Apache Spark 2 0 Ga Machine Learning Analytics Cloud

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

A "genotype" is essentially an organism's full hereditary information which is obtained from its parents. A "phenotype" is an organism's actual observed physical and behavioral properties. These may include traits such as morphology, size, height, eye color, metabolism, etc. One of the pressing challenges in computational and systems biology is genotype-to-phenotype prediction. This is challenging given the amount of data generated by modern Omics technologies. This "Big Data" is so large and complex that traditional data processing applications are not up to the task. Challenges arise in collection, analysis, mining, sharing, transfer, visualization, archiving, and integration of these data. In this Special Issue, there is a focus on the systems-level analysis of Omics data, recent developments in gene ontology annotation, and advances in biological pathways and network biology. The integration of Omics data with clinical and biomedical data using machine learning is explored. This Special Issue covers new methodologies in the context of gene–environment interactions, tissue-specific gene expression, and how external factors or host genetics impact the microbiome.

This book constitutes the refereed proceedings of the 6th IFIP TC 5 International Conference on Computational Intelligence and Its Applications, CIIA 2018, held in Oran, Algeria, in May 2018. The 56 full papers presented were carefully reviewed and selected from 202 submissions. They are organized in the following topical sections: data mining and information retrieval; evolutionary computation; machine learning; optimization; planning and scheduling; wireless communication and mobile computing; Internet of Things (IoT) and decision support systems; pattern recognition and image processing; and semantic web services.

The 4-volume set LNCS 11632 until LNCS 11635 constitutes the refereed proceedings of the 5th International Conference on Artificial Intelligence and Security, ICAIS 2019, which was held in New York, USA, in July 2019. The conference was formerly called "International Conference on Cloud Computing and Security" with the acronym ICCCS. The total of 230 full papers presented in this 4-volume proceedings was carefully reviewed and selected from 1529 submissions. The papers were organized in topical sections as follows: Part I: cloud computing; Part II: artificial intelligence; big data; and cloud computing and security; Part III: cloud computing and security; information hiding; IoT security; multimedia forensics; and encryption and cybersecurity; Part IV: encryption and cybersecurity.

Includes proceedings and reports of conferences of various financial organizations.

This open access book constitutes the refereed proceedings of the 15th International Conference on Semantic Systems, SEMANTiCS 2019, held in Karlsruhe, Germany, in September 2019. The 20 full papers and 8 short papers presented in this volume were carefully reviewed and selected from 88 submissions. They cover topics such as: web semantics and linked (open) data; machine learning and deep learning techniques; semantic information management and knowledge integration; terminology, thesaurus and ontology management; data mining and knowledge discovery; semantics in blockchain and distributed ledger technologies.

If you are ready to dive into the MapReduce framework for processing large datasets, this practical book takes you step by step through the algorithms and tools you need to build distributed MapReduce applications with Apache Hadoop or Apache Spark. Each chapter provides a recipe for solving a massive computational problem, such as building a recommendation system. You'll learn how to implement the appropriate MapReduce solution with code that you can use in your projects. Dr. Mahmoud Parsian covers basic design patterns, optimization techniques, and data mining and machine learning solutions for problems in bioinformatics, genomics, statistics, and social network analysis. This book also includes an overview of MapReduce, Hadoop, and Spark. Topics include: Market basket analysis for a large set of transactions Data mining algorithms (K-means, KNN, and Naive Bayes) Using huge genomic data to sequence DNA and RNA Naive Bayes theorem and Markov chains for data and market prediction Recommendation algorithms and pairwise document similarity Linear regression, Cox regression, and Pearson correlation Allelic frequency and mining DNA Social network analysis (recommendation systems, counting triangles, sentiment analysis)

This book constitutes the refereed proceedings of the 24th International Conference on Case-Based Reasoning Research and Development, ICCBR 2016, held in Atlanta, GA, USA, in October/November 2016. The 14 revised full papers presented were carefully reviewed and selected from 44 submissions. The papers cover a wide range of CBR topics that are of interest both to researchers and practitioners from foundations of Case-Based Reasoning; over CBR systems for specific tasks and related fields; up to CBR systems, applications and lessons learned in specific areas of expertise such as health; e-science; finance; energy, logistics, traffic; game/AI; cooking; diagnosis, technical support; as well as knowledge and experience management.

This book constitutes the refereed proceedings of the 11th International Conference on Simulated Evolution and Learning, SEAL 2017, held in Shenzhen, China, in November 2017. The 85 papers presented in this volume were carefully reviewed and selected from 145 submissions. They were organized in topical sections named: evolutionary optimisation; evolutionary multiobjective optimisation; evolutionary machine learning; theoretical developments; feature selection and dimensionality reduction; dynamic and uncertain environments; real-world applications; adaptive systems; and swarm intelligence.

This holistic book is an invaluable reference for addressing various practical challenges in architecting and engineering Intelligent IoT and eHealth solutions for industry practitioners, academic and

researchers, as well as for engineers involved in product development. The first part provides a comprehensive guide to fundamentals, applications, challenges, technical and economic benefits, and promises of the Internet of Things using examples of real-world applications. It also addresses all important aspects of designing and engineering cutting-edge IoT solutions using a cross-layer approach from device to fog, and cloud covering standards, protocols, design principles, reference architectures, as well as all the underlying technologies, pillars, and components such as embedded systems, network, cloud computing, data storage, data processing, big data analytics, machine learning, distributed ledger technologies, and security. In addition, it discusses the effects of Intelligent IoT, which are reflected in new business models and digital transformation. The second part provides an insightful guide to the design and deployment of IoT solutions for smart healthcare as one of the most important applications of IoT. Therefore, the second part targets smart healthcare-wearable sensors, body area sensors, advanced pervasive healthcare systems, and big data analytics that are aimed at providing connected health interventions to individuals for healthier lifestyles.

Learn the right cutting-edge skills and knowledge to leverage Spark Streaming to implement a wide array of real-time, streaming applications. This book walks you through end-to-end real-time application development using real-world applications, data, and code. Taking an application-first approach, each chapter introduces use cases from a specific industry and uses publicly available datasets from that domain to unravel the intricacies of production-grade design and implementation. The domains covered in Pro Spark Streaming include social media, the sharing economy, finance, online advertising, telecommunication, and IoT. In the last few years, Spark has become synonymous with big data processing. DStreams enhance the underlying Spark processing engine to support streaming analysis with a novel micro-batch processing model. Pro Spark Streaming by Zubair Nabi will enable you to become a specialist of latency sensitive applications by leveraging the key features of DStreams, micro-batch processing, and functional programming. To this end, the book includes ready-to-deploy examples and actual code. Pro Spark Streaming will act as the bible of Spark Streaming. What You'll Learn Discover Spark Streaming application development and best practices Work with the low-level details of discretized streams Optimize production-grade deployments of Spark Streaming via configuration recipes and instrumentation using Graphite, collectd, and Nagios Ingest data from disparate sources including MQTT, Flume, Kafka, Twitter, and a custom HTTP receiver Integrate and couple with HBase, Cassandra, and Redis Take advantage of design patterns for side-effects and maintaining state across the Spark Streaming micro-batch model Implement real-time and scalable ETL using data frames, SparkSQL, Hive, and SparkR Use streaming machine learning, predictive analytics, and recommendations Mesh batch processing with stream processing via the Lambda architecture Who This Book Is For Data scientists, big data experts, BI analysts, and data architects.

The two-volume set of LNCS 11778 and 11779 constitutes the refereed proceedings of the 18th International Semantic Web Conference, ISWC 2019, held in Auckland, New Zealand, in October 2019. The ISWC conference is the premier international forum for the Semantic Web / Linked Data Community. The total of 74 full papers included in this volume was selected from 283 submissions. The conference is organized in three tracks: for the Research Track 42 full papers were selected from 194 submissions; the Resource Track contains 21 full papers, selected from 64 submissions; and the In-Use Track features 11 full papers which were selected from 25 submissions to this track. The chapter "The SEPSES knowledge graph: An integrated resource for cybersecurity" is open access under a CC BY 4.0 license at link.springer.com.

Virtual, hands-on learning labs allow you to apply your technical skills in realistic environments. So Sybex has bundled AWS labs from XtremeLabs with our popular AWS Certified Data Analytics Study Guide to give you the same experience working in these labs as you prepare for the Certified Data Analytics Exam that you would face in a real-life application. These labs in addition to the book are a proven way to prepare for the certification and for work as an AWS Data Analyst. AWS Certified Data Analytics Study Guide: Specialty (DAS-C01) Exam is intended for individuals who perform in a data analytics-focused role. This UPDATED exam validates an examinee's comprehensive understanding of using AWS services to design, build, secure, and maintain analytics solutions that provide insight from data. It assesses an examinee's ability to define AWS data analytics services and understand how they integrate with each other; and explain how AWS data analytics services fit in the data lifecycle of collection, storage, processing, and visualization. The book focuses on the following domains: • Collection • Storage and Data Management • Processing • Analysis and Visualization • Data Security This is your opportunity to take the next step in your career by expanding and validating your skills on the AWS cloud. AWS is the frontrunner in cloud computing products and services, and the AWS Certified Data Analytics Study Guide: Specialty exam will get you fully prepared through expert content, and real-world knowledge, key exam essentials, chapter review questions, and much more. Written by an AWS subject-matter expert, this study guide covers exam concepts, and provides key review on exam topics. Readers will also have access to Sybex's superior online interactive learning environment and test bank, including chapter tests, practice exams, a glossary of key terms, and electronic flashcards. And included with this version of the book, XtremeLabs virtual labs that run from your browser. The registration code is included with the book and gives you 6 months of unlimited access to XtremeLabs AWS Certified Data Analytics Labs with 3 unique lab modules based on the book.

If you have a working knowledge of Hadoop 1.x but want to start afresh with YARN, this book is ideal for you. You will be able to install and administer a YARN cluster and also discover the configuration settings to fine-tune your cluster both in terms of performance and scalability. This book will help you develop, deploy, and run multiple applications/frameworks on the same shared YARN cluster.

Advance your skills in efficient data analysis and data processing using the powerful tools of Scala, Spark, and Hadoop About This Book This is a primer on functional-programming-style techniques to help you efficiently process and analyze all of your data Get acquainted with the best and newest tools available such as Scala, Spark, Parquet and MLlib for machine learning Learn the best practices to incorporate new Big Data machine learning in your data-driven enterprise to gain future scalability and maintainability Who This Book Is For Mastering Scala Machine Learning is intended for enthusiasts who want to plunge into the new pool of emerging techniques for machine learning. Some familiarity with standard statistical techniques is required. What You Will Learn Sharpen your functional programming skills in Scala using REPL Apply standard and advanced machine learning techniques using Scala Get acquainted with Big Data technologies and grasp why we need a functional approach to Big Data Discover new data structures, algorithms, approaches, and habits that will allow you to work effectively with large amounts of data Understand the principles of supervised and unsupervised learning in machine learning Work with unstructured data and serialize it using Kryo, Protobuf, Avro, and AvroParquet Construct reliable and robust data pipelines and manage data in a data-driven enterprise Implement scalable model monitoring and alerts with Scala In Detail Since the advent of object-oriented programming, new technologies related to Big Data are constantly popping up on the market. One such technology is Scala, which is considered to be a successor to Java in the area of Big Data by many, like Java was to C/C++ in the area of distributed programing. This book aims to take your knowledge to next level and help you impart that knowledge to build advanced applications such as social media mining, intelligent news portals, and more. After a quick refresher on functional programming concepts using REPL, you will see some practical examples of setting up the development environment and tinkering with data. We will then explore working with Spark and MLlib using k-means and decision trees. Most of the data that we produce today is unstructured and raw, and you will learn to tackle this type of data with advanced topics such as regression, classification, integration, and working with graph algorithms. Finally, you will discover at how to use Scala to perform complex concept analysis, to monitor model performance, and to build a model repository. By the end of this book, you will have gained expertise in performing Scala machine learning and will be able to build complex machine learning projects using Scala. Style and approach This hands-on guide dives straight into implementing Scala for machine

learning without delving much into mathematical proofs or validations. There are ample code examples and tricks that will help you sail through using the standard techniques and libraries. This book provides practical examples from the field on how to correctly tackle data analysis problems, particularly for modern Big Data datasets.

Intelligent Computing Theories and Application15th International Conference, ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part ISpringer

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

This book constitutes the proceedings of the 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, held in Porto, Portugal, in September 2018. The 51 full papers, 17 short papers, and 13 poster and tutorial papers presented in this volume were carefully reviewed and selected from 81 submissions. The general theme of TPDL 2018 was Digital Libraries for Open Knowledge. The papers present a wide range of the following topics: Metadata, Entity Disambiguation, Data Management, Scholarly Communication, Digital Humanities, User Interaction, Resources, Information Extraction, Information Retrieval, Recommendation.

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 hot only help you get up and running with Spark, but will also take you farther down the road to becoming a data scientist.

This two-volume set LNCS 11101 and 11102 constitutes the refereed proceedings of the 15th International Conference on Parallel Problem Solving from Nature, PPSN 2018, held in Coimbra, Portugal, in September 2018. The 79 revised full papers were carefully reviewed and selected from 205 submissions. The papers cover a wide range of topics in natural computing including evolutionary computation, artificial neural networks, artificial life, swarm intelligence, artificial immune systems, self-organizing systems, emergent behavior, molecular computing, evolutionary robotics, evolvable hardware, parallel implementations and applications to real-world problems. The papers are organized in the following topical sections: numerical optimization; combinatorial optimization; genetic programming; multi-objective optimization; parallel and distributed frameworks; runtime analysis and approximation results; fitness landscape modeling and analysis; algorithm configuration, selection, and benchmarking; machine learning and evolutionary algorithms; and applications. Also included are the descriptions of 23 tutorials and 6 workshops which took place in the framework of PPSN XV.

Design, implement, and deliver successful streaming applications, machine learning pipelines and graph applications using Spark SQL API About This Book Learn about the design and implementation of streaming applications, machine learning pipelines, deep learning, and large-scale graph processing applications using Spark SQL APIs and Scala. Learn data exploration, data munging, and how to process structured and semi-structured data using real-world datasets and gain hands-on exposure to the issues and challenges of working with noisy and "dirty" real-world data. Understand design considerations for scalability and performance in web-scale Spark application architectures. Who This Book Is For If you are a developer, engineer, or an architect and want to learn how to use Apache Spark in a web-scale project, then this is the book for you. It is assumed that you have prior knowledge of SQL querying. A basic programming knowledge with Scala, Java, R, or Python is all you need to get started with this book. What You Will Learn Familiarize yourself with Spark SQL programming, including working with DataFrame/Dataset API and SQL Perform a series of hands-on exercises with different types of data sources, including CSV, JSON, Avro, MySQL, and MongoDB Perform data quality checks, data visualization, and basic statistical analysis tasks Perform data munging tasks on publically available datasets Learn how to use Spark SQL and Apache Kafka to build streaming applications Learn key performance-tuning tips and tricks in Spark SQL applications Learn key architectural components and patterns in large-scale Spark SQL applications In Detail In the past year, Apache Spark has been increasingly adopted for the development of distributed applications. Spark SQL APIs provide an optimized interface that helps developers build such applications quickly and easily. However, designing web-scale production applications using Spark SQL APIs can be a complex task. Hence, understanding the design and implementation best practices before you start your project will help you avoid these problems. This book gives an insight into the engineering practices used to design and build real-world, Spark-based applications. The book's hands-on examples will give you the required confidence to work on any future projects you encounter in Spark SQL. It starts by familiarizing you with data exploration and data

munging tasks using Spark SQL and Scala. Extensive code examples will help you understand the methods used to implement typical use-cases for various types of applications. You will get a walkthrough of the key concepts and terms that are common to streaming, machine learning, and graph applications. You will also learn key performance-tuning details including Cost Based Optimization (Spark 2.2) in Spark SQL applications. Finally, you will move on to learning how such systems are architected and deployed for a successful delivery of your project. Style and approach This book is a hands-on guide to designing, building, and deploying Spark SQL-centric production applications at scale.

This book constitutes the refereed proceedings of the 17th International Semantic Web Conference, ESWC 2020, held in Heraklion, Crete, Greece.* The 39 revised full papers presented were carefully reviewed and selected from 166 submissions. The papers were submitted to three tracks: the research track, the resource track and the in-use track. These tracks showcase research and development activities, services and applications, and innovative research outcomes making their way into industry. The research track caters for both long standing and emerging research topics in the form of the following subtracks: ontologies and reasoning; natural language processing and information retrieval; semantic data management and data infrastructures; social and human aspects of the Semantic Web; machine learning; distribution and decentralization; science of science; security, privacy, licensing and trust; knowledge graphs; and integration, services and APIs. *The conference was held virtually due to the COVID-19 pandemic. Chapter ‘Piveau: A Large-scale Oopen Data Management Platform based on Semantic Web Technologies’ is available open access under a Creative Commons Attribution 4.0 International License via link.springer.com.

This two-volume set of LNCS 11643 and LNCS 11644 constitutes - in conjunction with the volume LNAI 11645 - the refereed proceedings of the 15th International Conference on Intelligent Computing, ICIC 2019, held in Nanchang, China, in August 2019. The 217 full papers of the three proceedings volumes were carefully reviewed and selected from 609 submissions. The ICIC theme unifies the picture of contemporary intelligent computing techniques as an integral concept that highlights the trends in advanced computational intelligence and bridges theoretical research with applications. The theme for this conference is “Advanced Intelligent Computing Methodologies and Applications.” Papers related to this theme are especially solicited, including theories, methodologies, and applications in science and technology.

WORD 2007 IN SIMPLE STEPS is a book that helps you to learn WORD 2007, the latest offering from Microsoft. Being precise and complete, it offers the reader a cutting edge in the field of Microsoft Office. With an easy to understand style, lots of examples to support the concepts, and use of practical approach in presentation are some of the features that make the book not only unique but also provides a sort of limited-edition look to the book.

.NET Black Book is the one-time reference and solid introduction, written from the programmer s point of view, containing hundreds of examples covering every aspect of VS 2005 programming. It will help you master the entire spectrum of VB 2005 from Visual basic language reference to creating Windows Applications to control docking, from basic database handling to Windows Services, from Windows Mobile Applications to directory services and My Object and much more. In C# 2005 from C# language reference to OOPS to delegates and events and error handling in .NET Framework from graphics and file Handling to Remoting, from collection and generics to security and cryptography in .NET Framework and much more. In ASP.NET 2.0 from features of ASP.NET 2.0 to standard and HTML controls from navigation controls to Login and Web Parts controls, from data driven web applications to master pages and themes, from Caching to web services and AJAX and much more. This unique book is designed to contain more VS 2005 coverage than any other no doubt every aspect of the book is worth the price of the entire book.