

1. Record Nr.	UNINA9910300655503321
Autore	Frampton Mike
Titolo	Big Data Made Easy : A Working Guide to the Complete Hadoop Toolset // by Michael Frampton
Pubbl/distr/stampa	Berkeley, CA : , : Apress : , : Imprint : Apress, , 2015
ISBN	9781484200940 1484200942
Edizione	[1st ed. 2015.]
Descrizione fisica	1 online resource (381 p.)
Collana	The expert's voice in big data
Disciplina	004 005.7 005.74
Soggetti	Big data Database management Computers Big Data Database Management Information Systems and Communication Service
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Note generali	Includes index.
Nota di contenuto	""Contents at a Glance""; ""Contents""; ""About the Author""; ""About the Technical Reviewer""; ""Acknowledgments""; ""Introduction""; ""Chapter 1: The Problem with Data""; ""A Definition of a€œBig Dataa€?""; ""The Potentials and Difficulties of Big Data""; ""Requirements for a Big Data System""; ""How Hadoop Tools Can Help""; ""My Approach""; ""Overview of the Big Data System""; ""Big Data Flow and Storage""; ""Benefits of Big Data Systems""; ""Whata€?s in This Book""; ""Storage: Chapter 2""; ""Data Collection: Chapter 3""; ""Processing: Chapter 4""; ""Scheduling: Chapter 5"" ""Data Movement: Chapter 6""""Monitoring: Chapter 7""; ""Cluster Management: Chapter 8""; ""Analysis: Chapter 9""; ""ETL: Chapter 10""; ""Reports: Chapter 11""; ""Summary""; ""Chapter 2: Storing and Configuring Data with Hadoop, YARN, and ZooKeeper""; ""An Overview of Hadoop""; ""The Hadoop V1 Architecture""; ""The Differences in Hadoop V2""; ""The Hadoop Stack""; ""Environment Management"";

""Hadoop V1 Installation""; ""Hadoop 1.2.1 Single-Node Installation""; ""1. Set up Bash shell file for hadoop HOME/.bashrc""; ""2. Set up conf/hadoop-env. sh""; ""3. Create Hadoop temporary directory""; ""4. Set up conf/core-site. xml""; ""5. Set up conf/mapred-site. xml""; ""6. Set up file conf/hdfs-site. xml""; ""7. Format the file system""; ""Setting up the Cluster""; ""Running a Map Reduce Job Check""; ""Hadoop User Interfaces""; ""Hadoop V2 Installation""; ""ZooKeeper Installation""; ""Manually Accessing the ZooKeeper Servers""; ""The ZooKeeper Client""; ""Hadoop MRv2 and YARN""; ""Running Another Map Reduce Job Test""; ""Hadoop Commands""; ""Hadoop Shell Commands""; ""Hadoop User Commands""; ""Hadoop Administration Commands""; ""Summary""; ""Chapter 3: Collecting Data with Nutch and Solr""; ""The Environment""; ""Stopping the Servers""; ""Changing the Environment Scripts""; ""Starting the Servers""; ""Architecture 1: Nutch 1.x""; ""Nutch Installation""; ""Solr Installation""; ""Running Nutch with Hadoop 1.8""; ""Architecture 2: Nutch 2.x""; ""Nutch and Solr Configuration""; ""HBase Installation""; ""Gora Configuration""; ""Running the Nutch Crawl""; ""Potential Errors""; ""A Brief Comparison""; ""Summary""; ""Chapter 4: Processing Data with Map Reduce""; ""An Overview of the Word-Count Algorithm""; ""Map Reduce Native""; ""Java Word-Count Example 1""; ""Describing the Example 1 Code""; ""Running the Example 1 Code""; ""Java Word-Count Example 2""; ""Describing the Example 2 Code""; ""Running the Example 2 Code""; ""Comparing the Examples""; ""Map Reduce with Pig""; ""Installing Pig""; ""Running Pig""; ""Pig User-Defined Functions""; ""Map Reduce with Hive""; ""Installing Hive""; ""Hive Word-Count Example""; ""Map Reduce with Perl""; ""Summary""; ""Chapter 5: Scheduling and Workflow""; ""An Overview of Scheduling""; ""The Capacity Scheduler""; ""The Fair Scheduler""; ""Scheduling in Hadoop V1""; ""V1 Capacity Scheduler""

Sommario/riassunto

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.
