Hadoop Application Architectures Designing Real World Big Data Applications

The ubiquity of technology has not only brought the need for computer knowledge to every aspect of the modern business world; it has also increased our need to safely store the data we are now creating at a rate never experienced before. Delivery and Adoption of Cloud Computing Services in Contemporary Organizations brings together the best practices for storing massive amounts of data. Highlighting ways cloud services can work effectively in production and in real time, this book is an essential reference source for professionals and academics of various disciplines, such as computer science, consulting, information technology, information and communication sciences, healthcare, and finance.

While traditional databases excel at complex queries over historical data, they are inherently pull-based and therefore ill-equipped to push new information to clients. Systems for data stream management and processing, on the other hand, are natively pushoriented and thus facilitate reactive behavior. However, they do not retain data indefinitely and are therefore not able to answer historical queries. The book provides an overview over the different (push-based) mechanisms for data retrieval in each system class and the semantic differences between them. It also provides a comprehensive overview over the current state of the art in real-time databases. It sfirst includes an in-depth system survey of today's real-time databases: Firebase, Meteor, RethinkDB, Parse, Baqend, and others. Second, the high-level classification scheme illustrated above provides a gentle introduction into the system space of data

management: Abstracting from the extreme system diversity in this field, it helps readers build a mental model of the available options.

"This book presents a closer look at the partnership between service oriented architecture and cloud computing environments while analyzing potential solutions to challenges related to the migration of legacy applications"--Provided by publisher.

Design, implement, and deliver successful streaming applications, machine learning pipelines and graph applications using Spark SQL API About This Book Learn about the design and implementation of streaming applications, machine learning pipelines, deep learning, and large-scale graph processing applications using Spark SQL APIs and Scala. Learn data exploration, data munging, and how to process structured and semi-structured data using real-world datasets and gain hands-on exposure to the issues and challenges of working with noisy and "dirty" real-world data. Understand design considerations for scalability and performance in web-scale Spark application architectures. Who This Book Is For If you are a developer, engineer, or an architect and want to learn how to use Apache Spark in a web-scale project, then this is the book for you. It is assumed that you have prior knowledge of SQL querying. A basic programming knowledge with Scala, Java, R, or Python is all you need to get started with this book. What You Will Learn Familiarize yourself with Spark SQL programming, including working with DataFrame/Dataset API and SQL Perform a series of hands-on exercises with different types of data sources, including CSV, JSON, Avro, MySQL, and MongoDB Perform data quality checks, data visualization, and basic statistical analysis tasks Perform data munging tasks on publically available datasets Learn how to use Spark SQL and Apache Kafka to build streaming applications Learn key performance-tuning tips and tricks in Spark

SQL applications Learn key architectural components and patterns in large-scale Spark SQL applications In Detail In the past year, Apache Spark has been increasingly adopted for the development of distributed applications. Spark SQL APIs provide an optimized interface that helps developers build such applications quickly and easily. However, designing web-scale production applications using Spark SQL APIs can be a complex task. Hence, understanding the design and implementation best practices before you start your project will help you avoid these problems. This book gives an insight into the engineering practices used to design and build real-world, Spark-based applications. The book's hands-on examples will give you the required confidence to work on any future projects you encounter in Spark SQL. It starts by familiarizing you with data exploration and data munging tasks using Spark SQL and Scala. Extensive code examples will help you understand the methods used to implement typical use-cases for various types of applications. You will get a walkthrough of the key concepts and terms that are common to streaming, machine learning, and graph applications. You will also learn key performance-tuning details including Cost Based Optimization (Spark 2.2) in Spark SQL applications. Finally, you will move on to learning how such systems are architected and deployed for a successful delivery of your project. Style and approach This book is a hands-on guide to designing, building, and deploying Spark SQL-centric production applications at scale. This book fills a gap between high–level overview texts that are often too general and low–level detail oriented technical handbooks that lose sight the "big picture". This book discusses SOA from the low–level perspective of middleware, various XML–based technologies, and basic service design. It also examines broader implications of SOA, particularly where it intersects with business process management and process modeling. Concrete overviews will be

provided of the methodologies in those fields, so that students will have a hands–on grasp of how they may be used in the context of SOA.

The theme of CUTE is focused on the various aspects of ubiquitous computing for advances in ubiquitous computing and provides an opportunity for academic and industry professionals to discuss the latest issues and progress in the area of ubiquitous computing. Therefore this book will be include the various theories and practical applications in ubiquitous computing Designing Software Architectures will teach you how to design any software architecture in a systematic, predictable, repeatable, and cost-effective way. This book introduces a practical methodology for architecture design that any professional software engineer can use, provides structured methods supported by reusable chunks of design knowledge, and includes rich case studies that demonstrate how to use the methods. Using realistic examples, you'll master the powerful new version of the proven Attribute-Driven Design (ADD) 3.0 method and will learn how to use it to address key drivers, including quality attributes, such as modifiability, usability, and availability, along with functional requirements and architectural concerns. Drawing on their extensive experience, Humberto Cervantes and Rick Kazman guide you through crafting practical designs that support the full software life cycle, from requirements to maintenance and evolution. You'll learn how to successfully integrate design in your organizational context, and how to design systems that will be built with agile methods. Comprehensive coverage includes Understanding what architecture design involves, and where it fits in the full software development life cycle Mastering core design concepts, principles, and processes Understanding how to perform the steps of the ADD method Scaling design and analysis up or down, including design for pre-sale processes or lightweight architecture reviews Recognizing

and optimizing critical relationships between analysis and design Utilizing proven, reusable design primitives and adapting them to specific problems and contexts Solving design problems in new domains, such as cloud, mobile, or big data

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

As more corporations turn to Hadoop to store and process their most valuable data, the risk of a potential breach of those systems increases exponentially. This practical book not only shows Hadoop administrators and security architects how to protect Hadoop data from

unauthorized access, it also shows how to limit the ability of an attacker to corrupt or modify data in the event of a security breach. Authors Ben Spivey and Joey Echeverria provide in-depth information about the security features available in Hadoop, and organize them according to common computer security concepts. You'll also get real-world examples that demonstrate how you can apply these concepts to your use cases. Understand the challenges of securing distributed systems, particularly Hadoop Use best practices for preparing Hadoop cluster hardware as securely as possible Get an overview of the Kerberos network authentication protocol Delve into authorization and accounting principles as they apply to Hadoop Learn how to use mechanisms to protect data in a Hadoop cluster, both in transit and at rest Integrate Hadoop data ingest into enterprise-wide security architecture Ensure that security architecture reaches all the way to end-user access

Unleash the power of cloud computing using Azure, AWS and Apache Hadoop Description With the advent of internet, there is a complete paradigm shift in the manner we comprehend computing. Need to enable ubiquity, convenient and on-demand access to resources in highly scalable and resilient environments that can be remotely accessed, gave birth to the concept of Cloud computing. The acceptance is so rapid that the notion influences sophisticated innovations in academia, industry and research world-wide and hereby change the landscape of information technology as we thought of. Through this book, the authors tried to incorporate core principles and basic notion of cloud computing in a step-by-step manner and tried to emphasize on key concepts for clear and thorough insight into the subject. Audience This book is intended for students of B.E., B.Tech., B.Sc., M.Sc., M.E., and M.Tech. as a text book. The content is designed keeping in mind the bench marked curriculum of various universities (both

National and International). The book covers not only the technical details of how cloud works but also exhibits the strategy, technical design, and in-depth knowledge required to migrate existing applications to the cloud. Therefore, it makes it relevant for the beginners who wants to learn cloud computing right from the foundation. Aspiring Cloud Computing Researchers Instructors, Academicians and Professionals, if they are familiar with cloud, can use this book to learn various open source cloud computing tools, applications, technologies. They will also get a flavor of various international certification exams available. What will you learn • Learn about the Importance of Cloud Computing in Current Digital Era • Understand the Core concepts and Principles of Cloud Computing with practical benefits • Learn about the Cloud Deployment models and Services • Discover how Cloud Computing Architecture works • Learn about the Load balancing approach and Mobile Cloud Computing (MCC) • Learn about the Virtualization and Service-Oriented Architecture (SOA) concepts • Learn about the various Cloud Computing applications, Platforms and Security concepts • Understand the adoption Cloud Computing technology and strategies for migration to the cloud • Case Studies for Cloud computing adoption - Sub-Saharan Africa and India Key Features • Provides a sound understanding of the Cloud computing concepts, architecture and its applications • Explores the practical benefits of Cloud computing services and deployment models in details • Cloud Computing Architecture, Cloud Computing Life Cycle (CCLC), Load balancing approach, Mobile Cloud Computing (MCC), Google App Engine (GAE) • Virtualization and Service-Oriented Architecture (SOA) • Cloud Computing applications - Google Apps, Dropbox Cloud and Apple iCloud and its uses in various sectors - Education, Healthcare, Politics, Business, and Agriculture • Cloud Computing platforms - Microsoft Azure, Amazon Web Services (AWS),

Open Nebulla, Eucalyptus, Open Stack, Nimbus and The Apache Hadoop Architecture • Adoption of Cloud Computing technology and strategies for migration to the cloud • Cloud computing adoption case studies - Sub-Saharan Africa and India • Chapter-wise Questions with Summary and Examination Model Question papers Table of Contents 1. Foundation of Cloud Computing 2. Cloud Services and Deployment Models 3. Cloud Computing Architecture 4. Virtualization & Service Oriented Architecture 5. Cloud Security and Privacy 6. Cloud Computing Applications 7. Cloud Computing Technologies, Platform and Services 8. Adoption of Cloud Computing 9. Model Paper 1 10. Model Paper 2 11. Model Paper 3 12. Model Paper 4

Explains how to use HDInsight along with HortonWorks Data Platform for Windows to store, manage, analyze, and share Big Data throughout the enterprise. Original.

Society is now completely driven by data with many industries relying on data to conduct business or basic functions within the organization. With the efficiencies that big data bring to all institutions, data is continuously being collected and analyzed. However, data sets may be too complex for traditional data-processing, and therefore, different strategies must evolve to solve the issue. The field of big data works as a valuable tool for many different industries. The Research Anthology on Big Data Analytics, Architectures, and Applications is a complete reference source on big data analytics that offers the latest, innovative architectures and frameworks and explores a variety of applications within various industries. Offering an international perspective, the applications discussed within this anthology feature global representation. Covering topics such as advertising curricula, driven supply chain, and smart cities, this research anthology is ideal for data scientists, data analysts, computer engineers,

software engineers, technologists, government officials, managers, CEOs, professors, graduate students, researchers, and academicians.

There's a lot of information about big data technologies, but splicing these technologies into an end-to-end enterprise data platform is a daunting task not widely covered. With this practical book, you'll learn how to build big data infrastructure both on-premises and in the cloud and successfully architect a modern data platform. Ideal for enterprise architects, IT managers, application architects, and data engineers, this book shows you how to overcome the many challenges that emerge during Hadoop projects. You'll explore the vast landscape of tools available in the Hadoop and big data realm in a thorough technical primer before diving into: Infrastructure: Look at all component layers in a modern data platform, from the server to the data center, to establish a solid foundation for data in your enterprise Platform: Understand aspects of deployment, operation, security, high availability, and disaster recovery, along with everything you need to know to integrate your platform with the rest of your enterprise IT Taking Hadoop to the cloud: Learn the important architectural aspects of running a big data platform in the cloud while maintaining enterprise security and high availability

More and more data-driven companies are looking to adopt stream processing and streaming analytics. With this concise ebook, you'll learn best practices for designing a reliable architecture that supports this emerging big-data paradigm. Authors Ted Dunning and Ellen Friedman (Real World Hadoop) help you explore some of the best technologies to handle stream processing and analytics, with a focus on the upstream queuing or message-passing layer. To illustrate the effectiveness of these technologies, this book also includes specific use cases. Ideal for developers and non-technical people alike, this book describes: Key elements

in good design for streaming analytics, focusing on the essential characteristics of the messaging layerNew messaging technologies, including Apache Kafka and MapR Streams, with links to sample codeTechnology choices for streaming analytics: Apache Spark Streaming, Apache Flink, Apache Storm, and Apache ApexHow stream-based architectures are helpful to support microservicesSpecific use cases such as fraud detection and geo-distributed data streams Ted Dunning is Chief Applications Architect at MapR Technologies, and active in the open source community. He currently serves as VP for Incubator at the Apache Foundation, as a champion and mentor for a large number of projects, and as committer and PMC member of the Apache ZooKeeper and Drill projects. Ted is on Twitter as @ted_dunning. Ellen Friedman, a committer for the Apache Drill and Apache Mahout projects, is a solutions consultant and well-known speaker and author, currently writing mainly about big data topics. With a PhD in Biochemistry, she has years of experience as a research scientist and has written about a variety of technical topics. Ellen is on Twitter as @Ellen_Friedman. 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.

Master alternative Big Data technologies that can do what Hadoop can't: real-time analytics and iterative machine learning. When most technical professionals think of Big Data analytics today, they think of Hadoop. But there are many cutting-edge applications that Hadoop isn't well suited for, especially real-time analytics and contexts requiring the use of iterative machine learning algorithms. Fortunately, several powerful new technologies have been developed specifically for use cases such as these. Big Data Analytics Beyond Hadoop is the first guide specifically designed to help you take the next steps beyond Hadoop. Dr. Vijay Srinivas Agneeswaran introduces the breakthrough Berkeley Data Analysis Stack (BDAS) in detail, including its motivation, design, architecture, Mesos cluster management, performance, and more. He presents realistic use cases and up-to-date example code for: Spark, the next generation in-memory computing technology from UC Berkeley Storm, the parallel real-time Big Data analytics technology from Twitter GraphLab, the next-generation graph processing

paradigm from CMU and the University of Washington (with comparisons to alternatives such as Pregel and Piccolo) Halo also offers architectural and design guidance and code sketches for scaling machine learning algorithms to Big Data, and then realizing them in real-time. He concludes by previewing emerging trends, including real-time video analytics, SDNs, and even Big Data governance, security, and privacy issues. He identifies intriguing startups and new research possibilities, including BDAS extensions and cutting-edge model-driven analytics. Big Data Analytics Beyond Hadoop is an indispensable resource for everyone who wants to reach the cutting edge of Big Data analytics, and stay there: practitioners, architects, programmers, data scientists, researchers, startup entrepreneurs, and advanced students.

"The expert's voice in big data"--Cover.

The go-to guidebook for deploying Big Data solutions withHadoop Today's enterprise architects need to understand how the Hadoopframeworks and APIs fit together, and how they can be integrated todeliver real-world solutions. This book is a practical, detailedguide to building and implementing those solutions, with code-levelinstruction in the popular Wrox tradition. It covers storing datawith HDFS and Hbase, processing data with MapReduce, and automatingdata processing with Oozie. Hadoop security, running Hadoop withAmazon Web Services, best practices, and automating Hadoopprocesses in real time are also covered in depth. With in-depth code examples in Java and XML and the latest onrecent additions to the Hadoop ecosystem, this complete resourcealso covers the use of APIs, exposing their inner workings andallowing architects and developers to better leverage and customizethem. The ultimate guide for developers,

designers, and architectswho need to build and deploy Hadoop applications Covers storing and processing data with various technologies,automating data processing, Hadoop security, and deliveringreal-time solutions Includes detailed, real-world examples and code-levelguidelines Explains when, why, and how to use these tools effectively Written by a team of Hadoop experts in theprogrammer-to-programmer Wrox style Professional Hadoop Solutions is the reference enterprisearchitects and developers need to maximize the power of Hadoop.

Every enterprise application creates data, whether it's log messages, metrics, user activity, outgoing messages, or something else. And how to move all of this data becomes nearly as important as the data itself. If you're an application architect, developer, or production engineer new to Apache Kafka, this practical guide shows you how to use this open source streaming platform to handle real-time data feeds. Engineers from Confluent and LinkedIn who are responsible for developing Kafka explain how to deploy production Kafka clusters, write reliable event-driven microservices, and build scalable stream-processing applications with this platform. Through detailed examples, you'll learn Kafka's design principles, reliability guarantees, key APIs, and architecture details, including the replication protocol, the controller, and the storage layer. Understand publish-subscribe messaging and how it fits in the big data ecosystem. Explore Kafka producers and consumers for writing and reading messages Understand Kafka patterns and use-case requirements to ensure

reliable data delivery Get best practices for building data pipelines and applications with Kafka Manage Kafka in production, and learn to perform monitoring, tuning, and maintenance tasks Learn the most critical metrics among Kafka's operational measurements Explore how Kafka's stream delivery capabilities make it a perfect source for stream processing systems

The 15th International Symposium on Distributed Computing and Artificial Intelligence 2018 (DCAI 2018) is a forum to present applications of innovative techniques for studying and solving complex problems. The exchange of ideas between scientists and technicians from both the academic and industrial sector is essential to facilitate the development of systems that can meet the ever-increasing demands of today's society. The present edition brings together past experience, current work and promising future trends associated with distributed computing, artificial intelligence and their application in order to provide efficient solutions to real problems. This symposium is organized by the University of Castilla-La Mancha, the Osaka Institute of Technology and the University of Salamanca. The present edition was held in Toledo, Spain, from 20th – 22nd June, 2018.

Data warehouses were developed for many good reasons, such as providing quick query and reporting for business operations, and business performance. However, over the years, due to the explosion of applications and data volume, many existing data warehouses have become difficult to manage. Extract, Transform, and Load (ETL)

processes are taking longer, missing their allocated batch windows. In addition, data types that are required for business analysis have expanded from structured data to unstructured data. The Apache open source Hadoop platform provides a great alternative for solving these problems. IBM® has committed to open source since the early years of open Linux. IBM and Hortonworks together are committed to Apache open source software more than any other company. IBM Power SystemsTM servers are built with open technologies and are designed for mission-critical data applications. Power Systems servers use technology from the OpenPOWER Foundation, an open technology infrastructure that uses the IBM POWER® architecture to help meet the evolving needs of big data applications. The combination of Power Systems with Hortonworks Data Platform (HDP) provides users with a highly efficient platform that provides leadership performance for big data workloads such as Hadoop and Spark. This IBM RedpaperTM publication provides details about Enterprise Data Warehouse (EDW) optimization with Hadoop on Power Systems. Many people know Power Systems from the IBM AIX® platform, but might not be familiar with IBM PowerLinuxTM, so part of this paper provides a Power Systems overview. A quick introduction to Hadoop is provided for those not familiar with the topic. Details of HDP on Power Reference architecture are included that will help both software architects and infrastructure architects understand the design. In the optimization chapter, we describe various topics: traditional EDW offload, sizing guidelines, performance tuning,

IBM Elastic StorageTM Server (ESS) for data-intensive workload, IBM Big SQL as the common structured query language (SQL) engine for Hadoop platform, and tools that are available on Power Systems that are related to EDW optimization. We also dedicate some pages to the analytics components (IBM Data Science Experience (IBM DSX) and IBM SpectrumTM Conductor for Spark workload) for the Hadoop infrastructure.

Get a solid grounding in Apache Oozie, the workflow scheduler system for managing Hadoop jobs. With this hands-on guide, two experienced Hadoop practitioners walk you through the intricacies of this powerful and flexible platform, with numerous examples and real-world use cases. Once you set up your Oozie server, you'll dive into techniques for writing and coordinating workflows, and learn how to write complex data pipelines. Advanced topics show you how to handle shared libraries in Oozie, as well as how to implement and manage Oozie's security capabilities. Install and configure an Oozie server, and get an overview of basic concepts Journey through the world of writing and configuring workflows Learn how the Oozie coordinator schedules and executes workflows based on triggers Understand how Oozie manages data dependencies Use Oozie bundles to package several coordinator apps into a data pipeline Learn about security features and shared library management Implement custom extensions and write your own EL functions and actions Debug workflows and manage Oozie's operational details

This book addresses the usefulness of knowledge discovery through data mining. With this aim, contributors from different fields propose concrete problems and applications showing how data mining and discovering embedded knowledge from raw data can be beneficial to social organizations, domestic spheres, and ICT markets. Data mining or knowledge discovery in databases (KDD) has received increasing interest due to its focus on transforming large amounts of data into novel, valid, useful, and structured knowledge by detecting concealed patterns and relationships. The concept of knowledge is broad and speculative and has promoted epistemological debates in western philosophies. The intensified interest in knowledge management and data mining stems from the difficulty in identifying computational models able to approximate human behaviors and abilities in resolving organizational, social, and physical problems. Current ICT interfaces are not yet adequately advanced to support and simulate the abilities of physicians, teachers, assistants or housekeepers in domestic spheres. And unlike in industrial contexts where abilities are routinely applied, the domestic world is continuously changing and unpredictable. There are challenging questions in this field: Can knowledge locked in conventions, rules of conduct, common sense, ethics, emotions, laws, cultures, and experiences be mined from data? Is it acceptable for automatic systems displaying emotional behaviors to govern complex interactions based solely on the mining of large volumes of data? Discussing multidisciplinary themes, the book proposes computational models able to approximate,

to a certain degree, human behaviors and abilities in resolving organizational, social, and physical problems. The innovations presented are of primary importance for: a. The academic research community b. The ICT market c. Ph.D. students and early stage researchers d. Schools, hospitals, rehabilitation and assisted-living centers e. Representatives from multimedia industries and standardization bodies Beschrijving van vijfentwintig open source applicaties. Big Data Application Architecture Pattern Recipes provides an insight into heterogeneous infrastructures, databases, and visualization and analytics tools used for realizing the architectures of big data solutions. Its problem-solution approach helps in selecting the right architecture to solve the problem at hand. In the process of reading through these problems, you will learn harness the power of new big data opportunities which various enterprises use to attain real-time profits. Big Data Application Architecture Pattern Recipes answers one of the most critical questions of this time 'how do you select the best end-to-end architecture to solve your big data problem?'. The book deals with various mission critical problems encountered by solution architects, consultants, and software architects while dealing with the myriad options available for implementing a typical solution, trying to extract insight from huge volumes of data in real–time and across multiple relational and non-relational data types for

clients from industries like retail, telecommunication, banking, and insurance. The patterns in this book provide the strong architectural foundation required to launch your next big data application. The architectures for realizing these opportunities are based on relatively less expensive and heterogeneous infrastructures compared to the traditional monolithic and hugely expensive options that exist currently. This book describes and evaluates the benefits of heterogeneity which brings with it multiple options of solving the same problem, evaluation of trade-offs and validation of 'fitness-for-purpose' of the solution. 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. Every enterprise application creates data, whether it consists of log messages, metrics, user activity, outgoing messages, or something else. Moving all of this data is just as important as the data itself. This bookâ??s updated second edition shows application architects, developers, and production engineers new to the Kafka open source streaming platform how to handle real-time data feeds. Additional chapters cover Kafkaâ??s AdminClient API, new security features, and tooling changes. Engineers from Confluent and LinkedIn responsible for

developing Kafka explain how to deploy production Kafka clusters, write reliable event-driven microservices, and build scalable stream processing applications with this platform. Through detailed examples, youâ??ll learn Kafkaâ??s design principles, reliability guarantees, key APIs, and architecture details, including the replication protocol, the controller, and the storage layer. Youâ??ll examine: How publish-subscribe messaging fits in the big data ecosystem Kafka producers and consumers for writing and reading messages Patterns and use-case requirements to ensure reliable data delivery Best practices for building data pipelines and applications with Kafka How to perform monitoring, tuning, and maintenance tasks with Kafka in production The most critical metrics among Kafkaâ??s operational measurements Kafkaâ??s delivery capabilities for stream processing systems

The increase in connected devices in the internet of things (IoT) is leading to an exponential increase in the data that an organization is required to manage. To successfully utilize IoT in businesses, big data analytics are necessary in order to efficiently sort through the increased data. The combination of big data and IoT can thus enable new monitoring services and powerful processing of sensory data streams. The Handbook of Research on Big Data and the IoT is a pivotal reference source that provides vital research on emerging trends and recent

innovative applications of big data and IoT, challenges facing organizations and the implications of these technologies on society, and best practices for their implementation. While highlighting topics such as bootstrapping, data fusion, and graph mining, this publication is ideally designed for IT specialists, managers, policymakers, analysts, software engineers, academicians, and researchers. Cloud computing has become integrated into all sectors, from business to quotidian life. Since it has revolutionized modern computing, there is a need for updated research related to the architecture and frameworks necessary to maintain its efficiency. The Handbook of Research on End-to-End Cloud Computing Architecture Design provides architectural design and implementation studies on cloud computing from an end-to-end approach, including the latest industrial works and extensive research studies of cloud computing. This handbook enumerates deep dive and systemic studies of cloud computing from architecture to implementation. This book is a comprehensive publication ideal for programmers, IT professionals, students, researchers, and engineers. Every enterprise application creates data, whether it consists of log messages, metrics, user activity, or outgoing messages. Moving all this data is just as important as the data itself. With this updated edition, application architects, developers, and production engineers new to the Kafka streaming platform will

learn how to handle data in motion. Additional chapters cover Kafka's AdminClient API, transactions, new security features, and tooling changes. Engineers from Confluent and LinkedIn responsible for developing Kafka explain how to deploy production Kafka clusters, write reliable event-driven microservices, and build scalable stream processing applications with this platform. Through detailed examples, you'll learn Kafka's design principles, reliability guarantees, key APIs, and architecture details, including the replication protocol, the controller, and the storage layer. You'll examine: Best practices for deploying and configuring Kafka Kafka producers and consumers for writing and reading messages Patterns and use-case requirements to ensure reliable data delivery Best practices for building data pipelines and applications with Kafka How to perform monitoring, tuning, and maintenance tasks with Kafka in production The most critical metrics among Kafka's operational measurements Kafka's delivery capabilities for stream processing systems

Big Data Imperatives, focuses on resolving the key questions on everyone's mind: Which data matters? Do you have enough data volume to justify the usage? How you want to process this amount of data? How long do you really need to keep it active for your analysis, marketing, and BI applications? Big data is emerging from the realm of one-off projects to mainstream business adoption;

however, the real value of big data is not in the overwhelming size of it, but more in its effective use. This book addresses the following big data characteristics: Very large, distributed aggregations of loosely structured data – often incomplete and inaccessible Petabytes/Exabytes of data Millions/billions of people providing/contributing to the context behind the data Flat schema's with few complex interrelationships Involves time-stamped events Made up of incomplete data Includes connections between data elements that must be probabilistically inferred Big Data Imperatives explains 'what big data can do'. It can batch process millions and billions of records both unstructured and structured much faster and cheaper. Big data analytics provide a platform to merge all analysis which enables data analysis to be more accurate, well-rounded, reliable and focused on a specific business capability. Big Data Imperatives describes the complementary nature of traditional data warehouses and big-data analytics platforms and how they feed each other. This book aims to bring the big data and analytics realms together with a greater focus on architectures that leverage the scale and power of big data and the ability to integrate and apply analytics principles to data which earlier was not accessible. This book can also be used as a handbook for practitioners; helping them on methodology,technical architecture, analytics techniques and best practices. At the same time, this book

intends to hold the interest of those new to big data and analytics by giving them a deep insight into the realm of big data.

Design documents and implement real world e-commerce applications with Couchbase About This Book Get acquainted with Couchbase architecture and design your document-based data schema Implement full text search using industry standard elastic search plugins Develop critical and high performance applications using this hands-on tutorial guide Who This Book Is For If you are new to the NoSQL document system or have little or no experience in NoSQL development and administration and are planning to deploy Couchbase for your next project, then this book is for you. It would be helpful to have a bit of familiarity with Java. What You Will Learn Get acquainted with the concept of NoSQL databases and configure your Couchbase database cluster Maintain Couchbase effectively using the web-based administrative console with ease Enable partition capabilities by making use of Buckets Analyze important design considerations for maintaining relationship between various documents Use Couchbase SDK Java API to store and retrieve document Write views using map/reduce to retrieve documents efficiently Get familiar with N1QL and how to use it in Java applications Integrate Couchbase with Elasticsearch to implement full text search Configure XDCR for disaster recovery and develop ecommerce

application using Couchbase In Detail NoSQL database systems have changed application development in terms of adaptability to dynamics schema and scalability. Compared with the currently available NoSQL database systems, Couchbase is the fastest. Its ease of configuration and powerful features for storing different schema structures, retrieval using map reduce and inbuilt disaster recovery by replicating document across the geographical region, make it one of the most powerful, scalable and comprehensive NoSQL in the market. Couchbase also introduces smart client API for various programming language to integrate the database with the application easily, yet providing very complex features like cluster health awareness. This book achieves its goal by taking up an end-to-end development structure, right from understanding NOSQL document design to implementing full fledged eCommerce application design using Couchbase as a backend. Starting with the architecture of Couchbase to get you up and running, this book quickly takes you through designing a NoSQL document and implementing highly scalable applications using Java API. You will then be introduced to document design and get to know the various ways to administer Couchbase. Followed by this, learn to store documents using bucket. Moving on, you will then learn to store, retrieve and delete documents using smart client base on Java API. You will then retrieve documents using SQL like

syntax call N1QL. Next, you will learn how to write map reduce base views. Finally, you will configure XDCR for disaster recovery and implement an eCommerce application using Couchbase. Style and approach The book starts from absolute basics and slowly moves to more advanced topics ensuring at every step that all concepts and terms are understood by the reader to have complete understanding at every stage. Technical and complex terms are explained in clear and simple language, thus making this book a perfect companion for those who have started their journey to NoSQL using Couchbase Cloud Enterprise Architecture examines enterprise architecture (EA) in the context of the surging popularity of Cloud computing. It explains the different kinds of desired transformations the architectural blocks of EA undergo in light of this strategically significant convergence. Chapters cover each of the contributing architectures of EA—business, information, application, integration, security, and technology—illustrating the current and impending implications of the Cloud on each. Discussing the implications of the Cloud paradigm on EA, the book details the perceptible and positive changes that will affect EA design, governance, strategy, management, and sustenance. The author ties these topics together with chapters on Cloud integration and composition architecture. He also examines the Enterprise Cloud, Federated Clouds, and the vision to establish the InterCloud. Laying out a comprehensive strategy

for planning and executing Cloud-inspired transformations, the book: Explains how the Cloud changes and affects enterprise architecture design, governance, strategy, management, and sustenance Presents helpful information on next-generation Cloud computing Describes additional architectural types such as enterprise-scale integration, security, management, and governance architectures This book is an ideal resource for enterprise architects, Cloud evangelists and enthusiasts, and Cloud application and service architects. Cloud center administrators, Cloud business executives, managers, and analysts will also find the book helpful and inspirational while formulating appropriate mechanisms and schemes for sound modernization and migration of traditional applications to Cloud infrastructures and platforms.

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 Decide whether you should migrate your relational applications to big data technologies or integrate them Transition your relational applications to Hadoop/NoSQL platforms in terms of logical design and physical implementation Discover RDBMS-to-HDFS integration, data transformation, and optimization techniques Consider when to use Lambda architecture and data lake solutions Select and implement Hadoop-based components and applications to speed transition, optimize integrated performance, and emulate relational functionalities Who This Book Is For 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.

Hadoop Application ArchitecturesDesigning Real-World Big Data Applications"O'Reilly Media, Inc."

Let Hadoop For Dummies help harness the power of your data and rein in the information overload Big data has become big business, and companies and organizations of all sizes are struggling to find ways to retrieve valuable information from their massive data sets with becoming overwhelmed. Enter Hadoop and this easy-to-understand For Dummies guide. Hadoop For Dummies helps readers understand the value of big data, make a business case for using Hadoop, navigate the Hadoop ecosystem, and build and manage Hadoop applications and clusters. Explains the origins of Hadoop, its economic benefits, and its functionality and practical applications Helps you find your way around the Hadoop ecosystem, program MapReduce, utilize design patterns, and get your Hadoop cluster up and running quickly and easily Details how to use Hadoop applications for data mining, web analytics and personalization, large-scale text processing, data science, and problem-solving Shows you how to improve the value of your Hadoop cluster, maximize your investment in Hadoop, and avoid common pitfalls when building your Hadoop cluster From programmers challenged with building and maintaining affordable, scaleable data systems to administrators who must deal with huge volumes of information effectively and efficiently, this how-to has something to help you with Hadoop.

Copyright: ef58939ebbba3e982b0a468162051f30