Talend Big Data Sandbox

Find the right big data solution for your business ororganization Big data management is one of the major challenges facingbusiness, industry, and not-for-profit organizations. Data setssuch as customer transactions for a mega-retailer, weather patternsmonitored by meteorologists, or social network activity can quicklyoutpace the capacity of traditional data management tools. If youneed to develop or manage big data solutions, you'll appreciate howthese four experts define, explain, and guide you through this newand often confusing concept. You'll learn what it is, why itmatters, and how to choose and implement solutions that work. Effectively managing big data is an issue of growing importanceto businesses, not-for-profit organizations, government, and ITprofessionals Authors are experts in information management, big data, and avariety of solutions Explains big data in detail and discusses how to select andimplement a solution, security concerns to consider, data storageand presentation issues, analytics, and much more Provides essential information in a no-nonsense,easy-to-understand style that is empowering Big Data For Dummies cuts through the confusion and helpsyou take charge of big data solutions for your organization.

This is the eBook of the printed book and may not include any media, website access codes, or print supplements that may come packaged with the bound book. The Comprehensive, Up-to-Date Apache Hadoop Administration Handbook and Reference “Sam Alapati has worked with production Hadoop clusters for six years. His unique depth of experience has enabled him to write the go-to resource for all administrators looking to spec, size, expand, and secure production Hadoop clusters of any size.” —Paul Dix, Series Editor In Expert Hadoop® Administration, leading Hadoop administrator Sam R. Alapati brings together authoritative knowledge for creating, configuring, securing, managing, and optimizing production Hadoop clusters in any environment. Drawing on his experience with large-scale Hadoop administration, Alapati integrates action-oriented advice with carefully researched explanations of both problems and solutions. He covers an unmatched range of topics and offers an unparalleled collection of realistic examples. Alapati demystifies complex Hadoop environments, helping you understand exactly what happens behind the scenes when you administer your cluster. You’ll gain unprecedented insight as you walk through building clusters from scratch and configuring high availability, performance, security, encryption, and other key attributes. The high-value administration skills you learn here will be indispensable no matter what Hadoop distribution you use or what Hadoop applications you run. Understand Hadoop’s architecture from an administrator’s standpoint Create simple and fully distributed clusters Run MapReduce and Spark applications in a Hadoop cluster Manage and protect Hadoop data and high availability Work with HDFS commands, file permissions, and storage management Move data, and use YARN to allocate resources and schedule jobs Manage job workflows with Oozie and Hue Secure, monitor, log, and optimize Hadoop Benchmark and troubleshoot Hadoop

If you’ve been asked to maintain large and complex Hadoop clusters, this book is a must. Demand for operations-specific material has skyrocketed now that Hadoop is becoming the de facto standard for truly large-scale data processing in the data center. Eric Sammer, Principal Solution Architect at Cloudera, shows you the particulars of running Hadoop in production, from planning, installing, and configuring the system to providing ongoing maintenance. Rather than run through all possible scenarios, this pragmatic operations guide calls out what works, as demonstrated in critical deployments. Get a high-level overview of HDFS and MapReduce: why they exist and how they work Plan a Hadoop deployment, from hardware and OS selection to network requirements Learn setup and configuration details with a list of critical

properties Manage resources by sharing a cluster across multiple groups Get a runbook of the most common cluster maintenance tasks Monitor Hadoop clusters—and learn troubleshooting with the help of real-world war stories Use basic tools and techniques to handle backup and catastrophic failure

If you are a system or application developer interested in learning how to solve practical problems using the Hadoop framework, then this book is ideal for you. This book is also meant for Hadoop professionals who want to find solutions to the different challenges they come across in their Hadoop projects.

The book explains core concepts while providing real world implementation specifics, detailing the administration-related activities with Oracle SOA Suite 11g with a step-by-step approach using real-world examples. The authors demonstrate the use of WLST scripts that administrators can reuse and extend to perform most administration tasks such as deployments, tuning, migration, and installation. If you are an Oracle SOA Suite administrator, WebLogic Server administrator, Database administrator, or developer that needs to administer and secure your Oracle SOA Suite services and applications, then this book is for you. Basic knowledge of Oracle SOA Suite Administration is beneficial, but not necessary.

Data Warehousing in the Age of the Big Data will help you and your organization make the most of unstructured data with your existing data warehouse. As Big Data continues to revolutionize how we use data, it doesn't have to create more confusion. Expert author Krish Krishnan helps you make sense of how Big Data fits into the world of data warehousing in clear and concise detail. The book is presented in three distinct parts. Part 1 discusses Big Data, its technologies and use cases from early adopters. Part 2 addresses data warehousing, its shortcomings, and new architecture options, workloads, and integration techniques for Big Data and the data warehouse. Part 3 deals with data governance, data visualization, information life-cycle management, data scientists, and implementing a Big Data–ready data warehouse. Extensive appendixes include case studies from vendor implementations and a special segment on how we can build a healthcare information factory. Ultimately, this book will help you navigate through the complex layers of Big Data and data warehousing while providing you information on how to effectively think about using all these technologies and the architectures to design the next-generation data warehouse. Learn how to leverage Big Data by effectively integrating it into your data warehouse. Includes real-world examples and use cases that clearly demonstrate Hadoop, NoSQL, HBASE, Hive, and other Big Data technologies Understand how to optimize and tune your current data warehouse infrastructure and integrate newer infrastructure matching data processing workloads and requirements

Hadoop in Action teaches readers how to use Hadoop and write MapReduce programs. The intended readers are programmers, architects, and project managers who have to process large amounts of data offline. Hadoop in Action will lead the reader from obtaining a copy of Hadoop to setting it up in a cluster and writing data analytic programs. The book begins by making the basic idea of Hadoop and MapReduce easier to grasp by applying the default Hadoop installation to a few easy-to-follow tasks, such as analyzing changes in word frequency across a body of documents. The book continues through the basic concepts of MapReduce applications developed using Hadoop, including a close look at framework components, use of Hadoop for a variety of data analysis tasks, and numerous examples of Hadoop in action. Hadoop in Action will explain how to use Hadoop and present design patterns and practices of programming MapReduce. MapReduce is a complex idea both conceptually and in its implementation, and Hadoop users are challenged to learn all the knobs and levers for running Hadoop. This book takes you beyond the mechanics of running Hadoop, teaching you to write meaningful programs in a MapReduce framework. This book assumes the reader will have a basic familiarity with Java, as most code examples will be written in Java. Familiarity with basic statistical

concepts (e.g. histogram, correlation) will help the reader appreciate the more advanced data processing examples. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book.

This volume brings together all the successful peer-reviewed papers submitted for the proceedings of the 43rd conference on Computer Applications and Quantitative Methods in Archaeology that took place in Siena (Italy) from March 31st to April 2nd 2015.

Discover how graph databases can help you manage and query highly connected data. With this practical book, you'll learn how to design and implement a graph database that brings the power of graphs to bear on a broad range of problem domains. Whether you want to speed up your response to user queries or build a database that can adapt as your business evolves, this book shows you how to apply the schema-free graph model to real-world problems. Learn how different organizations are using graph databases to outperform their competitors. With this book's data modeling, query, and code examples, you'll quickly be able to implement your own solution. Model data with the Cypher query language and property graph model Learn best practices and common pitfalls when modeling with graphs Plan and implement a graph database solution in test-driven fashion Explore real-world examples to learn how and why organizations use a graph database Understand common patterns and components of graph database architecture Use analytical techniques and algorithms to mine graph database information

Work with petabyte-scale datasets while building a collaborative, agile workplace in the process. This practical book is the canonical reference to Google BigQuery, the query engine that lets you conduct interactive analysis of large datasets. BigQuery enables enterprises to efficiently store, query, ingest, and learn from their data in a convenient framework. With this book, you'll examine how to analyze data at scale to derive insights from large datasets efficiently. Valliappa Lakshmanan, tech lead for Google Cloud Platform, and Jordan Tigani, engineering director for the BigQuery team, provide best practices for modern data warehousing within an autoscaled, serverless public cloud. Whether you want to explore parts of BigQuery you're not familiar with or prefer to focus on specific tasks, this reference is indispensable.

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

This is a step-by-step tutorial that deals with Microsoft Server 2012 reporting tools:SSRS and Power View. If you are a BI

developer, consultant, or architect who wishes to learn how to use SSRS and Power View, and want to understand the best use for each tool, then this book will get you up and running quickly. No prior experience is required with either tool!

"PostGIS in Action" is the first book devoted entirely to PostGIS. It will help both new and experienced users write spatial queries to solve real-world problems. It also discusses the new features available in PostgreSQL 8.4 and provides tutorials.

Pentaho Data Integration Cookbook Second Edition is written in a cookbook format, presenting examples in the style of recipes.This allows you to go directly to your topic of interest, or follow topics throughout a chapter to gain a thorough in-depth knowledge.Pentaho Data Integration Cookbook Second Edition is designed for developers who are familiar with the basics of Kettle but who wish to move up to the next level.It is also aimed at advanced users that want to learn how to use the new features of PDI as well as and best practices for working with Kettle.

The Globus Toolkit is a key technology in Grid Computing, the exciting new computing paradigm that allows users to share processing power, data, storage, and other computing resources across institutional and geographic boundaries. Globus Toolkit 4: Programming Java Services provides an introduction to the latest version of this widely acclaimed toolkit. Based on the popular web-based The Globus Toolkit 4 Programmer's Tutorial, this book far surpasses that document, providing greater detail, quick reference appendices, and many additional examples. If you’re making the leap into Grid Computing using the Globus Toolkit, you’ll want Globus Toolkit 4: Programming Java Services at your side as you take your first steps. Written for newcomers to Globus Toolkit, but filled with useful information for experienced users. Clearly situates Globus application development within the context of Web Services and evolving Grid standards. Provides detailed coverage of Web Services programming with the Globus Toolkit's Java WS Core component. Covers basic aspects of developing secure services using the Grid Security Infrastructure (GSI). Uses simple, didactic examples throughout the book, but also includes a more elaborate example, the FileBuy application, that showcases common design patterns found in Globus applications. Concludes with useful reference appendices.

This book presents a comprehensive and systematic introduction to transforming process-oriented data into information about the underlying business process, which is essential for all kinds of decision-making. To that end, the authors develop step-by-step models and analytical tools for obtaining high-quality data structured in such a way that complex analytical tools can be applied. The main emphasis is on process mining and data mining techniques and the combination of these methods for process-oriented data. After a general introduction to the business intelligence (BI) process and its constituent tasks in chapter 1, chapter 2 discusses different approaches to modeling in BI applications. Chapter 3 is an overview and provides details of data provisioning, including a section on big data. Chapter 4 tackles data description, visualization, and reporting. Chapter 5 introduces data mining techniques for cross-sectional data. Different techniques for the analysis of temporal data are then detailed in Chapter 6. Subsequently, chapter 7 explains techniques for the analysis of process data, followed by the introduction of analysis techniques for multiple BI perspectives in chapter 8. The book closes with a summary and discussion in chapter 9. Throughout the book, (mostly open source) tools are recommended, described and applied; a more detailed survey on tools can be found in the

appendix, and a detailed code for the solutions together with instructions on how to install the software used can be found on the accompanying website. Also, all concepts presented are illustrated and selected examples and exercises are provided. The book is suitable for graduate students in computer science, and the dedicated website with examples and solutions makes the book ideal as a textbook for a first course in business intelligence in computer science or business information systems. Additionally, practitioners and industrial developers who are interested in the concepts behind business intelligence will benefit from the clear explanations and many examples.

Integrating HadoopTechnics Publications

This book contains a selection of articles from The 2015 World Conference on Information Systems and Technologies (WorldCIST'15), held between the 1st and 3rd of April in Funchal, Madeira, Portugal, a global forum for researchers and practitioners to present and discuss recent results and innovations, current trends, professional experiences and challenges of modern Information Systems and Technologies research, technological development and applications. The main topics covered are: Information and Knowledge Management; Organizational Models and Information Systems; Intelligent and Decision Support Systems; Big Data Analytics and Applications; Software Systems, Architectures, Applications and Tools; Multimedia Systems and Applications; Computer Networks, Mobility and Pervasive Systems; Human-Computer Interaction; Health Informatics; Information Technologies in Education; Information Technologies in Radio communications.

If you are a system or application developer interested in learning how to solve practical problems using the Hadoop framework, then this book is ideal for you. You are expected to be familiar with the Unix/Linux command-line interface and have some experience with the Java programming language. Familiarity with Hadoop would be a plus.

This book highlights state-of-the-art research on big data and the Internet of Things (IoT), along with related areas to ensure efficient and Internet-compatible IoT systems. It not only discusses big data security and privacy challenges, but also energy-efficient approaches to improving virtual machine placement in cloud computing environments. Big data and the Internet of Things (IoT) are ultimately two sides of the same coin, yet extracting, analyzing and managing IoT data poses a serious challenge. Accordingly, proper analytics infrastructures/platforms should be used to analyze IoT data. Information technology (IT) allows people to upload, retrieve, store and collect information, which ultimately forms big data. The use of big data analytics has grown tremendously in just the past few years. At the same time, the IoT has entered the public consciousness, sparking people’s imaginations as to what a fully connected world can offer. Further, the book discusses the analysis of real-time big data to derive actionable intelligence in enterprise applications in several domains, such as in industry and agriculture. It explores possible automated solutions in daily life, including structures for smart cities and automated home systems based on IoT technology, as well as health care systems that manage large amounts of data (big data) to improve clinical decisions. The book addresses the security and privacy of the IoT and big data technologies, while also revealing the impact of IoT technologies on several scenarios in smart cities design. Intended as a comprehensive introduction, it offers in-depth analysis and provides scientists, engineers and professionals the latest techniques, frameworks and strategies used in IoT and big data technologies.

Unique prospective on the big data analytics phenomenon for both business and IT professionals The availability of Big Data, low-cost commodity hardware and new information management and analytics software has produced a unique moment in the history of business. The convergence of these trends means that we have the capabilities required to analyze astonishing data sets quickly and cost-effectively

for the first time in history. These capabilities are neither theoretical nor trivial. They represent a genuine leap forward and a clear opportunity to realize enormous gains in terms of efficiency, productivity, revenue and profitability. The Age of Big Data is here, and these are truly revolutionary times. This timely book looks at cutting-edge companies supporting an exciting new generation of business analytics. Learn more about the trends in big data and how they are impacting the business world (Risk, Marketing, Healthcare, Financial Services, etc.) Explains this new technology and how companies can use them effectively to gather the data that they need and glean critical insights Explores relevant topics such as data privacy, data visualization, unstructured data, crowd sourcing data scientists, cloud computing for big data, and much more.

Hands-On Low-Code Application Development with Salesforce follows the “Clicks, not Code” mantra to develop business applications within the Salesforce environment. This book will help you increase your productivity by understanding the core concepts of metadata-driven development.

A complete guide to Pentaho Kettle, the Pentaho Data lntegration toolset for ETL This practical book is a complete guide to installing, configuring, and managing Pentaho Kettle. If you’re a database administrator or developer, you’ll first get up to speed on Kettle basics and how to apply Kettle to create ETL solutions—before progressing to specialized concepts such as clustering, extensibility, and data vault models. Learn how to design and build every phase of an ETL solution. Shows developers and database administrators how to use the open-source Pentaho Kettle for enterprise-level ETL processes (Extracting, Transforming, and Loading data) Assumes no prior knowledge of Kettle or ETL, and brings beginners thoroughly up to speed at their own pace Explains how to get Kettle solutions up and running, then follows the 34 ETL subsystems model, as created by the Kimball Group, to explore the entire ETL lifecycle, including all aspects of data warehousing with Kettle Goes beyond routine tasks to explore how to extend Kettle and scale Kettle solutions using a distributed “cloud” Get the most out of Pentaho Kettle and your data warehousing with this detailed guide—from simple single table data migration to complex multisystem clustered data integration tasks.

Tom Mohr's book, Scaling the Revenue Engine, has already garnered over 12,000 online readers. This is the book author Geoffrey Moore (Crossing the Chasm) has challenged execs to read ("You really want to read this..."). Same with Tien Tzuo, the CEO of Zuora ("...read this book"). So too with Victor Ho, CEO of FiveStars ("...the most complete resource on driving real growth I've ever seen."). And many more. In Scaling the Revenue Engine, the revenue engine is seen as a whole system, bounded by unit economics. It stretches beyond marketing and sales to also incorporate product, technology, and even accounting. At every stage of revenue engine growth, you uplift maturity by leveraging your deployment of people, tools, workflows and metrics-- always working outward from a clear understanding of customer value.

Plan and Implement Hadoop Virtualization for Maximum Performance, Scalability, and Business Agility Enterprises running Hadoop must absorb rapid changes in big data ecosystems, frameworks, products, and workloads. Virtualized approaches can offer important advantages in speed, flexibility, and elasticity. Now, a world-class team of enterprise virtualization and big data experts guide you through the choices, considerations, and tradeoffs surrounding Hadoop virtualization. The authors help you decide whether to virtualize Hadoop, deploy Hadoop in the cloud, or integrate conventional and virtualized approaches in a blended solution. First, Virtualizing Hadoop reviews big data and Hadoop

from the standpoint of the virtualization specialist. The authors demystify MapReduce, YARN, and HDFS and guide you through each stage of Hadoop data management. Next, they turn the tables, introducing big data experts to modern virtualization concepts and best practices. Finally, they bring Hadoop and virtualization together, guiding you through the decisions you'll face in planning, deploying, provisioning, and managing virtualized Hadoop. From security to multitenancy to day-to-day management, you'll find reliable answers for choosing your best Hadoop strategy and executing it. Coverage includes the following: • Reviewing the frameworks, products, distributions, use cases, and roles associated with Hadoop • Understanding YARN resource management, HDFS storage, and I/O • Designing data ingestion, movement, and organization for modern enterprise data platforms • Defining SQL engine strategies to meet strict SLAs • Considering security, data isolation, and scheduling for multitenant environments • Deploying Hadoop as a service in the cloud • Reviewing the essential concepts, capabilities, and terminology of virtualization • Applying current best practices, guidelines, and key metrics for Hadoop virtualization • Managing multiple Hadoop frameworks and products as one unified system • Virtualizing master and worker nodes to maximize availability and performance • Installing and configuring Linux for a Hadoop environment

Use Hadoop to solve business problems by learning from a rich set of real-life case studies About This Book Solve real-world business problems using Hadoop and other Big Data technologies Build efficient data lakes in Hadoop, and develop systems for various business cases like improving marketing campaigns, fraud detection, and more Power packed with six case studies to get you going with Hadoop for Business Intelligence Who This Book Is For If you are interested in building efficient business solutions using Hadoop, this is the book for you This book assumes that you have basic knowledge of Hadoop, Java, and any scripting language. What You Will Learn Learn about the evolution of Hadoop as the big data platform Understand the basics of Hadoop architecture Build a 360 degree view of your customer using Sqoop and Hive Build and run classification models on Hadoop using BigML Use Spark and Hadoop to build a fraud detection system Develop a churn detection system using Java and MapReduce Build an IoT-based data collection and visualization system Get to grips with building a Hadoop-based Data Lake for large enterprises Learn about the coexistence of NoSQL and In-Memory databases in the Hadoop ecosystem In Detail If you have a basic understanding of Hadoop and want to put your knowledge to use to build fantastic Big Data solutions for business, then this book is for you. Build six real-life, end-to-end solutions using the tools in the Hadoop ecosystem, and take your knowledge of Hadoop to the next level. Start off by understanding various business problems which can be solved using Hadoop. You will also get acquainted with the common architectural patterns which are used to build Hadoop-based solutions. Build a 360-degree view of the customer by working with different types of data, and build an efficient fraud detection system for

a financial institution. You will also develop a system in Hadoop to improve the effectiveness of marketing campaigns. Build a churn detection system for a telecom company, develop an Internet of Things (IoT) system to monitor the environment in a factory, and build a data lake – all making use of the concepts and techniques mentioned in this book. The book covers other technologies and frameworks like Apache Spark, Hive, Sqoop, and more, and how they can be used in conjunction with Hadoop. You will be able to try out the solutions explained in the book and use the knowledge gained to extend them further in your own problem space. Style and approach This is an example-driven book where each chapter covers a single business problem and describes its solution by explaining the structure of a dataset and tools required to process it. Every project is demonstrated with a step-by-step approach, and explained in a very easy-to-understand manner.

R for Business Analytics looks at some of the most common tasks performed by business analysts and helps the user navigate the wealth of information in R and its 4000 packages. With this information the reader can select the packages that can help process the analytical tasks with minimum effort and maximum usefulness. The use of Graphical User Interfaces (GUI) is emphasized in this book to further cut down and bend the famous learning curve in learning R. This book is aimed to help you kick-start with analytics including chapters on data visualization, code examples on web analytics and social media analytics, clustering, regression models, text mining, data mining models and forecasting. The book tries to expose the reader to a breadth of business analytics topics without burying the user in needless depth. The included references and links allow the reader to pursue business analytics topics. This book is aimed at business analysts with basic programming skills for using R for Business Analytics. Note the scope of the book is neither statistical theory nor graduate level research for statistics, but rather it is for business analytics practitioners. Business analytics (BA) refers to the field of exploration and investigation of data generated by businesses. Business Intelligence (BI) is the seamless dissemination of information through the organization, which primarily involves business metrics both past and current for the use of decision support in businesses. Data Mining (DM) is the process of discovering new patterns from large data using algorithms and statistical methods. To differentiate between the three, BI is mostly current reports, BA is models to predict and strategize and DM matches patterns in big data. The R statistical software is the fastest growing analytics platform in the world, and is established in both academia and corporations for robustness, reliability and accuracy. The book utilizes Albert Einstein’s famous remarks on making things as simple as possible, but no simpler. This book will blow the last remaining doubts in your mind about using R in your business environment. Even non-technical users will enjoy the easy-to-use examples. The interviews with creators and corporate users of R make the book very readable. The author firmly believes Isaac Asimov was a better writer in spreading science than any textbook

or journal author.

Integrating Hadoop leverages the discipline of data integration and applies it to the Hadoop open-source software framework for storing data on clusters of commodity hardware. It is packed with the need-to-know for managers, architects, designers, and developers responsible for populating Hadoop in the enterprise, allowing you to harness big data and do it in such a way that the solution: · Complies with (and even extends) enterprise standards · Integrates seamlessly with the existing information infrastructure · Fills a critical role within enterprise architecture. Integrating Hadoop covers the gamut of the setup, architecture and possibilities for Hadoop in the organization, including: · Supporting an enterprise information strategy · Organizing for a successful Hadoop rollout · Loading and extracting of data in Hadoop · Managing Hadoop data once it's in the cluster · Utilizing Spark, streaming data, and master data in Hadoop processes - examples are provided to reinforce concepts.

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.

This book features high-quality papers presented at the International Conference on Computational Intelligence and Communication Technology (CICT 2019) organized by ABES Engineering College, Ghaziabad, India, and held from February 22 to 23, 2019. It includes the latest advances and research findings in fields of computational science and communication such as communication & networking, web & informatics, hardware and software designs, distributed & parallel processing, advanced software engineering, advanced database management systems and bioinformatics. As such, it is of interest to research scholars, students, and engineers around the globe.

Data is arriving faster than you can process it and the overall volumes keep growing at a rate that keeps you awake at night. Hadoop can help you tame the data beast. Effective use of Hadoop however requires a mixture of programming, design, and system administration skills. "Hadoop Beginner's Guide" removes the mystery from Hadoop, presenting Hadoop and related technologies with a focus on building working systems and getting the job done, using cloud services to do so when it makes sense. From basic concepts and initial setup through developing applications and keeping the system running as the data grows, the book gives the understanding needed to effectively use Hadoop to solve real world problems. Starting with the basics of installing and configuring Hadoop, the book explains how to develop applications, maintain the system, and how to use additional products to integrate with other systems. While learning different ways to develop applications to run on Hadoop the book also covers tools such as Hive, Sqoop, and Flume that show how Hadoop can be integrated with relational databases and log collection. In addition to examples on Hadoop clusters on Ubuntu uses of cloud services such as Amazon, EC2 and Elastic MapReduce are covered.

Data analytics is core to business and decision making. The rapid increase in data volume, velocity and variety offers both opportunities and challenges. While open source solutions to store big data, like Hadoop, offer platforms for exploring value and insight from big data, they were not originally developed with data security and governance in mind. Big Data Management discusses numerous policies, strategies and recipes for managing big data. It addresses data security, privacy, controls and life cycle management offering modern principles and open source architectures for successful governance of big data. The author has collected best practices from the world’s leading organizations that have successfully implemented big data platforms. The topics discussed cover the entire data management life cycle, data quality, data stewardship, regulatory considerations, data council, architectural and operational models are presented for successful management of big data. The book is a must-read for data scientists, data engineers and corporate leaders who are implementing big data platforms in their organizations.

Get expert guidance on architecting end-to-end data management solutions with Apache Hadoop. While many sources explain how to use

various components in the Hadoop ecosystem, this practical book takes you through architectural considerations necessary to tie those components together into a complete tailored application, based on your particular use case. To reinforce those lessons, the book's second section provides detailed examples of architectures used in some of the most commonly found Hadoop applications. Whether you're designing a new Hadoop application, or planning to integrate Hadoop into your existing data infrastructure, Hadoop Application Architectures will skillfully guide you through the process. This book covers: Factors to consider when using Hadoop to store and model data Best practices for moving data in and out of the system Data processing frameworks, including MapReduce, Spark, and Hive Common Hadoop processing patterns, such as removing duplicate records and using windowing analytics Giraph, GraphX, and other tools for large graph processing on Hadoop Using workflow orchestration and scheduling tools such as Apache Oozie Near-real-time stream processing with Apache Storm, Apache Spark Streaming, and Apache Flume Architecture examples for clickstream analysis, fraud detection, and data warehousing

Data Quality: The Accuracy Dimension is about assessing the quality of corporate data and improving its accuracy using the data profiling method. Corporate data is increasingly important as companies continue to find new ways to use it. Likewise, improving the accuracy of data in information systems is fast becoming a major goal as companies realize how much it affects their bottom line. Data profiling is a new technology that supports and enhances the accuracy of databases throughout major IT shops. Jack Olson explains data profiling and shows how it fits into the larger picture of data quality. * Provides an accessible, enjoyable introduction to the subject of data accuracy, peppered with real-world anecdotes. * Provides a framework for data profiling with a discussion of analytical tools appropriate for assessing data accuracy. * Is written by one of the original developers of data profiling technology. * Is a must-read for any data management staff, IT management staff, and CIOs of companies with data assets.

A practical guide to implementing a scalable and fast state-of-the-art analytical data estate Key Features Store and analyze data with enterprise-grade security and auditing Perform batch, streaming, and interactive analytics to optimize your big data solutions with ease Develop and run parallel data processing programs using real-world enterprise scenarios Book Description Azure Data Lake, the modern data warehouse architecture, and related data services on Azure enable organizations to build their own customized analytical platform to fit any analytical requirements in terms of volume, speed, and quality. This book is your guide to learning all the features and capabilities of Azure data services for storing, processing, and analyzing data (structured, unstructured, and semi-structured) of any size. You will explore key techniques for ingesting and storing data and perform batch, streaming, and interactive analytics. The book also shows you how to overcome various challenges and complexities relating to productivity and scaling. Next, you will be able to develop and run massive data workloads to perform different actions. Using a cloud-based big data-modern data warehouse-analytics setup, you will also be able to build secure, scalable data estates for enterprises. Finally, you will not only learn how to develop a data warehouse but also understand how to create enterprise-grade security and auditing big data programs. By the end of this Azure book, you will have learned how to develop a powerful and efficient analytical platform to meet enterprise needs. What you will learn Implement data governance with Azure services Use integrated monitoring in the Azure Portal and integrate Azure Data Lake Storage into the Azure Monitor Explore the serverless feature for ad-hoc data discovery, logical data warehousing, and data wrangling Implement networking with Synapse Analytics and Spark pools Create and run Spark jobs with Databricks clusters Implement streaming using Azure Functions, a serverless runtime environment on Azure Explore the predefined ML services in Azure and use them in your app Who this book is for This book is for data architects, ETL developers, or anyone who wants to get well-versed with Azure data services to implement an analytical data estate for their enterprise. The book will also appeal to

data scientists and data analysts who want to explore all the capabilities of Azure data services, which can be used to store, process, and analyze any kind of data. A beginner-level understanding of data analysis and streaming will be required.

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.

Semantic Web technology is already changing how we interact with data on the Web. By connecting random information on the Internet in new ways, Web 3.0, as it is sometimes called, represents an exciting online evolution. Whether you’re a consumer doing research online, a business owner who wants to offer your customers the most useful Web site, or an IT manager eager to understand Semantic Web solutions, Semantic Web For Dummies is the place to start! It will help you: Know how the typical Internet user will recognize the effects of the Semantic Web Explore all the benefits the data Web offers to businesses and decide whether it’s right for your business Make sense of the technology and identify applications for it See how the Semantic Web is about data while the “old” Internet was about documents Tour the architectures, strategies, and standards involved in Semantic Web technology Learn a bit about the languages that make it all work: Resource Description Framework (RDF) and Web Ontology Language (OWL) Discover the variety of information-based jobs that could become available in a data-driven economy You’ll also find a quick primer on tech specifications, some key priorities for CIOs, and tools to help you sort the hype from the reality. There are case studies of early Semantic Web successes and a list of common myths you may encounter. Whether you’re incorporating the Semantic Web in the workplace or using it at home, Semantic Web For Dummies will help you define, develop, implement, and use Web 3.0.

Learn how to use Zoho Creator effectively to benefit your business. This book takes you through a number of real-life scenarios and teaches you how to use Zoho Creator to create solutions for your business, with no technical background needed and with little to no coding required. Sound too good to be true? Not with Zoho Creator. With the help of this book you can create a fully-functional cloud-based app that manages your company information, is elegant to use, and cost-effective to maintain. Get started today. Technology makes our lives easier and there are a large number of resources on offer to help with various tasks, including managing business information. With all the tools, apps, and services to choose from, it is still a daunting and often expensive undertaking for businesses to create solutions that fit their specific requirements. Mastering Zoho Creator will guide you through all of this. What You'll Learn Build Zoho Creator applications properly from the

ground up Design with the user in mind Design with the data in mind Create and launch real world business applications, such as real estate management system Integrate your app with external tools and services Extend the capabilities of other Zoho offerings such as CRM Add advanced features by coding in Deluge scripting language Who This Book Is For Small business owners and solopreneurs who want to create business applications and solution to solve their day-to-day problems without the need for prior technical knowledge, coding, or the help of programmers and expensive external consultants. Solution providers and consultants who want to learn the ins and outs of Zoho tools and create world-class business applications for their clients quickly and efficiently.

The Encyclopedia of Big Data Technologies provides researchers, educators, students and industry professionals with a comprehensive authority over the most relevant Big Data Technology concepts. With over 300 articles written by worldwide subject matter experts from both industry and academia, the encyclopedia covers topics such as big data storage systems, NoSQL database, cloud computing, distributed systems, data processing, data management, machine learning and social technologies, data science. Each peer-reviewed, highly structured entry provides the reader with basic terminology, subject overviews, key research results, application examples, future directions, cross references and a bibliography. The entries are expository and tutorial, making this reference a practical resource for students, academics, or professionals. In addition, the distinguished, international editorial board of the encyclopedia consists of well-respected scholars, each developing topics based upon their expertise.