Spark Architecture Distributed Systems Architecture

Right here, we have countless books **Spark Architecture Distributed Systems Architecture** and collections to check out. We additionally give variant types and plus type of the books to browse. The agreeable book, fiction, history, novel, scientific research, as skillfully as various extra sorts of books are readily approachable here.

As this Spark Architecture Distributed Systems Architecture , it ends happening monster one of the favored books Spark Architecture Distributed Systems Architecture collections that we have. This is why you remain in the best website to look the unbelievable books to have.

Information Systems Design and Intelligent Applications - Vikrant Bhateja 2018-03-01
The book is a collection of high-quality peer-reviewed research papers presented at International Conference on Information System Design and Intelligent Applications (INDIA 2017) held at Duy Tan University, Da Nang, Vietnam during 15-17 June 2017. The book covers a wide range of topics of computer science and information technology discipline ranging from image processing, database application, data mining, grid and cloud computing, bioinformatics and many others. The various intelligent tools like swarm intelligence, artificial intelligence, evolutionary algorithms, bio-inspired algorithms have been well applied in different domains for solving various challenging problems.
Computer Vision and Machine Learning in Agriculture, Volume 2 - Mohammad Shorif Uddin 2022-03-13
This book is as an extension of previous book “Computer Vision and Machine Learning in Agriculture” for academicians, researchers, and professionals interested in solving the problems of agricultural plants and products for boosting production by rendering the advanced machine learning including deep learning tools and techniques to computer vision algorithms. The book contains 15 chapters. The first three chapters are devoted to crops harvesting, weed, and multi-class crops detection with the help of robots and UAVs through machine learning and deep learning algorithms for smart agriculture. Next, two chapters describe agricultural data retrievals and data collections. Chapters 6, 7, 8 and 9 focuses on yield estimation, crop maturity detection, agri-food product quality assessment, and medicinal plant recognition, respectively. The remaining six chapters concentrates on optimized disease recognition through computer vision-based machine and deep learning strategies.
Proceeding of the Second International Conference on Microelectronics, Computing & Communication Systems (MCCS 2017) - Vijay Nath 2018-07-30
The volume presents high quality papers presented at the Second International Conference on Microelectronics, Computing & Communication Systems (MCCS 2017). The book discusses recent trends in technology and advancement in MEMS and nanoelectronics, wireless communications, optical communication, instrumentation, signal processing, image processing, bioengineering, green energy, hybrid vehicles, environmental science, weather forecasting, cloud computing, renewable energy,

RFID, CMOS sensors, actuators, transducers, telemetry systems, embedded systems, and sensor network applications. It includes original papers based on original theoretical, practical, experimental, simulations, development, application, measurement, and testing. The applications and solutions discussed in the book will serve as a good reference material for future works.

Mastering Spark for Data Science - Andrew Morgan 2017-03-29

Master the techniques and sophisticated analytics used to construct Spark-based solutions that scale to deliver production-grade data science products About This Book Develop and apply advanced analytical techniques with Spark Learn how to tell a compelling story with data science using Spark's ecosystem Explore data at scale and work with cutting edge data science methods Who This Book Is For This book is for those who have beginner-level familiarity with the Spark architecture and data science applications, especially those who are looking for a challenge and want to learn cutting edge techniques. This book assumes working knowledge of data science, common machine learning methods, and popular data science tools, and assumes you have previously run proof of concept studies and built prototypes. What You Will Learn Learn the design patterns that integrate Spark into industrialized data science pipelines See how commercial data scientists design scalable code and reusable code for data science services Explore cutting edge data science methods so that you can study trends and causality Discover advanced programming techniques using RDD and the DataFrame and Dataset APIs Find out how Spark can be used as a universal ingestion engine tool and as a web scraper Practice the implementation of advanced topics in graph processing, such as community detection and contact chaining Get to know the best practices when performing Extended Exploratory Data Analysis, commonly used in commercial data science teams Study advanced Spark concepts, solution design patterns, and integration architectures Demonstrate powerful data science pipelines In Detail Data science seeks to transform the world using data, and this is typically achieved through disrupting and changing real processes in real industries. In order to operate at this level you need to build data science solutions of substance –solutions that solve real problems. Spark has emerged as the big data platform of choice for data scientists due to its speed, scalability, and easy-to-use APIs. This book deep dives into using Spark to deliver production-grade data science solutions. This process is demonstrated by exploring the construction of a sophisticated global news analysis service that uses Spark to generate continuous geopolitical and current affairs insights.You will learn all about the core Spark APIs and take a comprehensive tour of advanced libraries, including Spark SQL, Spark Streaming, MLlib, and more. You will be introduced to advanced techniques and methods that will help you to construct commercial-grade data products. Focusing on a sequence of tutorials that deliver a working news intelligence service, you will learn about advanced Spark architectures, how to work with geographic data in Spark, and how to tune Spark algorithms so they scale linearly. Style and approach This is an advanced guide for those with beginner-level familiarity with the Spark architecture and working with Data Science applications. Mastering Spark for Data Science is a practical tutorial that uses core Spark APIs and takes a deep dive into advanced libraries including: Spark SQL, visual streaming, and MLlib. This book expands on titles like: Machine Learning with Spark and Learning Spark. It is the next learning curve for those comfortable with Spark and looking to improve their skills.

Proceedings of the Fifth Euro-China Conference on Intelligent Data Analysis and Applications - Pavel Krömer 2018-12-24

This volume of Advances in Intelligent Systems and Computing highlights papers presented at the Fifth Euro-China Conference on Intelligent Data Analysis and Applications (ECC2018), held in Xi’an,

China from October 12 to 14 2018. The conference was co-sponsored by Springer, Xi'an University of Posts and Telecommunications, VSB Technical University of Ostrava (Czech Republic), Fujian University of Technology, Fujian Provincial Key Laboratory of Digital Equipment, Fujian Provincial Key Lab of Big Data Mining and Applications, and Shandong University of Science and Technology in China. The conference was intended as an international forum for researchers and professionals engaged in all areas of computational intelligence, intelligent control, intelligent data analysis, pattern recognition, intelligent information processing, and applications.

Scaling Machine Learning with Spark - Adi Polak 2023-03-07

Learn how to build end-to-end scalable machine learning solutions with Apache Spark. With this practical guide, author Adi Polak introduces data and ML practitioners to creative solutions that supersede today's traditional methods. You'll learn a more holistic approach that takes you beyond specific requirements and organizational goals--allowing data and ML practitioners to collaborate and understand each other better. Scaling Machine Learning with Spark examines several technologies for building end-to-end distributed ML workflows based on the Apache Spark ecosystem with Spark MLlib, MLflow, TensorFlow, and PyTorch. If you're a data scientist who works with machine learning, this book shows you when and why to use each technology. You will: Explore machine learning, including distributed computing concepts and terminology Manage the ML lifecycle with MLflow Ingest data and perform basic preprocessing with Spark Explore feature engineering, and use Spark to extract features Train a model with MLlib and build a pipeline to reproduce it Build a data system to combine the power of Spark with deep learning Get a step-by-step example of working with distributed TensorFlow Use PyTorch to scale machine learning and its internal architecture

Evolutionary Decision Trees in Large-Scale Data Mining - Marek Kretowski 2019-06-05

This book presents a unified framework, based on specialized evolutionary algorithms, for the global induction of various types of classification and regression trees from data. The resulting univariate or oblique trees are significantly smaller than those produced by standard top-down methods, an aspect that is critical for the interpretation of mined patterns by domain analysts. The approach presented here is extremely flexible and can easily be adapted to specific data mining applications, e.g. cost-sensitive model trees for financial data or multi-test trees for gene expression data. The global induction can be efficiently applied to large-scale data without the need for extraordinary resources. With a simple GPU-based acceleration, datasets composed of millions of instances can be mined in minutes. In the event that the size of the datasets makes the fastest memory computing impossible, the Spark-based implementation on computer clusters, which offers impressive fault tolerance and scalability potential, can be applied.

Distributed Multi-label Learning on Apache Spark - Jorge Gonzalez Lopez 2019

This thesis proposes a series of multi-label learning algorithms for classification and feature selection implemented on the Apache Spark distributed computing model. Five approaches for determining the optimal architecture to speed up multi-label learning methods are presented. These approaches range from local parallelization using threads to distributed computing using independent or shared memory spaces. It is shown that the optimal approach performs hundreds of times faster than the baseline method. Three distributed multi-label k nearest neighbors methods built on top of the Spark architecture are proposed: an exact iterative method that computes pair-wise distances, an approximate tree-based method that indexes the instances across multiple nodes, and an approximate local sensitive hashing method that builds multiple hash tables to index the data. The results indicated that the predictions of the tree-based method are

on par with those of an exact method while reducing the execution times in all the scenarios. The aforementioned method is then used to evaluate the quality of a selected feature subset. The optimal adaptation for a multi-label feature selection criterion is discussed and two distributed feature selection methods for multi-label problems are proposed: a method that selects the feature subset that maximizes the Euclidean norm of individual information measures, and a method that selects the subset of features maximizing the geometric mean. The results indicate that each method excels in different scenarios depending on type of features and the number of labels. Rigorous experimental studies and statistical analyses over many multi-label metrics and datasets confirm that the proposals achieve better performances and provide better scalability to bigger data than the methods compared in the state of the art.

Designing Distributed Systems - Brendan Burns 2018-02-20

Without established design patterns to guide them, developers have had to build distributed systems from scratch, and most of these systems are very unique indeed. Today, the increasing use of containers has paved the way for core distributed system patterns and reusable containerized components. This practical guide presents a collection of repeatable, generic patterns to help make the development of reliable distributed systems far more approachable and efficient. Author Brendan Burns—Director of Engineering at Microsoft Azure—demonstrates how you can adapt existing software design patterns for designing and building reliable distributed applications. Systems engineers and application developers will learn how these long-established patterns provide a common language and framework for dramatically increasing the quality of your system. Understand how patterns and reusable components enable the rapid development of reliable distributed systems Use the side-car, adapter, and ambassador patterns to split your application into a group of containers on a single machine Explore loosely coupled multi-node distributed patterns for replication, scaling, and communication between the components Learn distributed system patterns for large-scale batch data processing covering work-queues, event-based processing, and coordinated workflows

Agile Business Architecture for Digital Transformation - Dr Mehmet Yildiz 2021-05-01

We are in a frenetic and a convoluted digital age. Every organisation strives to transform its business to stay competitive in this exponentially growing digital world. Digital transformation became pervasive and ubiquitous in all business ventures. This new norm of constant transformation requires architecting our business and underlying technology stacks rapidly. Establishing agile business architecture frameworks are fundamental requirements to achieve successful digital transformation outcomes. In this book, I attempt to share my knowledge and experience using a rigorous yet agile architectural method. My aim is to add accelerated value to the broader business architecture and digital transformation communities by focusing on the practical aspect with minimal emphasis on the theoretical aspect. The content in this book is based on my architectural thought leadership experience gained in multiple large business and enterprise architecture initiatives, focusing on business capabilities, digital transformation initiatives, and enterprise modernisation engagements, reflecting hard lessons learned in these applied settings. In this book I attempt to redefine the role of business architects as primary leaders for digital transformation programs. The content reflects my experience and observations from the field. As a caveat, this book is not based on theories in the traditional business architecture textbooks which may conflict with my experience. My beta readers found this as a unique guide reflecting reality from the field. Hope it adds new insights for your role in the business digital transformation initiatives.

Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing - Management

Association, Information Resources 2021-01-25
Distributed systems intertwine with our everyday lives. The benefits and current shortcomings of the underpinning technologies are experienced by a wide range of people and their smart devices. With the rise of large-scale IoT and similar distributed systems, cloud bursting technologies, and partial outsourcing solutions, private entities are encouraged to increase their efficiency and offer unparalleled availability and reliability to their users. The Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing is a vital reference source that provides valuable insight into current and emergent research occurring within the field of distributed computing. It also presents architectures and service frameworks to achieve highly integrated distributed systems and solutions to integration and efficient management challenges faced by current and future distributed systems. Highlighting a range of topics such as data sharing, wireless sensor networks, and scalability, this multi-volume book is ideally designed for system administrators, integrators, designers, developers, researchers, academicians, and students.

Integrating the Internet of Things Into Software Engineering Practices - Mala, D. Jeya 2019-01-25
To provide the necessary security and quality assurance activities into Internet of Things (IoT)-based software development, innovative engineering practices are vital. They must be given an even higher level of importance than most other events in the field. Integrating the Internet of Things Into Software Engineering Practices provides research on the integration of IoT into the software development life cycle (SDLC) in terms of requirements management, analysis, design, coding, and testing, and provides security and quality assurance activities to IoT-based software development. The content within this publication covers agile software, language specification, and collaborative software and is designed for analysts, security experts, IoT software programmers, computer and software engineers, students, professionals, and researchers.

Data Algorithms - Mahmoud Parsian 2015-07-13
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)

Artificial Intelligence and Soft Computing - Leszek Rutkowski 2016-05-30
The two-volume set LNAI 9692 and LNAI 9693 constitutes the refereed proceedings of the 15th International Conference on Artificial Intelligence and Soft Computing, ICAISC 2016, held in Zakopane, Poland in June 2016. The 134 revised full papers presented were carefully reviewed and selected from 343 submissions. The papers included in the first volume are organized in the following topical sections: neural networks and their applications; fuzzy systems and their applications; evolutionary

algorithms and their applications; agent systems, robotics and control; and pattern classification. The second volume is divided in the following parts: bioinformatics, biometrics and medical applications; data mining; artificial intelligence in modeling and simulation; visual information coding meets machine learning; and various problems of artificial intelligence.

Deep Learning - Stephane S. Tuffery 2023-01-10

A concise and practical exploration of key topics and applications in data science In Deep Learning, from Big Data to Artificial Intelligence, expert researcher Dr. Stéphane Tufféry delivers an insightful discussion of the applications of deep learning and big data that focuses on practical instructions on various software tools and deep learning methods relying on three major libraries: MXNet, PyTorch, and Keras-TensorFlow. In the book, numerous, up-to-date examples are combined with key topics relevant to modern data scientists, including processing optimization, neural network applications, natural language processing, and image recognition. This is a thoroughly revised and updated edition of a book originally released in French, with new examples and methods included throughout. Classroom-tested and intuitively organized, Deep Learning, from Big Data to Artificial Intelligence offers complimentary access to a companion website that provides R and Python source code for the examples offered in the book. Readers will also find: A thorough introduction to practical deep learning techniques with explanations and examples for various programming libraries Comprehensive explorations of a variety of applications for deep learning, including image recognition and natural language processing Discussions of the theory of deep learning, neural networks, and artificial intelligence linked to concrete techniques and strategies commonly used to solve real-world problems Perfect for graduate students studying data science, big data, deep learning, and artificial intelligence, Deep Learning, from Big Data to Artificial Intelligence will also earn a place in the libraries of data science researchers and practicing data scientists.

The Social Sciences Empowered - Ford Lumban Gaol 2020-04-14

The Social Sciences Empowered contains papers presented at the 7th International Congress on Interdisciplinary Behavior and Social Science 2018 (ICIBSoS 2018), held 21-22 July 2018, Bangkok, Thailand, 22-23 September 2018, Bali, Indonesia, 6-7 October 2018, Kuta, Bali, Indonesia, and 24-25 November 2018, Yogyakarta, Indonesia. ICIBSoS 2018 provided the economic and social analysis necessary for addressing issues in Humanities disciplines such as Education, Sociology, Anthropology, Politics, History, Philosophy, Psychology as well as food security. Contributions to these proceedings give necessary insight into the cultural and human dimension of such diverse research areas as transport, climate change, energy and agriculture. ICIBSoS 2018 also analyses the cultural, behavioural, psychological, social and institutional drivers that transform people’s behaviour and the global environment. ICIBSoS 2018 proposes new ideas, strategies and governance structures for overcoming the crisis from a global perspective, innovating the public sector and business models, promoting social innovation and fostering creativity in the development of services and product design.

Hybrid Intelligent Systems - Ajith Abraham 2020-08-12

This book highlights the recent research on hybrid intelligent systems and their various practical applications. It presents 34 selected papers from the 18th International Conference on Hybrid Intelligent Systems (HIS 2019) and 9 papers from the 15th International Conference on Information Assurance and Security (IAS 2019), which was held at VIT Bhopal University, India, from December 10 to 12, 2019. A premier conference in the field of artificial intelligence, HIS - IAS 2019 brought together researchers, engineers and practitioners whose work involves intelligent systems, network security and their applications in industry. Including contributions by authors from 20 countries, the book offers a valuable reference guide for all researchers, students and

practitioners in the fields of Computer Science and Engineering.

Big Data Processing with Apache Spark - Srini Penchikala 2018-03-13

Apache Spark is a popular open-source big-data processing framework thatÕs built around speed, ease of use, and unified distributed computing architecture. Not only it supports developing applications in different languages like Java, Scala, Python, and R, itÕs also hundred times faster in memory and ten times faster even when running on disk compared to traditional data processing frameworks. Whether you are currently working on a big data project or interested in learning more about topics like machine learning, streaming data processing, and graph data analytics, this book is for you. You can learn about Apache Spark and develop Spark programs for various use cases in big data analytics using the code examples provided. This book covers all the libraries in Spark ecosystem: Spark Core, Spark SQL, Spark Streaming, Spark ML, and Spark GraphX.

Algorithms and Architectures for Parallel Processing - Sheng Wen 2020-01-21

The two-volume set LNCS 11944-11945 constitutes the proceedings of the 19th International Conference on Algorithms and Architectures for Parallel Processing, ICA3PP 2019, held in Melbourne, Australia, in December 2019. The 73 full and 29 short papers presented were carefully reviewed and selected from 251 submissions. The papers are organized in topical sections on: Parallel and Distributed Architectures, Software Systems and Programming Models, Distributed and Parallel and Network-based Computing, Big Data and its Applications, Distributed and Parallel Algorithms, Applications of Distributed and Parallel Computing, Service Dependability and Security, IoT and CPS Computing, Performance Modelling and Evaluation.

Data Intensive Computing Applications for Big Data - M. Mittal 2018-01-31

The book ‘Data Intensive Computing Applications for Big Data’ discusses the technical concepts of big data, data intensive computing through machine learning, soft computing and parallel computing paradigms. It brings together researchers to report their latest results or progress in the development of the above mentioned areas. Since there are few books on this specific subject, the editors aim to provide a common platform for researchers working in this area to exhibit their novel findings. The book is intended as a reference work for advanced undergraduates and graduate students, as well as multidisciplinary, interdisciplinary and transdisciplinary research workers and scientists on the subjects of big data and cloud/parallel and distributed computing, and explains didactically many of the core concepts of these approaches for practical applications. It is organized into 24 chapters providing a comprehensive overview of big data analysis using parallel computing and addresses the complete data science workflow in the cloud, as well as dealing with privacy issues and the challenges faced in a data-intensive cloud computing environment. The book explores both fundamental and high-level concepts, and will serve as a manual for those in the industry, while also helping beginners to understand the basic and advanced aspects of big data and cloud computing.

Transactions on Large-Scale Data- and Knowledge-Centered Systems XLVII - Abdelkader Hameurlain 2021-01-16

The LNCS journal Transactions on Large-Scale Data- and Knowledge-Centered Systems focuses on data management, knowledge discovery, and knowledge processing, which are core and hot topics in computer science. Since the 1990s, the Internet has become the main driving force behind application development in all domains. An increase in the demand for resource sharing across different sites connected through networks has led to an evolution of data- and knowledge-management systems from centralized systems to decentralized systems enabling large-scale distributed applications providing

high scalability. This, the 47th issue of Transactions on Large-Scale Data- and Knowledge-Centered Systems, constitutes a special issue focusing on Digital Ecosystems and Social Networks. The 9 revised selected papers cover topics that include Social Big Data, Data Analysis, Cloud-Based Feedback, Experience Ecosystems, Pervasive Environments, and Smart Systems.

The Role of Technology in Education - Fahriye Altınay 2020-03-11

This book has three sections on the role of technology in education. The first section covers the merits of online learning and environment. The second section of the book gives insight on new technologies in learning and teaching. The third section of the book underlines the importance of new tendencies for the technology in education. I have a firm belief that readers can find great insights on the role of technology in education from different reflections and research.

Data Algorithms with Spark - Mahmoud Parsian 2022-04-08

Apache Spark's speed, ease of use, sophisticated analytics, and multilanguage support makes practical knowledge of this cluster-computing framework a required skill for data engineers and data scientists. With this hands-on guide, anyone looking for an introduction to Spark will learn practical algorithms and examples using PySpark. In each chapter, author Mahmoud Parsian shows you how to solve a data problem with a set of Spark transformations and algorithms. You'll learn how to tackle problems involving ETL, design patterns, machine learning algorithms, data partitioning, and genomics analysis. Each detailed recipe includes PySpark algorithms using the PySpark driver and shell script. With this book, you will: Learn how to select Spark transformations for optimized solutions Explore powerful transformations and reductions including reduceByKey(), combineByKey(), and mapPartitions() Understand data partitioning for optimized queries Build and apply a model using PySpark design patterns Apply motif-finding algorithms to graph data Analyze graph data by using the GraphFrames API Apply PySpark algorithms to clinical and genomics data Learn how to use and apply feature engineering in ML algorithms Understand and use practical and pragmatic data design patterns

Distributed Applications and Interoperable Systems - Anne Remke 2020-06-08

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

Creating Autonomous Vehicle Systems - Shaoshan Liu 2020-09-11

This is one of the first technical overviews of autonomous vehicles written for a general computing and engineering audience. Students will find a comprehensive overview of the entire autonomous technology stack and practitioners will find many practical techniques. Throughout the book, the authors share their practical experiences designing autonomous vehicle systems. These systems are complex, consisting of three major subsystems: (1) algorithms for localization, perception, and planning and control; (2) client systems, such as the robotics operating system and hardware platform; and (3) the cloud platform, which includes data storage, simulation, high-definition (HD) mapping, and deep learning model training. The algorithm subsystem extracts meaningful information from sensor raw data to understand its environment and make decisions as to its future actions. The

client subsystem integrates these algorithms to meet real-time and reliability requirements. The cloud platform provides offline computing and storage capabilities for autonomous vehicles. Using the cloud platform, new algorithms can be tested so as to update the HD map in addition to training better recognition, tracking, and decision models. Since the first edition of this book was released, many universities have adopted it in their autonomous driving classes, and the authors received many helpful comments and feedback from readers. Based on this, the second edition was improved by extending and rewriting multiple chapters and adding two commercial test case studies. In addition, a new section entitled “Teaching and Learning from this Book” was added to help instructors better utilize this book in their classes. The second edition captures the latest advances in autonomous driving and that it also presents usable real-world case studies to help readers better understand how to utilize their lessons in commercial autonomous driving projects.

Advances in Knowledge Discovery and Data Mining - Qiang Yang 2019-04-03

The three-volume set LNAI 11439, 11440, and 11441 constitutes the thoroughly refereed proceedings of the 23rd Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2019, held in Macau, China, in April 2019. The 137 full papers presented were carefully reviewed and selected from 542 submissions. The papers present new ideas, original research results, and practical development experiences from all KDD related areas, including data mining, data warehousing, machine learning, artificial intelligence, databases, statistics, knowledge engineering, visualization, decision-making systems, and the emerging applications. They are organized in the following topical sections: classification and supervised learning; text and opinion mining; spatio-temporal and stream data mining; factor and tensor analysis; healthcare, bioinformatics and related topics; clustering and anomaly detection; deep learning models and applications; sequential pattern mining; weakly supervised learning; recommender system; social network and graph mining; data pre-processing and featureselection; representation learning and embedding; mining unstructured and semi-structured data; behavioral data mining; visual data mining; and knowledge graph and interpretable data mining.

Natural Language Processing with Spark NLP - Alex Thomas 2020-06-25

If you want to build an enterprise-quality application that uses natural language text but aren’t sure where to begin or what tools to use, this practical guide will help get you started. Alex Thomas, principal data scientist at Wisecube, shows software engineers and data scientists how to build scalable natural language processing (NLP) applications using deep learning and the Apache Spark NLP library. Through concrete examples, practical and theoretical explanations, and hands-on exercises for using NLP on the Spark processing framework, this book teaches you everything from basic linguistics and writing systems to sentiment analysis and search engines. You’ll also explore special concerns for developing text-based applications, such as performance. In four sections, you’ll learn NLP basics and building blocks before diving into application and system building: Basics: Understand the fundamentals of natural language processing, NLP on Apache Stark, and deep learning Building blocks: Learn techniques for building NLP applications—including tokenization, sentence segmentation, and named-entity recognition—and discover how and why they work Applications: Explore the design, development, and experimentation process for building your own NLP applications Building NLP systems: Consider options for productionizing and deploying NLP models, including which human languages to support

An Architecture for Fast and General Data Processing on Large Clusters - Matei Zaharia 2016-05-01

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.

Spark in Action - Jean-Georges Perrin 2020-05-12

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

Hands-On Deep Learning with Apache Spark - Guglielmo Iozzia 2019-01-31

Speed up the design and implementation of deep learning solutions using Apache Spark Key FeaturesExplore the world of distributed deep learning with Apache SparkTrain neural networks with deep learning libraries such as BigDL and TensorFlowDevelop Spark deep learning applications to intelligently handle large and complex datasetsBook Description Deep learning is a subset of machine learning where datasets with several layers of complexity can be processed. Hands-On Deep Learning with Apache Spark addresses the sheer complexity of technical and analytical parts and the speed at which deep learning solutions can be implemented on Apache Spark. The book starts with the fundamentals of Apache Spark and deep learning. You will set up Spark for deep learning, learn principles of distributed modeling, and understand different types of neural nets. You will then implement deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) on Spark. As you progress through the book, you will gain hands-on experience of what it takes to understand the complex datasets you are dealing with. During the course of this book, you will use popular deep learning frameworks, such as TensorFlow, Deeplearning4j, and Keras to train your distributed models. By the end of this book, you'll have gained experience with the implementation of your models on a variety of use cases. What you will learnUnderstand the basics of deep learningSet up Apache Spark for deep learningUnderstand the principles of distribution modeling and different types of neural networksObtain an understanding of deep learning algorithmsDiscover textual analysis and deep learning with SparkUse popular deep learning frameworks, such as Deeplearning4j, TensorFlow, and KerasExplore popular deep learning algorithms Who this book is for If you are a Scala developer, data scientist, or data analyst who wants to learn how to use Spark for implementing efficient deep learning models, Hands-On Deep Learning with Apache Spark is for you. Knowledge of the core machine learning concepts and some exposure to Spark will be helpful.

Algorithms and Architectures for Parallel Processing - Guojun Wang 2015-11-16

This four volume set LNCS 9528, 9529, 9530 and 9531 constitutes the refereed proceedings of the 15th International Conference on Algorithms and Architectures for Parallel Processing, ICA3PP 2015, held in Zhangjiajie, China, in November 2015. The 219 revised full papers presented together with 77 workshop papers in these four volumes were carefully reviewed and selected from 807 submissions (602 full papers and 205 workshop papers). The first volume comprises the following topics: parallel and distributed

architectures; distributed and network-based computing and internet of things and cyber-physical-social computing. The second volume comprises topics such as big data and its applications and parallel and distributed algorithms. The topics of the third volume are: applications of parallel and distributed computing and service dependability and security in distributed and parallel systems. The covered topics of the fourth volume are: software systems and programming models and performance modeling and evaluation.

Optimizing Databricks Workloads - Anirudh Kala 2021-12-24

Accelerate computations and make the most of your data effectively and efficiently on Databricks Key FeaturesUnderstand Spark optimizations for big data workloads and maximizing performanceBuild efficient big data engineering pipelines with Databricks and Delta LakeEfficiently manage Spark clusters for big data processingBook Description Databricks is an industry-leading, cloud-based platform for data analytics, data science, and data engineering supporting thousands of organizations across the world in their data journey. It is a fast, easy, and collaborative Apache Spark-based big data analytics platform for data science and data engineering in the cloud. In Optimizing Databricks Workloads, you will get started with a brief introduction to Azure Databricks and quickly begin to understand the important optimization techniques. The book covers how to select the optimal Spark cluster configuration for running big data processing and workloads in Databricks, some very useful optimization techniques for Spark DataFrames, best practices for optimizing Delta Lake, and techniques to optimize Spark jobs through Spark core. It contains an opportunity to learn about some of the real-world scenarios where optimizing workloads in Databricks has helped organizations increase performance and save costs across various domains. By the end of this book, you will be prepared with the necessary toolkit to speed up your Spark jobs and process your data more efficiently. What you will learnGet to grips with Spark fundamentals and the Databricks platformProcess big data using the Spark DataFrame API with Delta LakeAnalyze data using graph processing in DatabricksUse MLflow to manage machine learning life cycles in DatabricksFind out how to choose the right cluster configuration for your workloadsExplore file compaction and clustering methods to tune Delta tablesDiscover advanced optimization techniques to speed up Spark jobsWho this book is for This book is for data engineers, data scientists, and cloud architects who have working knowledge of Spark/Databricks and some basic understanding of data engineering principles. Readers will need to have a working knowledge of Python, and some experience of SQL in PySpark and Spark SQL is beneficial.

Innovative Computing - Jason C. Hung 2022-01-04

This book comprises select proceedings of the 4th International Conference on Innovative Computing (IC 2021) focusing on cutting-edge research carried out in the areas of information technology, science, and engineering. Some of the themes covered in this book are cloud communications and networking, high performance computing, architecture for secure and interactive IoT, satellite communication, wearable network and system, infrastructure management, etc. The essays are written by leading international experts, making it a valuable resource for researchers and practicing engineers alike.

Distributed Object-Oriented Architectures - Josef Stepisnik 2007

This document intends to offer a detailed discussion of selected distributed object-oriented architectures at conceptual level. The first part of the discussion offers a comprehensive overview of the Socket architecture in Java 2 and Berkeley UNIX and the distributed object model of Java Remote Method Invocation and the Common Object Request Broker Architecture. The second part concludes the discussion with a comparative study of selected features with emphasis on the Common Object Request Broker Architecture and Java Remote Method Invocation. Major Issues Include The TCP/IP Protocol Suite. We provide an introductory

overview of the TCP/IP protocol suite and its architecture including layers and protocols. The TCP/IP architecture is based on three concepts: processes, layers and protocols. Sockets in Berkeley Unix. We present the Berkeley UNIX socket architecture in relation to the Internet communication domain and illustrate connection-oriented and a connectionless models of communication. Sockets in Java 2. We describe the Java 2 socket architecture, outline selected socket operations, introduce related packages and classes and conclude with a framework for a connection-oriented and connectionless model of communication. Remote Method Invocation in Java 2. We present a distributed object model in Java RMI, provide an overview of related interfaces, classes and packages and discuss security related issues. We conclude with the development of a framework for a distributed object application. Common Object Request Broker Architecture. We introduce a distributed object model for the Common Object Request Broker Architecture and outline design concepts including the Interface Definition Language and the Interoperable Naming Service. We conclude with the development of a framework for a distributed object application. Comparative Study of Distributed Architectures. We present a comparative study of socket architectures and distributed object models introduced in part o

Creating Autonomous Vehicle Systems - Liu Shaoshan 2017-10-25

This book is the first technical overview of autonomous vehicles written for a general computing and engineering audience. The authors share their practical experiences of creating autonomous vehicle systems. These systems are complex, consisting of three major subsystems: (1) algorithms for localization, perception, and planning and control; (2) client systems, such as the robotics operating system and hardware platform; and (3) the cloud platform, which includes data storage, simulation, high-definition (HD) mapping, and deep learning model training. The algorithm subsystem extracts meaningful information from sensor raw data to understand its environment and make decisions about its actions. The client subsystem integrates these algorithms to meet real-time and reliability requirements. The cloud platform provides offline computing and storage capabilities for autonomous vehicles. Using the cloud platform, we are able to test new algorithms and update the HD map—plus, train better recognition, tracking, and decision models. This book consists of nine chapters. Chapter 1 provides an overview of autonomous vehicle systems; Chapter 2 focuses on localization technologies; Chapter 3 discusses traditional techniques used for perception; Chapter 4 discusses deep learning based techniques for perception; Chapter 5 introduces the planning and control sub-system, especially prediction and routing technologies; Chapter 6 focuses on motion planning and feedback control of the planning and control subsystem; Chapter 7 introduces reinforcement learning-based planning and control; Chapter 8 delves into the details of client systems design; and Chapter 9 provides the details of cloud platforms for autonomous driving. This book should be useful to students, researchers, and practitioners alike. Whether you are an undergraduate or a graduate student interested in autonomous driving, you will find herein a comprehensive overview of the whole autonomous vehicle technology stack. If you are an autonomous driving practitioner, the many practical techniques introduced in this book will be of interest to you. Researchers will also find plenty of references for an effective, deeper exploration of the various technologies.

Beginning Spark - Madhukara Phatak 2016-01

With the rise in popularity of distributed systems like Hadoop, more and more people are working in big data processing. A growing number of companies want to build dataflow systems, which can churn huge amounts of data to gain insights for their business. Since Hadoop was a first generation, open source distributed system, there is a need for a next generation distributed system to take data processing to next level. Apache Spark is the next step in that direction. Spark brings a great

flexibility and compositional system to the big data world by revolutionizing the field itself. In this book, the author takes a deep dive into Spark and the big data ecosystem. The author discusses and illustrates how different concepts of Spark are brought together in order to solve complex issues with a data flow system. The reader will acquire an understanding of the Next generation of distribution systems, Apache Spark architecture and abstraction, and the Spark ecosystem including Spark QL, GraphX and MLlib.

Spark - Ilya Ganelin 2016-03-21

Production-targeted Spark guidance with real-world use cases Spark: Big Data Cluster Computing in Production goes beyond general Spark overviews to provide targeted guidance toward using lightning-fast big-data clustering in production. Written by an expert team well-known in the big data community, this book walks you through the challenges in moving from proof-of-concept or demo Spark applications to live Spark in production. Real use cases provide deep insight into common problems, limitations, challenges, and opportunities, while expert tips and tricks help you get the most out of Spark performance. Coverage includes Spark SQL, Tachyon, Kerberos, ML Lib, YARN, and Mesos, with clear, actionable guidance on resource scheduling, db connectors, streaming, security, and much more. Spark has become the tool of choice for many Big Data problems, with more active contributors than any other Apache Software project. General introductory books abound, but this book is the first to provide deep insight and real-world advice on using Spark in production. Specific guidance, expert tips, and invaluable foresight make this guide an incredibly useful resource for real production settings. Review Spark hardware requirements and estimate cluster size Gain insight from real-world production use cases Tighten security, schedule resources, and fine-tune performance Overcome common problems encountered using Spark in production Spark works with other big data tools including MapReduce and Hadoop, and uses languages you already know like Java, Scala, Python, and R. Lightning speed makes Spark too good to pass up, but understanding limitations and challenges in advance goes a long way toward easing actual production implementation. Spark: Big Data Cluster Computing in Production tells you everything you need to know, with real-world production insight and expert guidance, tips, and tricks.

Distributed Systems - Ratan K. Ghosh 2023-02-22

Distributed Systems Comprehensive textbook resource on distributed systems—integrates foundational topics with advanced topics of contemporary importance within the field Distributed Systems: Theory and Applications is organized around three layers of abstractions: networks, middleware tools, and application framework. It presents data consistency models suited for requirements of innovative distributed shared memory applications. The book also focuses on distributed processing of big data, representation of distributed knowledge and management of distributed intelligence via distributed agents. To aid in understanding how these concepts apply to real-world situations, the work presents a case study on building a P2P Integrated E-Learning system. Downloadable lecture slides are included to help professors and instructors convey key concepts to their students. Additional topics discussed in Distributed Systems: Theory and Applications include: Network issues and high-level communication tools Software tools for implementations of distributed middleware. Data sharing across distributed components through publish and subscribe-based message diffusion, gossip protocol, P2P architecture and distributed shared memory. Consensus, distributed coordination, and advanced middleware for building large distributed applications Distributed data and knowledge management Autonomy in distributed systems, multi-agent architecture Trust in distributed systems, distributed ledger, Blockchain and related technologies. Researchers, industry professionals, and students in the fields of science, technology, and medicine will be able to use Distributed

Systems: Theory and Applications as a comprehensive textbook resource for understanding distributed systems, the specifics behind the modern elements which relate to them, and their practical applications.

Fast Data Processing Systems with SMACK Stack - Raul Estrada 2016-12-22

Combine the incredible powers of Spark, Mesos, Akka, Cassandra, and Kafka to build data processing platforms that can take on even the hardest of your data troubles! About This Book This highly practical guide shows you how to use the best of the big data technologies to solve your response-critical problems Learn the art of making cheap-yet-effective big data architecture without using complex Greek-letter architectures Use this easy-to-follow guide to build fast data processing systems for your organization Who This Book Is For If you are a developer, data architect, or a data scientist looking for information on how to integrate the Big Data stack architecture and how to choose the correct technology in every layer, this book is what you are looking for. What You Will Learn Design and implement a fast data Pipeline architecture Think and solve programming challenges in a functional way with Scala Learn to use Akka, the actors model implementation for the JVM Make on memory processing and data analysis with Spark to solve modern business demands Build a powerful and effective cluster infrastructure with Mesos and Docker Manage and consume unstructured and No-SQL data sources with Cassandra Consume and produce messages in a massive way with Kafka In Detail SMACK is an open source full stack for big data architecture. It is a combination of Spark, Mesos, Akka, Cassandra, and Kafka. This stack is the newest technique developers have begun to use to tackle critical real-time analytics for big data. This highly practical guide will teach you how to integrate these technologies to create a highly efficient data analysis system for fast data processing. We'll start off with an introduction to SMACK and show you when to use it. First you'll get to grips with functional thinking and problem solving using Scala. Next you'll come to understand the Akka architecture. Then you'll get to know how to improve the data structure architecture and optimize resources using Apache Spark. Moving forward, you'll learn how to perform linear scalability in databases with Apache Cassandra. You'll grasp the high throughput distributed messaging systems using Apache Kafka. We'll show you how to build a cheap but effective cluster infrastructure with Apache Mesos. Finally, you will deep dive into the different aspect of SMACK using a few case studies. By the end of the book, you will be able to integrate all the components of the SMACK stack and use them together to achieve highly effective and fast data processing. Style and approach With the help of various industry examples, you will learn about the full stack of big data architecture, taking the important aspects in every technology. You will learn how to integrate the technologies to build effective systems rather than getting incomplete information on single technologies. You will learn how various open source technologies can be used to build cheap and fast data processing systems with the help of various industry examples

Distributed Computing and Artificial Intelligence, 14th International Conference - Sigeru Omatu 2017-06-19

The 14th International Symposium on Distributed Computing and Artificial Intelligence 2017 (DCAI 2017) provided a forum for presenting the application of innovative techniques to study and solve complex problems. The exchange of ideas between scientists and technicians from both the academic and industrial sector is essential to advancing the development of systems that can meet the ever-growing demands of today’s society. The book brings together past experience, current work and promising future trends in distributed computing, artificial intelligence and their applications to efficiently solve real-world problems. It combines contributions in well-established and evolving areas of research, including the content of the DCAI 17 Special Sessions, which focused on multi-disciplinary and transversal aspects, such

as AI-driven methods for multimodal networks and processes modeling, and secure management towards smart buildings and smart grids. The symposium was jointly organized by the Polytechnic of Porto, the Osaka Institute of Technology and the University of Salamanca. The latest event was held in Porto, Portugal, from 21st to 23rd June 2017.