ANALISIS DATA

Tingkat Lanjut:

Map Reduce, Hadoop, dan Analisis Basisdata

Wahyudi
Anasari

eureka
media aksara
Anggota IKAPI
No. 225/JTE/2021

0858 5343 1992
eurekamediaaksara@gmail.com
Jl. Banjaran RT.20 RW.10
Bojongsari - Purbalingga 53362

ISBN 978-623-120-854-5

ANALISIS DATA TINGKAT LANJUT: MAP REDUCE, HADOOP, DAN ANALISIS BASISDATA

Wahyudi
Anasari

PENERBIT CV.EUREKA MEDIA AKSARA

ANALISIS DATA TINGKAT LANJUT: MAP REDUCE, HADOOP, DAN ANALISIS BASISDATA

Penulis : Wahyudi
Anasari

Desain Sampul : Eri Setiawan

Tata Letak : Rizki Rose Mardiana

ISBN : 978-623-120-854-5

Diterbitkan oleh : **EUREKA MEDIA AKSARA, JUNI 2024**
ANGGOTA IKAPI JAWA TENGAH
NO. 225/JTE/2021

Redaksi:
Jalan Banjaran, Desa Banjaran RT 20 RW 10 Kecamatan Bojongsari Kabupaten Purbalingga Telp. 0858-5343-1992
Surel : eurekamediaaksara@gmail.com
Cetakan Pertama : 2024

KATA PENGANTAR

Puji syukur selalu terucap kepada Allah SWT yang sampai saat ini telah memberikan nikmat sehat, sehingga penulis dapat menyelesaikan buku yang berjudul "Analisis Data Tingkat Lanjut: Map Reduce, Hadoop, dan Analisis Basisdata". Penulis mengucapkan banyak terima kasih pada semua pihak yang sudah terlibat dalam proses pembuatan buku ini, sehingga buku ini bisa hadir di hadapan pembaca.

Buku yang berada di tangan pembaca ini terdiri dari 3 Bab, yaitu:

Bab 1 Mapreduce dan Hadoop
Bab 2 Analisis Basis Data
Bab 3 Mengoperasionalkan Proyek Analisis Data

Penulis menyadari bahwa buku ini masih jauh dari kesempurnaan, sejatinya kesempurnaan hanya milik Allah yang Maha Kuasa, Oleh karena itu kritik dan saran yang membangun demi penyempurnaan buku ini sangatlah dibutuhkan. Akhir kata penulis mengucapkan terima kasih, semoga buku ini bisa membawa manfaat bagi pengembangan ilmu pengetahuan.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

ANALISIS DATA TINGKAT LANJUT: MAP REDUCE, HADOOP, DAN ANALISIS BASISDATA

Wahyudi
Anasari

BAB 1 MAPREDUCE DAN HADOOP

Konsep Utama Ekosistem Hadoop MapReduce NoSQL

Pada buku sebelumnya dengan judul, "Association Rules dan Supervised Learning Menggunakan R," hingga buku yang berjudul, "ANALISIS DATA: TIME SERIES DAN TEXT," membahas beberapa metode analisis yang berguna untuk mengklasifikasikan, memprediksi, dan memeriksa hubungan dalam data. Bab ini dan Bab berikutnya " Analisis dalam Basisdata," membahas beberapa aspek pengumpulan, penyimpanan, dan pemrosesan data tidak terstruktur dan data terstruktur. Bab ini menyajikan beberapa teknologi dan alat utama yang terkait dengan perangkat lunak Apache Hadoop, "kerangka kerja yang memungkinkan pemrosesan dataset besar secara terdistribusi di seluruh kelompok komputer yang menggunakan model pemrograman sederhana".

Bab ini berfokus bagaimana Hadoop menyimpan data dalam sistem terdistribusi dan bagaimana Hadoop mengimplementasikan paradigma pemrograman sederhana yang dikenal sebagai MapReduce. Meskipun bab ini membuat beberapa referensi khusus Java, namun referensi khusus Java, satu-satunya pengetahuan prasyarat yang dimaksudkan adalah pemahaman dasar tentang

BAB 2 ANALISIS BASIS DATA

Konsep-konsep Utama MADlib Ekspresi reguler SQL yang ditentukan pengguna Fungsi-fungsi Window

Analisis basisdata adalah istilah yang luas yang menggambarkan pemrosesan data di dalam repositori. Dalam banyak contoh R sebelumnya, data diekstraksi dari sumber data dan dimuat ke dalam R. Salah satu keuntungan dari **analisis basisdata** adalah bahwa kebutuhan untuk memindahkan data ke dalam alat analitik dihilangkan. Selain itu, dengan melakukan **analisis basisdata**, dimungkinkan untuk mendapatkan hasil yang hampir real-time. Aplikasi analisis basisdata termasuk deteksi penipuan transaksi kartu kredit, rekomendasi produk, dan pemilihan iklan web yang disesuaikan untuk pengguna tertentu.

Basisdata sumber terbuka yang populer adalah PostgreSQL. Nama ini merujuk pada sebuah analitik penting dalam basisdata yang dikenal sebagai **Structured Query Language** (SQL). Bab ini membahas dasar dan juga tingkat lanjut dasar maupun lanjutan dalam SQL. Contoh-contoh kode SQL yang disediakan diuji terhadap basisdata Greenplum, yang yang didasarkan pada PostgreSQL. Namun, konsep-konsep yang disajikan dapat diterapkan pada lingkungan SQL lainnya.

BAB 3

MENGOPERASIONALKAN PROYEK ANALISIS DATA

Konsep Utama Mengomunikasikan dan mengoperasionalkan proyek analitik Membuat kiriman akhi Menggunakan kumpulan materi inti untuk audiens yang berbeda Membandingkan area fokus utama untuk sponsor dan analis Memahami prinsip visualisasi data sederhana Membersihkan bagan atau visualisasi

Bab ini berfokus pada fase terakhir dari Siklus Hidup Analisis Data: mengoperasionalkan. Dalam fase ini, tim tim proyek memberikan laporan akhir, kode, dan dokumentasi teknis. Pada akhir fase ini, tim umumnya mencoba untuk membuat proyek percontohan dan mengimplementasikan model yang dikembangkan dari Fase 4 dalam lingkungan produksi. Seperti yang dinyatakan dalam Bab 2, "Siklus Hidup Analisis Data," tim dapat melakukan analisis yang akurat secara teknis, tetapi jika mereka tidak dapat menerjemahkan hasilnya ke dalam bahasa yang audiens mereka, orang lain tidak akan melihat nilainya, dan upaya serta sumber daya yang signifikan akan terbuang percuma. Bab ini berfokus pada bagaimana membuat ringkasan narasi yang jelas dan kerangka kerja untuk menyampaikan narasi tersebut kepada para pemangku kepentingan utama.

DAFTAR PUSTAKA

"apache.org," [Online]. Available: http://www.apache.org/. [Accessed 11 February 2014].

"BSP Tutorial," [Online]. Available: http://hama.apache.org/hama_bsp_tutorial .html. [Accessed 20 February 2014].

"Hadoop Pipes," [Online]. Available: http://hadoop.apache.org/docs/r1.2.1/api/org/apache/hadoop/mapred/pipes/package-summary.html. [Accessed 19 February 2014].

"Hadoop Wiki Disk Setup," [Online]. Available: http://wiki.apache.org/hadoop/ DiskSetup. [Accessed 20 February 2014].

"Hama," [Online]. Available: http://hama.apache.org/. [Accessed 20 February 2014].

"HBase Key Value," [Online]. Available: http://hbase.apache.org/book/regions .arch.html. [Accessed 28 February 2014].

"HBase Regionserver," [Online]. Available: http://hbase.apache.org/book/ regionserver.arch.html. [Accessed 3 March 2014].

"HBase Rowkey," [Online]. Available: http://hbase.apache.org/book/rowkey .design.html. [Accessed 4 March 2014].

"HDFS Design," [Online]. Available: http://hadoop.apache.org/docs/stable1/hdfs_design.html. [Accessed 19 February 2014].

"HDFS High Availability," [Online]. Available: http://hadoop.apache.org/docs/ c urrent/hadoop-yarn/hadoop-yarn-site/

"Mahout," [Online]. Available: http://mahout.apache.org/users/basics/ algorithms.html. [Accessed 19 February 2014].

"Netcraft," [Online]. Available: http://news.netcraft.com/archives/2014/02/03/ february-2014-web-server-survey.html. [Accessed 21 February 2014].

"Number of Column Families," [Online]. Available: http://hbase.apache.org/book/ number.of.cfs.html.

"Pig," [Online]. Available: http://pig.apache.org/. [Accessed 11 Feb 2014].

"pig.apache.org," [Online]. Available: http://pig.apache.org/.

"Piggybank," [Online]. Available: https://cwiki.apache.org/confluence/ display/PIG/PiggyBank. [Accessed 28 February 2014].

"Pivotal HD," [Online]. Available: http://www.gopivotal.com/big-data/ pivotal-hd. [Accessed 8 May 2014].

"PoweredByYarn," [Online]. Available: http://wiki.apache.org/hadoop/ PoweredByYarn. [Accessed 20 February 2014].

"wiki.apache.org/hadoop," [Online]. Available: http://wiki.apache.org/hadoop/ NameNode. [Accessed 11 February 2014].

"Zookeeper," [Online]. Available: http://hbase.apache.org/book/zookeeper .html. [Accessed 21 February 2014].

"Zookeeper," [Online]. Available: http://zookeeper.apache.org/. [Accessed 11 Feb 2014].

Apache, "Apache Hadoop," [Online]. Available: http://hadoop.apache.org/. [Accessed 8 May 2014].

Apache, "Hadoop Streaming," [Online]. Available: https://wiki.apache.org/hadoop/ HadoopStreaming. [Accessed 8 May 2014].

B. Minto, The Minto Pyramid Principle: Logic in Writing, Thinking, and Problem Solving, Prentice Hall, 2010.

D. Cutting, "Free Search: Rambilings About Lucene, Nutch, Hadoop and Other Stuff," [Online]. Available: http://cutting.wordpress.com. [Accessed 11 February 2014].

D. Davidian, "IBM.com," 14 February 2011. [Online]. Available: https://www-304.ibm.com/ connections/ blogs/davidian/tags/hadoop?lang=en_us. Accessed 11 February 2014].

D. Gottfrid, "Self-Service, Prorated Supercomputing Fun!," 01 November 2007. [Online]. Available: http://open.blogs.nytimes.com/2007/11/01/self- service-prorated-super-computing-fun/. [Accessed 11 February 2014].

E. Baldeschwieler, “http://www.slideshare.net,” [Online]. Available: http://www .slideshare.net/ydn/hadoop-yahoo-internet-scale-data-processing. [Accessed 11 February 2014].

E. Dumbill, “The Data Lake Dream,” Forbes, 14 January 2014. [Online]. Available: http://www .forbes.com/sites/edddumbill/2014/01/14/the-data-lake-dream/ . [Accessed 4 June 2014].

Eclipse. [Online]. Available: https://www.eclipse. org/downloads/. [Accessed 27 February 2014].

F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R.E. Gruber Fay Chang, “Bigtable: A Distributed Storage System for Structured Data,” [Online]. Available: http://research.google.com/archive/bigtable.html. [Accessed 11 February 2014].

G. Reynolds, Presentation Zen: Simple Ideas on Presentation Design and Delivery, Berkeley: New Riders, 2011.

G. Zelazny, Say It with Charts: The Executive’s Guide to Visual Communication, McGraw-Hill, 2001.

HDFSHighAvailabilityWithNFS.html. [Accessed 8 May 2014].

IBM, “IBM.com,” [Online]. Available: http://www-03.ibm.com/innovation/us/ watson/watson_in_healthcare.shtml. [Accessed 11 February 2014].

J. Cohen, B. Dolan, M. Dunlap, J. Hellerstein, and C. Welton, “MAD Skills: New Analysis Practices for Big Data,” in

Proceedings of the VLDB Endowment Volume 2 Issue 2, August 2009.

J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," [Online]. Available: http://research.google.com/archive/mapreduce.html. [Accessed 11 February 2014].

K. Muthukkaruppan, "The Underlying Technology of Messages," 15 November 2010. [Online]. Available: http://www.facebook.com/notes/facebook-engineering/the-underlying-technology-of-messages/454991608919. [Accessed 11 February 2014].

LinkedIn, "Hadoop," [Online]. Available: http://data.linkedin.com/projects/ hadoop. [Accessed 11 February 2014].

Linkedin, "LinkedIn," [Online]. Available: http://www.linkedin.com/about-us. [Accessed 11 February 2014].

MADlib, "MADlib Modules" [Online]. Available: http://doc.madlib.net/latest/ modules.html. [Accessed 10 April 2014]

MADlib, "MADlib" [Online]. Available: http://madlib.net/download/. [Accessed 10 April 2014].

N. Spiegelberg. [Online]. Available: http://www.slideshare.net/brizzzdotcom/ facebook-messages-hbase. [Accessed 11 February 2014].

N. Yau, "flowingdata.com" [Online]. Available: http://flowingdata.com.

N. Yau, Visualize This, Indianapolis: Wiley, 2011.

P. J. Sadalage and M. Fowler, *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot*, Upper Saddle River, NJ: Addison Wesley, 2013.

PostgreSQl.org, "Window Functions" [Online]. Available: http://www.postgresql.org/ docs/9.3/static/functions-window.html. [Accessed 10 April 2014].

S. Few, Now You See It: Simple Visualization Techniques for Quantitative Analysis, Analytics Press, 2009.

S. Ghemawat, H. Gobioff, and S.-T. Leung, "The Google File System," [Online]. Available: http:// static.googleusercontent.com/media/research.google.com/en/us/ archive/gfs-sosp2003.pdf. [Accessed 11 February 2014].

S. Singh, "http://developer.yahoo.com/," [Online]. Available: http://developer .yahoo.com/blogs/hadoop/apache-hbase-yahoo-multi-tenancy- helm-again-171710422.html. [Accessed 11 February 2014].

Wikipedia, "IBM Watson," [Online]. Available: http://en.wikipedia.org/wiki/IBM_ Watson. [Accessed 11 Februry 2014].