biographer of Clint Eastwood, Cary Grant, The Eagles, and more. Merle Haggard was one of the most important country music musicians who ever lived. His astonishing musical career stretched across the second half of the 20th Century and into the first two decades of the next, during which he released an extraordinary 63

albums, 38 that made it on to Billboard's Country Top Ten, 13 that went to #1, and 37 #1 hit singles. With his ample songbook, unique singing voice and brilliant phrasing that illuminated his uncompromising commitment to individual freedom, cut with the monkey of personal despair on his back and a chip the size of Monument Valley on his shoulder, Merle's music and his extraordinary charisma helped change the look, the sound, and the fury of American music. The Hag tells, without compromise, the extraordinary life of Merle Haggard, augmented by deep secondary research, sharp detail and ample anecdotal material that biographer Marc Eliot is known for, and enriched and deepened by over 100 new and far-ranging interviews. It explores the uniquely American life of an angry rebellious boy from the wrong side

of the tracks bound for a life of crime and a permanent home in a penitentiary, who found redemption through the music of "the common man." Merle Haggard's story is a great American saga of a man who lifted himself out of poverty, oppression, loss and wanderlust, to catapult himself into the pantheon of American artists admired around the world. Eliot has interviewed more than 100 people who knew Haggard, worked with him, were influenced by him, loved him or hated him. The book celebrates the accomplishments and explore the singer's infamous dark side: the self-created turmoil that expressed itself through drugs, women, booze, and betrayal. The Hag offers a richly anecdotal narrative that will elevate the life and work of Merle Haggard to where both properly belong, in the pantheon of American music and letters. The

Hag is the definitive account of this unique American original, and will speak to readers of country music and rock biographies alike.

A teen outcast must work together with new friends to keep her family and town safe from murderous Fae while also dealing with panic attacks, family issues, and a lesbian love triangle in C.M. McGuires's kick-butt paranormal YA debut, Ironspark. For the past nine years, ever since a bunch of those evil Tinkerbells abducted her mother, cursed her father, and forced her family into hiding, Bryn has devoted herself to learning everything she can about killing the Fae. Now it's time to put those lessons to use. Then the Court Fae finally show up, and Bryn realizes she can't handle this on her own. Thankfully, three friends offer to help: Gwen, a kindhearted

water witch; Dom, a new foster kid pulled into her world; and Jasika, a schoolmate with her own grudge against the Fae. But trust is hard-won, and what little Bryn has gained is put to the test when she uncovers a book of Fae magic that belonged to her mother. With the Fae threat mounting every day, Bryn must choose between faith in her friends and power from a magic that could threaten her very humanity.

Big Data Processing Made Simple

High Performance Spark

Trino: The Definitive Guide

Create scalable pipelines that ingest, curate, and aggregate complex data in a timely and secure way

Apache Spark in 24 Hours, Sams Teach Yourself

Ironspark

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 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

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 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 LearnDiscover the functional programming features of Scala Understand the complete architecture of Spark and its componentsIntegrate 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.

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 HBase

Kafka: The Definitive Guide

The Complete Guide to Large-Scale Analysis and Modeling

Frank Kane's Taming Big Data with Apache Spark and Python

Lightning-Fast Big Data Analysis

Unleashing Large Cluster Analytics in the Cloud

The past few years have seen a major change in computing systems, as growing data volumes and stalling processor speeds require more and more applications to scale out to clusters. Today, a myriad data sources, from the Internet to business operations to scientific instruments, produce large and valuable data streams. However, the processing capabilities of single machines have not kept up with the size of data. As a result, organizations increasingly need to scale out their computations over clusters. At the same time, the speed and sophistication required of data processing have grown. In addition to simple queries, complex algorithms like machine learning and graph analysis are becoming common. And in addition to batch

processing, streaming analysis of real-time data is required to let organizations take timely action. Future computing platforms will need to not only scale out traditional workloads, but support these new applications too. This book, a revised version of the 2014 ACM Dissertation Award winning dissertation, proposes an architecture for cluster computing systems that can tackle emerging data processing workloads at scale. Whereas early cluster computing systems, like MapReduce, handled batch processing, our architecture also enables streaming and interactive queries, while keeping MapReduce's scalability and fault tolerance. And whereas most deployed systems only support simple one-pass computations (e.g.,

SQL queries), ours also extends to the multi-pass algorithms required for complex analytics like machine learning. Finally, unlike the specialized systems proposed for some of these workloads, our architecture allows these computations to be combined, enabling rich new applications that intermix, for example, streaming and batch processing. We achieve these results through a simple extension to MapReduce that adds primitives for data sharing, called Resilient Distributed Datasets (RDDs). We show that this is enough to capture a wide range of workloads. We implement RDDs in the open source Spark system, which we evaluate using synthetic and real workloads. Spark matches or exceeds the performance of

specialized systems in many domains, while offering stronger fault tolerance properties and allowing these workloads to be combined. Finally, we examine the generality of RDDs from both a theoretical modeling perspective and a systems perspective. This version of the dissertation makes corrections throughout the text and adds a new section on the evolution of Apache Spark in industry since 2014. In addition, editing, formatting, and links for the references have been added.

Summary The Spark distributed data processing platform provides an easy-to-implement tool for ingesting, streaming, and processing data from any source. In Spark in Action, Second Edition, you'll learn to take advantage of

Spark's core features and incredible processing speed, with applications including real-time computation, delayed evaluation, and machine learning. Spark skills are a hot commodity in enterprises worldwide, and with Spark's powerful and flexible Java APIs, you can reap all the benefits without first learning Scala or Hadoop. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Analyzing enterprise data starts by reading, filtering, and merging files and streams from many sources. The Spark data processing engine handles this varied volume like a champ, delivering speeds 100 times faster than Hadoop systems. Thanks to SQL support, an

intuitive interface, and a straightforward multilanguage API, you can use Spark without learning a complex new ecosystem. About the book Spark in Action, Second Edition, teaches you to create end-to-end analytics applications. In this entirely new book, you'll learn from interesting Java-based examples, including a complete data pipeline for processing NASA satellite data. And you'll discover Java, Python, and Scala code samples hosted on GitHub that you can explore and adapt, plus appendixes that give you a cheat sheet for installing tools and understanding Spark-specific terms. What's inside Writing Spark applications in Java Spark application architecture Ingestion through files, databases, streaming,

and Elasticsearch Querying distributed datasets with Spark SQL About the reader This book does not assume previous experience with Spark, Scala, or Hadoop. About the author Jean-Georges Perrin is an experienced data and software architect. He is France's first IBM Champion and has been honored for 12 consecutive years. Table of Contents PART 1 - THE THEORY CRIPPLED BY AWESOME EXAMPLES 1 So, what is Spark, anyway? 2 Architecture and flow 3 The majestic role of the dataframe 4 Fundamentally lazy 5 Building a simple app for deployment 6 Deploying your simple app PART 2 - INGESTION 7 Ingestion from files 8 Ingestion from databases 9 Advanced ingestion: finding data sources and

building your own 10 Ingestion through structured streaming PART 3 - TRANSFORMING YOUR DATA 11 Working with SQL 12 Transforming your data 13 Transforming entire documents 14 Extending transformations with user-defined functions 15 Aggregating your data PART 4 - GOING FURTHER 16 Cache and checkpoint: Enhancing Spark's performances 17 Exporting data and building full data pipelines 18 Exploring deployment

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

Practical Data Science with Hadoop and Spark

The Hag

Implementing End-to-End Real-Time Data Pipelines: From Ingest to Machine Learning