1. Record Nr. UNINA9911006763703321

Autore	Owens Jonathan R
Titolo	Hadoop real-world solutions cookbook : Realistic, simple code examples to solve problems at scale with Hadoop and related technologies / / Jonathan R. Owens, Jon Lentz, Brian Femiano
Pubbl/distr/stampa	Birmingham [England], : Packt Pub., 2013
ISBN	1-62198-910-0 1-84951-913-7 1-299-18393-X
Edizione	[1st edition]
Descrizione fisica	1 online resource (316 p.)
Altri autori (Persone)	LentzJon FemianoBrian
Disciplina	004.6 005.74
Soggetti	Electronic data processing - Distributed processing Open source software
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Note generali	Includes index.
Nota di contenuto	Cover; Copyright; Credits; About the Authors; About the Reviewers; www.packtpub.com; Table of Contents; Preface; Chapter 1: Hadoop Distributed File System - Importing and Exporting Data; Introduction; Importing and exporting data into HDFS using Hadoop shell commands; Moving data efficiently between clusters using Distributed Copy; Importing data from MySQL into HDFS using Sqoop; Exporting data from HDFS into MySQL using Sqoop; Configuring Sqoop for Microsoft SQL Server; Exporting data from HDFS into MongoDB; Importing data from MongoDB into HDFS Exporting data from HDFS into MongoDB using PigUsing HDFS in a Greenplum external table; Using Flume to load data into HDFS; Chapter 2: HDFS; Introduction; Reading and writing data to HDFS; Compressing data using LZO; Reading and writing data to SequenceFiles; Using Apache Avro to serialize data; Using Apache Thrift to serialize data; Using Protocol Buffers to serialize data; Setting the replication factor for HDFS; Setting the block size for HDFS; Chapter 3: Extracting and Transforming Data; Introduction; Transforming Apache logs into TSV

format using MapReduce

Using Apache Pig to filter bot traffic from web server logsUsing Apache Pig to sort web server log data by timestamp; Using Apache Pig to sessionize web server log data; Using Python to extend Apache Pig functionality; Using MapReduce and secondary sort to calculate page views; Using Hive and Python to clean and transform geographical event data; Using Python and Hadoop Streaming to perform a time series analytic; Using Multiple Outputs in MapReduce to name output files; Creating custom Hadoop Writable and InputFormat to read geographical event data

Chapter 4: Performing Common Tasks Using Hive, Pig, and MapReduce Introduction; Using Hive to map an external table over weblog data in HDFS; Using Hive to dynamically create tables from the results of a weblog query; Using the Hive string UDFs to concatenate fields in weblog data; Using Hive to intersect weblog IPs and determine the country; Generating n-grams over news archives using MapReduce; Using the distributed cache in MapReduce; to find lines that contain matching keywords over news archives; Using Pig to load a table and perform a SELECT operation with GROUP BY

Chapter 5: Advanced Joins Introduction; Joining data in the Mapper using MapReduce; Joining data using Apache Pig replicated join; Joining sorted data using Apache Pig merge join; Joining skewed data using Apache Pig skewed join; Using a map-side join in Apache Hive to analyze geographical events; Using optimized full outer joins in Apache Hive to analyze geographical events; Joining data using an external key-value store (Redis); Chapter 6: Big Data Analysis; Introduction; Counting distinct IPs in web log data using MapReduce and Combiners Using Hive date UDFs to transform and sort event dates from geographic event data

Sommario/riassunto

Realistic, simple code examples to solve problems at scale with Hadoop and related technologies.