Beginning Apache Pig Springer

This book constitutes the refereed proceedings of the 6th International Conference on Big Data analytics, BDA 2018, held in Warangal, India, in December 2018. The 29 papers presented in this volume were carefully reviewed and selected from 93 submissions. The papers are organized in topical sections named: big data analytics: vision and perspectives; financial data analytics and data streams; web and social media data; big data systems and frameworks; predictive analytics in healthcare and agricultural domains; and machine learning and pattern mining.

This book constitutes the thoroughly refereed post-conference proceedings of the 8th TPC Technology Conference, on Performance Evaluation and Benchmarking, TPCTC 2016, held in conjunction with the 41st International Conference on Very Large Databases (VLDB 2016) in New Delhi, India, in September 2016. The 9 papers presented were carefully reviewed and selected from 20 submissions. They reflect the rapid pace at which industry experts and researchers develop innovative techniques for evaluation, measurement and characterization of complex systems.

The book aims to provide a broad overview of various topics of the Internet of Things (IoT) from the research and development priorities to enabling technologies, architecture, security, privacy, interoperability and industrial applications. It is intended to be a stand-alone book in a series that covers the Internet of Things activities of the IERC - Internet of Things European Research Cluster - from technology to international cooperation and the global "state of play." The book builds on the ideas put forward by the European Research Cluster on the Internet of Things Strategic Research and Innovation Agenda and presents views and state of the art results on the challenges facing the research, development and deployment of IoT at the global level. Today we see the integration of Industrial, Business and Consumer Internet which is bringing together the Internet of People, Internet of Things, Internet of Energy, Internet of Vehicles, Internet of Media, Services and Enterprises in forming the backbone of the digital economy, the digital society and the foundation for the future knowledge and innovation based economy. These developments are supporting solutions for the emerging challenges of public health, aging population, environmental protection and climate change, the conservation of energy and scarce materials, enhancements to safety and security and the continuation and growth of economic prosperity. Penetration of smartphones and advances in nanoelectronics, cyber-physical systems, wireless communication, software, and Cloud computing technology will be the main drivers for IoT development. The IoT contribution is seen in the increased value of information created by the number of interconnections among things and the transformation of the processed information into knowledge shared into the Internet of Everything. The connected devices are part of ecosystems connecting people, processes, data, and things which are communicating in the Cloud using the increased storage and computing power while attempting to standardize communication and metadata. In this context, the next generation of Cloud computing technologies will need to be flexible enough to scale autonomously, adaptive enough to handle constantly changing connections and resilient enough to stand up to the huge flows of data that will occur. In 2025, analysts forecast that there will be six devices per human on the planet, which means around 50 billion more connected devices over the next 12 years. The Internet of Things market is connected to this anticipated device growth from industrial Machine to Machine (M2M) systems, smart meters and wireless sensors. Internet of Things technology will generate new services and new interfaces by creating smart environments and smart spaces with applications ranging from Smart Cities, Smart Transport, Buildings, Energy, Grid, to Smart Health and Life.

This book constitutes the refereed proceedings of six workshops of the 14th International Conference on Web-Age Information Management, WAIM 2013, held in Beidaihe, China, June

2013. The 37 revised full papers are organized in topical sections on the six following workshops: The International Workshop on Big Data Management on Emerging Hardware (HardBD 2013), the Second International Workshop on Massive Data Storage and Processing (MDSP 2013), the First International Workshop on Emergency Management in Big Data Age (BigEM 2013), the International Workshop on Trajectory Mining in Social Networks (TMSN 2013), the First International Workshop on Location-based Query Processing in Mobile Environments (LQPM 2013), and the First International Workshop on Big Data Management and Service (BDMS 2013).

Beginning Apache PigBig Data Processing Made EasyApress

Develop applications for the big data landscape with Spark and Hadoop. This book also explains the role of Spark in developing scalable machine learning and analytics applications with Cloud technologies. Beginning Apache Spark 2 gives you an introduction to Apache Spark and shows you how to work with it. Along the way, you'll discover resilient distributed datasets (RDDs); use Spark SQL for structured data; and learn stream processing and build real-time applications with Spark Structured Streaming. Furthermore, you'll learn the fundamentals of Spark ML for machine learning and much more. After you read this book, you will have the fundamentals to become proficient in using Apache Spark and know when and how to apply it to your big data applications. What You Will Learn Understand Spark unified data processing platform How to run Spark in Spark Shell or Databricks Use and manipulate RDDs Deal with structured data using Spark SQL through its operations and advanced functions Build real-time applications using Spark Structured Streaming Develop intelligent applications with the Spark Machine Learning library Who This Book Is For Programmers and developers active in big data, Hadoop, and Java but who are new to the Apache Spark platform.

Learn all you need to know about seven key innovations disrupting business analytics today. These innovations—the open source business model, cloud analytics, the Hadoop ecosystem, Spark and in-memory analytics, streaming analytics, Deep Learning, and self-service analytics—are radically changing how businesses use data for competitive advantage. Taken together, they are disrupting the business analytics value chain, creating new opportunities. Enterprises who seize the opportunity will thrive and prosper, while others struggle and decline: disrupt or be disrupted. Disruptive Business Analytics provides strategies to profit from disruption. It shows you how to organize for insight, build and provision an open source stack, how to practice lean data warehousing, and how to assimilate disruptive innovations into an organization. Through a short history of business analytics and a detailed survey of products and services, analytics authority Thomas W. Dinsmore provides a practical explanation of the most compelling innovations available today. What You'll Learn Discover how the open source business model works and how to make it work for you See how cloud computing completely changes the economics of analytics Harness the power of Hadoop and its ecosystem Find out why Apache Spark is everywhere Discover the potential of streaming and real-time analytics Learn what Deep Learning can do and why it matters See how self-service analytics can change the way organizations do business Who This Book Is For Corporate actors at all levels of responsibility for analytics: analysts, CIOs, CTOs, strategic decision makers, managers, systems architects, technical marketers, product developers, IT personnel, and consultants.

The four-volume set LNCS 8517, 8518, 8519 and 8520 constitutes the proceedings of the Third International Conference on Design, User Experience, and Usability, DUXU 2014, held as part of the 16th International Conference on Human-Computer Interaction, HCII 2014, held in Heraklion, Crete, Greece in June 2014, jointly with 13 other thematically similar conferences. The total of 1476 papers and 220 posters presented at the HCII 2014 conferences were carefully reviewed and selected from 4766 submissions. These papers address the latest research and development efforts and highlight the human aspects of design and use of computing systems. The papers accepted for presentation thoroughly cover the entire field of

Human-Computer Interaction, addressing major advances in knowledge and effective use of computers in a variety of application areas. The total of 256 contributions included in the DUXU proceedings were carefully reviewed and selected for inclusion in this four-volume set. The 76 papers included in this volume are organized in topical sections on design for the web, design for the mobile experience, design of visual information, design for novel interaction techniques and realities, games and gamification.

Now in its second edition, this book focuses on practical algorithms for mining data from even the largest datasets.

This book constitutes the refereed proceedings of the 19th International Conference on Data Analytics and Management in Data Intensive Domains, DAMDID/RCDL 2017, held in Moscow, Russia, in October 2017. The 16 revised full papers presented together with three invited papers were carefully reviewed and selected from 75 submissions. The papers are organized in the following topical sections: data analytics; next generation genomic sequencing: challenges and solutions; novel approaches to analyzing and classifying of various astronomical entities and events; ontology population in data intensive domains; heterogeneous data integration issues; data curation and data provenance support; and temporal summaries generation.

In order to carry out data analytics, we need powerful and flexible computing software. However the software available for data analytics is often proprietary and can be expensive. This book reviews Apache tools, which are open source and easy to use. After providing an overview of the background of data analytics, covering the different types of analysis and the basics of using Hadoop as a tool, it focuses on different Hadoop ecosystem tools, like Apache Flume, Apache Spark, Apache Storm, Apache Hive, R, and Python, which can be used for different types of analysis. It then examines the different machine learning techniques that are useful for data analytics, and how to visualize data with different graphs and charts. Presenting data analytics from a practice-oriented viewpoint, the book discusses useful tools and approaches for data analytics, supported by concrete code examples. The book is a valuable reference resource for graduate students and professionals in related fields, and is also of interest to general readers with an understanding of data analytics.

This book discusses the revolution of cycles and rhythms that is expected to take place in different branches of science and engineering in the 21st century, with a focus on communication and information processing. It presents high-quality papers in vibration sciences, rhythms and oscillations, neurosciences, mathematical sciences, and communication. It includes major topics in engineering and structural mechanics, computer sciences, biophysics and biomathematics, as well as other related fields. Offering valuable insights, it also inspires researchers to work in these fields. The papers included in this book were presented at the 1st International Conference on Engineering Vibration, Communication and Information Processing (ICoEVCI-2018), India.

Take a journey toward discovering, learning, and using Apache Spark 3.0. In this book, you will gain expertise on the powerful and efficient distributed data processing engine inside of Apache Spark; its user-friendly, comprehensive, and flexible programming model for processing data in batch and streaming; and the scalable machine learning algorithms and practical utilities to build machine learning applications. Beginning Apache Spark 3 begins by explaining different ways of interacting with Apache Spark, such as Spark Concepts and Architecture, and Spark Unified Stack. Next, it offers an

overview of Spark SQL before moving on to its advanced features. It covers tips and techniques for dealing with performance issues, followed by an overview of the structured streaming processing engine. It concludes with a demonstration of how to develop machine learning applications using Spark MLlib and how to manage the machine learning development lifecycle. This book is packed with practical examples and code snippets to help you master concepts and features immediately after they are covered in each section. After reading this book, you will have the knowledge required to build your own big data pipelines, applications, and machine learning applications. What You Will Learn Master the Spark unified data analytics engine and its various components Work in tandem to provide a scalable, fault tolerant and performant data processing engine Leverage the user-friendly and flexible programming model to perform simple to complex data analytics using dataframe and Spark SQL Develop machine learning applications using Spark MLlib Manage the machine learning development lifecycle using MLflow Who This Book Is For Data scientists, data engineers and software developers.

This book is a selection of results obtained within two years of research per- formed under SYNAT - a nation-wide scientific project aiming at creating an infrastructure for scientific content storage and sharing for academia, education and open knowledge society in Poland. The selection refers to the research in artificial intelligence, knowledge discovery and data mining, information retrieval and natural language processing, addressing the problems of implementing intelligent tools for building a scientific information platform.This book is a continuation and extension of the ideas presented in “Intelligent Tools for Building a Scientific Information Platform” published as volume 390 in the same series in 2012. It is based on the SYNAT 2012 Workshop held in Warsaw. The papers included in this volume present an overview and insight into information retrieval, repository systems, text processing, ontology-based systems, text mining, multimedia data processing and advanced software engineering.

Data is at the center of many challenges in system design today. Difficult issues need to be figured out, such as scalability, consistency, reliability, efficiency, and maintainability. In addition, we have an overwhelming variety of tools, including relational databases, NoSQL datastores, stream or batch processors, and message brokers. What are the right choices for your application? How do you make sense of all these buzzwords? In this practical and comprehensive guide, author Martin Kleppmann helps you navigate this diverse landscape by examining the pros and cons of various technologies for processing and storing data. Software keeps changing, but the fundamental principles remain the same. With this book, software engineers and architects will learn how to apply those ideas in practice, and how to make full use of data in modern applications. Peer under the hood of the systems you already use, and learn how to use and operate them more effectively Make informed decisions by identifying the strengths and weaknesses of different tools Navigate the trade-offs around consistency, scalability, fault tolerance, and complexity Understand the distributed systems research upon which modern databases are built Peek behind the scenes of major online services, and learn from their architectures

In this book readers will find technological discussions on the existing and emerging technologies across the different stages of the big data value chain. They will learn about legal aspects of big data, the social impact, and about education needs and

requirements. And they will discover the business perspective and how big data technology can be exploited to deliver value within different sectors of the economy. The book is structured in four parts: Part I “The Big Data Opportunity” explores the value potential of big data with a particular focus on the European context. It also describes the legal, business and social dimensions that need to be addressed, and briefly introduces the European Commission’s BIG project. Part II “The Big Data Value Chain” details the complete big data lifecycle from a technical point of view, ranging from data acquisition, analysis, curation and storage, to data usage and exploitation. Next, Part III “Usage and Exploitation of Big Data” illustrates the value creation possibilities of big data applications in various sectors, including industry, healthcare, finance, energy, media and public services. Finally, Part IV “A Roadmap for Big Data Research” identifies and prioritizes the cross-sectorial requirements for big data research, and outlines the most urgent and challenging technological, economic, political and societal issues for big data in Europe. This compendium summarizes more than two years of work performed by a leading group of major European research centers and industries in the context of the BIG project. It brings together research findings, forecasts and estimates related to this challenging technological context that is becoming the major axis of the new digitally transformed business environment.

This book discusses the advanced databases for the cloud-based application known as NoSQL. It will explore the recent advancements in NoSQL database technology. Chapters on structured, unstructured and hybrid databases will be included to explore bigdata analytics, bigdata storage and processing. The book is likely to cover a wide range of topics such as cloud computing, social computing, bigdata and advanced databases processing techniques.

From cloud computing to big data to mobile technologies, there is a vast supply of information being mined and collected. With an abundant amount of information being accessed, stored, and saved, basic controls are needed to protect and prevent security incidents as well as ensure business continuity. Applications of Security, Mobile, Analytic, and Cloud (SMAC) Technologies for Effective Information Processing and Management is a vital resource that discusses various research findings and innovations in the areas of big data analytics, mobile communication and mobile applications, distributed systems, and information security. With a focus on big data, the internet of things (IoT), mobile technologies, cloud computing, and information security, this book proves a vital resource for computer engineers, IT specialists, software developers, researchers, and graduate-level students seeking current research on SMAC technologies and information security management systems.

Learn to use Apache Pig to develop lightweight big data applications easily and quickly. This book shows you many optimization techniques and covers every context where Pig is used in big data analytics. Beginning Apache Pig shows you how Pig is easy to learn and requires relatively little time to develop big data applications.The book is divided into four parts: the complete features of Apache Pig; integration with other tools; how to solve complex business problems; and optimization of tools.You'll discover topics such as MapReduce and why it cannot

meet every business need; the features of Pig Latin such as data types for each load, store, joins, groups, and ordering; how Pig workflows can be created; submitting Pig jobs using Hue; and working with Oozie. You'll also see how to extend the framework by writing UDFs and custom load, store, and filter functions. Finally you'll cover different optimization techniques such as gathering statistics about a Pig script, joining strategies, parallelism, and the role of data formats in good performance. What You Will Learn• Use all the features of Apache Pig• Integrate Apache Pig with other tools• Extend Apache Pig• Optimize Pig Latin code• Solve different use cases for Pig LatinWho This Book Is ForAll levels of IT professionals: architects, big data enthusiasts, engineers, developers, and big data administrators

This book constitutes the refereed proceedings of the ACM/IFIP/USENIX 14th International Middleware Conference, held in Beijing, China, in December 2013. The 24 revised full papers presented were carefully reviewed and selected from 189 submissions. The papers cover a wide range of topics including design, implementation, deployment and evaluation of middleware for next-generation platforms such as cloud computing, social networks and large-scale storage and distributed systems. The middleware solutions introduced provide features such as availability, efficiency, scalability, fault-tolerance, trustworthy operation and support security and privacy needs.

Re-architect relational applications to NoSQL, integrate relational database management systems with the Hadoop ecosystem, and transform and migrate relational data to and from Hadoop components. This book covers the best-practice design approaches to re-architecting your relational applications and transforming your relational data to optimize concurrency, security, denormalization, and performance. Winner of IBM’s 2012 Gerstner Award for his implementation of big data and data warehouse initiatives and author of Practical Hadoop Security, author Bhushan Lakhe walks you through the entire transition process. First, he lays out the criteria for deciding what blend of re-architecting, migration, and integration between RDBMS and HDFS best meets your transition objectives. Then he demonstrates how to design your transition model. Lakhe proceeds to cover the selection criteria for ETL tools, the implementation steps for migration with SQOOP- and Flume-based data transfers, and transition optimization techniques for tuning partitions, scheduling aggregations, and redesigning ETL. Finally, he assesses the pros and cons of data lakes and Lambda architecture as integrative solutions and illustrates their implementation with real-world case studies. Hadoop/NoSQL solutions do not offer by default certain relational technology features such as role-based access control, locking for concurrent updates, and various tools for measuring and enhancing performance. Practical Hadoop Migration shows how to use open-source tools to emulate such relational functionalities in Hadoop ecosystem components. What You'll Learn The requirements and design methodologies of relational data and NoSQL models How to decide whether you should migrate your relational

applications to big data technologies or integrate them How to transition your relational applications to Hadoop/NoSQL platforms in terms of logical design and physical implementation RDBMS-to-HDFS integration, data transformation, and optimization techniques The situations in which Lambda architecture and data lake solutions should be considered How to select and implement Hadoop-based components and applications to speed transition, optimize integrated performance, and emulate relational functionalities Who This Book Is For The primary readership for Practical Hadoop Migration is database developers, database administrators, enterprise architects, Hadoop/NoSQL developers, and IT leaders. Its secondary readership is project and program managers and advanced students of database and management information systems.

This book presents selected peer-reviewed papers from the International Conference on Artificial Intelligence and Data Engineering (AIDE 2019). The topics covered are broadly divided into four groups: artificial intelligence, machine vision and robotics, ambient intelligence, and data engineering. The book discusses recent technological advances in the emerging fields of artificial intelligence, machine learning, robotics, virtual reality, augmented reality, bioinformatics, intelligent systems, cognitive systems, computational intelligence, neural networks, evolutionary computation, speech processing, Internet of Things, big data challenges, data mining, information retrieval, and natural language processing. Given its scope, this book can be useful for students, researchers, and professionals interested in the growing applications of artificial intelligence and data engineering.

Build, implement and scale distributed deep learning models for large-scale datasets About This Book Get to grips with the deep learning concepts and set up Hadoop to put them to use Implement and parallelize deep learning models on Hadoop's YARN framework A comprehensive tutorial to distributed deep learning with Hadoop Who This Book Is For If you are a data scientist who wants to learn how to perform deep learning on Hadoop, this is the book for you. Knowledge of the basic machine learning concepts and some understanding of Hadoop is required to make the best use of this book. What You Will Learn Explore Deep Learning and various models associated with it Understand the challenges of implementing distributed deep learning with Hadoop and how to overcome it Implement Convolutional Neural Network (CNN) with deeplearning4j Delve into the implementation of Restricted Boltzmann Machines (RBM) Understand the mathematical explanation for implementing Recurrent Neural Networks (RNN) Get hands on practice of deep learning and their implementation with Hadoop. In Detail This book will teach you how to deploy large-scale dataset in deep neural networks with Hadoop for optimal performance. Starting with understanding what deep learning is, and what the various models associated with deep neural networks are, this book will then show you how to set up the Hadoop environment for deep learning. In this book, you will also learn how to overcome the challenges that you face while implementing distributed deep

learning with large-scale unstructured datasets. The book will also show you how you can implement and parallelize the widely used deep learning models such as Deep Belief Networks, Convolutional Neural Networks, Recurrent Neural Networks, Restricted Boltzmann Machines and autoencoder using the popular deep learning library deeplearning4j. Get in-depth mathematical explanations and visual representations to help you understand the design and implementations of Recurrent Neural network and Denoising AutoEncoders with deeplearning4j. To give you a more practical perspective, the book will also teach you the implementation of large-scale video processing, image processing and natural language processing on Hadoop. By the end of this book, you will know how to deploy various deep neural networks in distributed systems using Hadoop. Style and approach This book takes a comprehensive, step-by-step approach to implement efficient deep learning models on Hadoop. It starts from the basics and builds the readers' knowledge as they strengthen their understanding of the concepts. Practical examples are included in every step of the way to supplement the theory.

This volume comprises the select proceedings of the annual convention of the Computer Society of India. Divided into 10 topical volumes, the proceedings present papers on state-of-the-art research, surveys, and succinct reviews. The volumes cover diverse topics ranging from communications networks to big data analytics, and from system architecture to cyber security. This volume focuses on Big Data Analytics. The contents of this book will be useful to researchers and students alike.

This book constitutes the proceedings of the 22nd International Conference on Tools and Algorithms for the Construction and Analysis of Systems, TACAS 2016, which took place in Eindhoven, The Netherlands, in April 2016, held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2016. The 44 full papers presented in this volume were carefully reviewed and selected from 175 submissions. They were organized in topical sections named: abstraction and verification; probabilistic and stochastic systems; synthesis; tool papers; concurrency; tool demos; languages and automata; security; optimization; and competition on software verification – SV-COMP.

The 14 contributed chapters in this book survey the most recent developments in high-performance algorithms for NGS data, offering fundamental insights and technical information specifically on indexing, compression and storage; error correction; alignment; and assembly. The book will be of value to researchers, practitioners and students engaged with bioinformatics, computer science, mathematics, statistics and life sciences.

This book highlights several gaps that have not been addressed in existing cyber security research. It first discusses the recent attack prediction techniques that utilize one or more aspects of information to create attack prediction models. The second part is dedicated to new trends on information fusion and their applicability to cyber security; in particular, graph data analytics for cyber security, unwanted traffic detection and control based on trust management software defined networks, security in wireless

sensor networks & their applications, and emerging trends in security system design using the concept of social behavioral biometric. The book guides the design of new commercialized tools that can be introduced to improve the accuracy of existing attack prediction models. Furthermore, the book advances the use of Knowledge-based Intrusion Detection Systems (IDS) to complement existing IDS technologies. It is aimed towards cyber security researchers.

For many organizations, Hadoop is the first step for dealing with massive amounts of data. The next step? Processing and analyzing datasets with the Apache Pig scripting platform. With Pig, you can batch-process data without having to create a full-fledged application, making it easy to experiment with new datasets. Updated with use cases and programming examples, this second edition is the ideal learning tool for new and experienced users alike. You'll find comprehensive coverage on key features such as the Pig Latin scripting language and the Grunt shell. When you need to analyze terabytes of data, this book shows you how to do it efficiently with Pig. Delve into Pig's data model, including scalar and complex data types Write Pig Latin scripts to sort, group, join, project, and filter your data Use Grunt to work with the Hadoop Distributed File System (HDFS) Build complex data processing pipelines with Pig's macros and modularity features Embed Pig Latin in Python for iterative processing and other advanced tasks Use Pig with Apache Tez to build high-performance batch and interactive data processing applications Create your own load and store functions to handle data formats and storage mechanisms

This book comprises select proceedings of the international conference ETAEERE 2020, and primarily focuses on renewable energy resources and smart grid technologies. The book provides valuable information on the technology and design of power grid integration on microgrids of green energy sources. Some of the topics covered include solar PV array, hybrid microgrid, daylight harvesting, green computing, photovoltaic applications, nanogrid applications, AC/DC/AC converter for wind energy systems, solar photovoltaic panels, PEM fuel cell system, and biogas run dual-fueled diesel engine. The contents of this book will be useful for researchers and practitioners working in the areas of smart grids and renewable energy generation, distribution, and management.

Cloud computing has become a significant technology trend. Experts believe cloud computing is currently reshaping information technology and the IT marketplace. The advantages of using cloud computing include cost savings, speed to market, access to greater computing resources, high availability, and scalability. Handbook of Cloud Computing includes contributions from world experts in the field of cloud computing from academia, research laboratories and private industry. This book presents the systems, tools, and services of the leading providers of cloud computing; including Google, Yahoo, Amazon, IBM, and Microsoft. The basic concepts of cloud computing and cloud computing applications are also introduced. Current and future technologies applied in cloud computing are also discussed. Case studies, examples, and exercises are provided throughout. Handbook of Cloud Computing is intended for advanced-level students and researchers in computer science and electrical engineering as a reference book. This handbook is also beneficial to computer and system infrastructure designers, developers, business managers, entrepreneurs and investors within the cloud computing related industry.

Knowledge Discovery in Big Data from Astronomy and Earth Observation: Astrogeoinformatics bridges the gap between astronomy and geoscience in the context of applications, techniques and key principles of big data. Machine learning and parallel computing are increasingly becoming cross-disciplinary as the phenomena of Big Data is becoming common place. This book provides insight into the common workflows and data science tools used for big data in astronomy and geoscience. After establishing similarity in data gathering, pre-processing and handling, the data science aspects are illustrated in the context of both fields. Software, hardware and algorithms of big data are addressed. Finally, the book offers insight into the emerging science which combines data and expertise from both fields in studying the effect of cosmos on the earth and its inhabitants.

Get a jump start on using Azure HDInsight and Hadoop Ecosystem components. As most Hadoop and Big Data projects are written in either Java, Scala, or Python, this book minimizes the effort to learn another language and is written from the perspective of a .NET developer. Hadoop components are covered, including Hive, Pig, HBase, Storm, and Spark on Azure HDInsight, and code samples are written in .NET only. Processing Big Data with Azure HDInsight covers the fundamentals of big data, how businesses are using it to their advantage, and how Azure HDInsight fits into the big data world. This book introduces Hadoop and big data concepts and then dives into creating different solutions with HDInsight and the Hadoop Ecosystem. It covers concepts with real-world scenarios and code examples, making sure you get hands-on experience. The best way to utilize this book is to practice while reading. After reading this book you will be familiar with Azure HDInsight and how it can be utilized to build big data solutions, including batch processing, stream analytics, interactive processing, and storing and retrieving data in an efficient manner. What You'll Learn Understand the fundamentals of HDInsight and Hadoop Work with HDInsight cluster Query with Apache Hive and Apache Pig Store and retrieve data with Apache HBase Stream data processing using Apache Storm Work with Apache Spark Who This Book Is For Software developers, technical architects, data scientists/analyts, and Hadoop administrators who want to develop on Microsoft’s managed Hadoop offering, HDInsight

This book features selected papers presented at the 3rd International Conference on Recent Innovations in Computing (ICRIC 2020), held on 20-21 March 2020 at the Central University of Jammu, India, and organized by the university's Department of Computer Science & Information Technology. It includes the latest research in the areas of software engineering, cloud computing, computer networks and Internet technologies, artificial intelligence, information security, database and distributed computing, and digital India.

This book presents high-quality research papers that demonstrate how emerging technologies in the field of intelligent systems can be used to effectively meet global needs. The respective papers highlight a wealth of innovations and experimental results, while also addressing proven IT governance, standards and practices, and new designs and tools that facilitate rapid information flows to the user. The book is divided into five major sections, namely: “Advances in High Performance Computing”, “Advances in Machine and Deep Learning”, “Advances in Networking and Communication”, “Advances in Circuits and Systems in Computing” and “Advances in

Control and Soft Computing".

Learn advanced analytical techniques and leverage existing tool kits to make your analytic applications more powerful, precise, and efficient. This book provides the right combination of architecture, design, and implementation information to create analytical systems that go beyond the basics of classification, clustering, and recommendation. Pro Hadoop Data Analytics emphasizes best practices to ensure coherent, efficient development. A complete example system will be developed using standard third-party components that consist of the tool kits, libraries, visualization and reporting code, as well as support glue to provide a working and extensible end-to-end system. The book also highlights the importance of end-to-end, flexible, configurable, high-performance data pipeline systems with analytical components as well as appropriate visualization results. You'll discover the importance of mix-and-match or hybrid systems, using different analytical components in one application. This hybrid approach will be prominent in the examples. What You'll Learn Build big data analytic systems with the Hadoop ecosystem Use libraries, tool kits, and algorithms to make development easier and more effective Apply metrics to measure performance and efficiency of components and systems Connect to standard relational databases, noSQL data sources, and more Follow case studies with example components to create your own systems Who This Book Is For Software engineers, architects, and data scientists with an interest in the design and implementation of big data analytical systems using Hadoop, the Hadoop ecosystem, and other associated technologies.

Today's malware mutates randomly to avoid detection, but reactively adaptive malware is more intelligent, learning and adapting to new computer defenses on the fly. Using the same algorithms that antivirus software uses to detect viruses, reactively adaptive malware deploys those algorithms to outwit antivirus defenses and to go undetected. This book provides details of the tools, the types of malware the tools will detect, implementation of the tools in a cloud computing framework and the applications for insider threat detection.

This book offers a comprehensible overview of Big Data Preprocessing, which includes a formal description of each problem. It also focuses on the most relevant proposed solutions. This book illustrates actual implementations of algorithms that helps the reader deal with these problems. This book stresses the gap that exists between big, raw data and the requirements of quality data that businesses are demanding. This is called Smart Data, and to achieve Smart Data the preprocessing is a key step, where the imperfections, integration tasks and other processes are carried out to eliminate superfluous information. The authors present the concept of Smart Data through data preprocessing in Big Data scenarios and connect it with the emerging paradigms of IoT and edge computing, where the end points generate Smart Data without completely relying on the cloud. Finally, this book provides some novel areas of study that are gathering a deeper attention on the Big Data preprocessing. Specifically, it considers the relation with Deep Learning (as of a technique that also relies in large volumes of data), the difficulty of finding the appropriate selection and concatenation of preprocessing techniques applied and some other open problems. Practitioners and data scientists who work in this field, and want to introduce themselves to preprocessing in large data volume scenarios will want to purchase this book. Researchers that work in this field, who want to know which algorithms are currently

implemented to help their investigations, may also be interested in this book.
Leverage Phoenix as an ANSI SQL engine built on top of the highly distributed and scalable NoSQL framework HBase. Learn the basics and best practices that are being adopted in Phoenix to enable a high write and read throughput in a big data space. This book includes real-world cases such as Internet of Things devices that send continuous streams to Phoenix, and the book explains how key features such as joins, indexes, transactions, and functions help you understand the simple, flexible, and powerful API that Phoenix provides. Examples are provided using real-time data and data-driven businesses that show you how to collect, analyze, and act in seconds. Pro Apache Phoenix covers the nuances of setting up a distributed HBase cluster with Phoenix libraries, running performance benchmarks, configuring parameters for production scenarios, and viewing the results. The book also shows how Phoenix plays well with other key frameworks in the Hadoop ecosystem such as Apache Spark, Pig, Flume, and Sqoop. You will learn how to: Handle a petabyte data store by applying familiar SQL techniques Store, analyze, and manipulate data in a NoSQL Hadoop echo system with HBase Apply best practices while working with a scalable data store on Hadoop and HBase Integrate popular frameworks (Apache Spark, Pig, Flume) to simplify big data analysis Demonstrate real-time use cases and big data modeling techniques Who This Book Is For Data engineers, Big Data administrators, and architects.
The growth of a global digital economy has enabled rapid communication, instantaneous movement of funds, and availability of vast amounts of information. With this come challenges such as the vulnerability of digitalized sociotechnological systems (STSs) to destructive events (earthquakes, disease events, terrorist attacks). Similar issues arise for disruptions to complex linked natural and social systems (from changing climates, evolving urban environments, etc.). This book explores new approaches to the resilience of sociotechnological and natural-social systems in a digital world of big data, extraordinary computing capacity, and rapidly developing methods of Artificial Intelligence. Most of the book's papers were presented at the Workshop on Big Data and Systems Analysis held at the International Institute for Applied Systems Analysis in Laxenburg, Austria in February, 2020. Their authors are associated with the Task Group "Advanced mathematical tools for data-driven applied systems analysis" created and sponsored by CODATA in November, 2018. The world-wide COVID-19 pandemic illustrates the vulnerability of our healthcare systems, supply chains, and social infrastructure, and confronts our notions of what makes a system resilient. We have found that use of AI tools can lead to problems when unexpected events occur. On the other hand, the vast amounts of data available from sensors, satellite images, social media, etc. can also be used to make modern systems more resilient. Papers in the book explore disruptions of complex networks and algorithms that minimize departure from a previous state after a disruption; introduce a multigrammatical framework for the technological and resource bases of today's large-scale industrial systems and the transformations resulting from disruptive events; and explain how robotics can enhance pre-emptive measures or post-disaster responses to increase resiliency. Other papers explore current directions in data processing and handling and principles of FAIRness in data; how the availability of large amounts of data can aid in the development of resilient STSs and challenges to overcome in doing so. The book also addresses interactions between humans and built environments,

focusing on how AI can inform today's smart and connected buildings and make them resilient, and how AI tools can increase resilience to misinformation and its dissemination.

Many corporations are finding that the size of their data sets are outgrowing the capability of their systems to store and process them. The data is becoming too big to manage and use with traditional tools. The solution: implementing a big data system. As Big Data Made Easy: A Working Guide to the Complete Hadoop Toolset shows, Apache Hadoop offers a scalable, fault-tolerant system for storing and processing data in parallel. It has a very rich toolset that allows for storage (Hadoop), configuration (YARN and ZooKeeper), collection (Nutch and Solr), processing (Storm, Pig, and Map Reduce), scheduling (Oozie), moving (Sqoop and Avro), monitoring (Chukwa, Ambari, and Hue), testing (Big Top), and analysis (Hive). The problem is that the Internet offers IT pros wading into big data many versions of the truth and some outright falsehoods born of ignorance. What is needed is a book just like this one: a wide-ranging but easily understood set of instructions to explain where to get Hadoop tools, what they can do, how to install them, how to configure them, how to integrate them, and how to use them successfully. And you need an expert who has worked in this area for a decade—someone just like author and big data expert Mike Frampton. Big Data Made Easy approaches the problem of managing massive data sets from a systems perspective, and it explains the roles for each project (like architect and tester, for example) and shows how the Hadoop toolset can be used at each system stage. It explains, in an easily understood manner and through numerous examples, how to use each tool. The book also explains the sliding scale of tools available depending upon data size and when and how to use them. Big Data Made Easy shows developers and architects, as well as testers and project managers, how to: Store big data Configure big data Process big data Schedule processes Move data among SQL and NoSQL systems Monitor data Perform big data analytics Report on big data processes and projects Test big data systems Big Data Made Easy also explains the best part, which is that this toolset is free. Anyone can download it and—with the help of this book—start to use it within a day. With the skills this book will teach you under your belt, you will add value to your company or client immediately, not to mention your career.