

Dr. Wahyudi, S.T., M.T.



PENGANTAR SAINS DATA

big data dan visualisasi data

BIG DATA

PENGANTAR SAINS DATA

big data dan visualisasi data

Buku Pengantar Sains Data: big data dan visualisasi data terdiri dari tiga bab. Bab pertama yaitu Menangani Data Besar Pada Satu Komputer menjelaskan Masalah utama yang akan dihadapi ketika bekerja dengan kumpulan data yang besar, solusi dari masalah tersebut, struktur data di sains data, modifikasi algoritma untuk data besar serta tools python yang membantu Anda menangani kumpulan data yang besar. Bab kedua menjelaskan proses big data. Pada bab ini akan dibahas dua framework sains data untuk data besar yaitu Hadoop dan Spark. Framework tersebut bisa dikombinasikan dengan bahasa Python seperti pywebhdfs untuk hadoop dan PySpark untuk spark Bab terakhir menjelaskan tentang Visualisasi data kepada End User. Bab ini berfokus pada bagian terakhir dari proses sains data, dan tujuan kita adalah untuk membangun aplikasi sains data di mana pengguna akhir diberikan dasbor interaktif. Setelah melalui semua langkah proses sains data, kita akan mendapatkan data yang bersih, sering kali dipadatkan atau padat informasi. Dengan cara ini kita dapat meminta lebih sedikit data dan mendapatkan wawasan yang kita inginkan.



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



PENGANTAR SAINS DATA:
BIG DATA DAN VISUALISASI DATA

Dr. Wahyudi,S.T.,M.T.



PENERBIT CV.EUREKA MEDIA AKSARA

**PENGANTAR SAINS DATA:
BIG DATA DAN VISUALISASI DATA**

Penulis : Dr. Wahyudi,S.T.,M.T.

Desain Sampul : Eri Setiawan

Tata Letak : Meilita Anggie Nurlatifah

ISBN : 978-623-151-504-9

Diterbitkan oleh : **EUREKA MEDIA AKSARA,
SEPTEMBER 2023
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 : 2023

All right reserved

Hak Cipta dilindungi undang-undang

Dilarang memperbanyak atau memindahkan sebagian atau seluruh isi buku ini dalam bentuk apapun dan dengan cara apapun, termasuk memfotokopi, merekam, atau dengan teknik perekaman lainnya tanpa seizin tertulis dari penerbit.

KATA PENGANTAR

Alhamdulillahirobbil'alamin. Penulis bersyukur kehadiran Allah SWT berkat rahmat, karunia dan pertolonganNya, penulis dapat menyelesaikan buku berjudul "**Pengantar Sains Data: big data dan visualisasi data**". Shalawat serta salam semoga senantiasa tercurah atas Nabi Muhammad SAW, para kerabat, serta pengikutnya hingga hari kiamat nanti.

Buku ini hadir untuk menambah literasi tentang teknologi informasi di lingkungan Universitas andalas. Buku ini merupakan seri kedua dari beberapa buku sains data. Buku ini menjelaskan tentang munculnya bidang sains data, hubungan sains data dan big data, proses yang dilakukan di sains data dan pembelajaran mesin yang merupakan bagian paling penting di sains data.

Penulis menyadari bahwa dalam penulisan buku ini masih banyak terdapat kekurangan, untuk itu penulis mengharapkan kritik dan sarannya guna penyempurnaan buku ini di masa mendatang

Padang, Agustus 2023

Penulis

DAFTAR ISI

KATA PENGANTAR	iii
DAFTAR ISI	iv
DAFTAR GAMBAR.....	v
DAFTAR TABEL.....	x
BAB 1 MENANGANI DATA BESAR PADA SATU KOMPUTER.....	1
A. Masalah Saat Menangani Data Besar.....	2
B. Teknik Umum Untuk Menangani Volume Data yang Besar	4
C. Kiat Umum Menangani Data Besar	27
D. Studi kasus 1: Memprediksi URL Berbahaya	32
E. Studi kasus 2: Membangun Sistem Pemberi Rekomendasi di Dalam Database	40
F. Rangkuman.....	57
BAB 2 LANGKAH PERTAMA DALAM BIG DATA	59
A. Mendistribusikan Penyimpanan dan Pemrosesan Data Dengan Framework	61
B. Studi kasus: Menilai Risiko Saat Meminjamkan Uang.....	69
C. Rangkuman.....	103
BAB 3 VISUALISASI DATA KEPADA END USER.....	104
A. Opsi Visualisasi Data	107
B. Crossfilter, Library JavaScript MapReduce.....	112
C. Membuat Dasbor Interaktif dengan dc.js.....	128
D. Tool Pengembangan Dasbor	135
E. Ringkasan.....	139
DAFTAR PUSTAKA	142

DAFTAR GAMBAR

Gambar 1.1.	Overview masalah yang dihadapi saat bekerja dengan lebih banyak data.....	3
Gambar 1.2.	Overview untuk menangani kumpulan data besar	5
Gambar 1.3	Overview teknik untuk mengadaptasi algoritma ke kumpulan data besar	6
Gambar 1.4	Blok matriks dapat digunakan untuk menghitung jumlah matriks A dan B.	14
Gambar 1.5	Overview struktur data yang sering digunakan dalam sains data ketika bekerja dengan data besar.....	19
Gambar 1.6	Contoh matriks jarang: hampir semuanya bernilai 0; nilai lainnya adalah pengecualian dalam matriks jarang	19
Gambar 1.7	Contoh struktur data pohon: aturan keputusan seperti kategori usia dapat digunakan untuk menemukan seseorang dengan cepat di dalam silsilah keluarga.....	21
Gambar 1.8	Overview tool yang dapat digunakan saat bekerja dengan data yang besar ..	23
Gambar 1.9	Overview tentang praktik terbaik pemrograman ketika bekerja dengan data yang besar.....	27
Gambar 1.10	Kesalahan memori saat mencoba mengambil kumpulan data yang besar ke dalam memori	33

Gambar 1.11	Informasi 2 pelanggan pertama pada semua 32 film setelah konversi string bit ke angka.....	50
Gambar 1.12	Informasi dari proses kompresi string bit dan 9 sampel film	51
Gambar 2.1	Dasbor interaktif Qlik.....	60
Gambar 2.2	Contoh dari ekosistem aplikasi yang muncul di sekitar Framework Hadoop Core	63
Gambar 2.3	Contoh yang disederhanakan dari alur MapReduce untuk menghitung jumlah warna dalam teks masukan	64
Gambar 2.4	Contoh alur MapReduce untuk menghitung warna dalam teks input .	65
Gambar 2.5	Framework Spark ketika digunakan bersama dengan framework Hadoop.	68
Gambar 2.6	Menghubungkan ke Horton Sandbox menggunakan PuTTY.....	70
Gambar 2.7	Hasil akhir dari latihan ini adalah dasbor penjelasan untuk membandingkan peluang pinjaman dengan peluang serupa.....	73
Gambar 2.8	Keluaran dari perintah Listing Hadoop: <code>hadoop fs -ls /</code> . Folder root Hadoop sudah terdaftar.....	74
Gambar 2.9	Layar pembuka Spark untuk penggunaan interaktif dengan Python.....	76
Gambar 2.10	Mengambil status file di Hadoop melalui konsol PySpark.....	79

Gambar 2.11	Konfigurasi ODBC Windows Hortonworks	88
Gambar 2.12	Layar selamat datang Qlik Sense	89
Gambar 2.13	Kotak pesan Buat aplikasi baru.....	90
Gambar 2.14	Sebuah kotak yang mengonfirmasi bahwa aplikasi berhasil dibuat.....	90
Gambar 2.15	Sebuah pop up Layar menambahkan data saat Anda membuka aplikasi baru.....	90
Gambar 2.16	Pilih ODBC sebagai sumber data di layar Select a data source.....	91
Gambar 2.17	Pilih Hortonworks pada DSN Pengguna dan tentukan nama pengguna dan kata sandi.....	92
Gambar 2.18	Gambaran umum kolom raw data interface Hive	92
Gambar 2.19	Konfirmasi bahwa data telah dimuat di Qlik.....	93
Gambar 2.20	Layar editor untuk laporan yang terbuka.....	93
Gambar 2.21	Seret judul dari panel Bidang sebelah kiri ke panel laporan.....	94
Gambar 2.22	Contoh bagan KPI.....	95
Gambar 2.23	Empat langkah untuk menambahkan bagan KPI ke laporan Qlik.....	95
Gambar 2. 24	Contoh diagram batang.....	96
Gambar 2.25	Menambahkan diagram batang membutuhkan lima langkah.....	97
Gambar 2.26	Contoh tabel pivot, yang menunjukkan tingkat bunga rata-rata	

	yang dibayarkan per jabatan/kombinasi tingkat risiko	98
Gambar 2.27	Menambahkan tabel pivot membutuhkan enam langkah	99
Gambar 2. 28	Hasil akhir dalam mode edit	100
Gambar 2.29	Ketika kita memilih direktur, kita dapat melihat bahwa mereka membayar rata-rata 11,97% untuk sebuah pinjaman	102
Gambar 2.30	Ketika kami memilih artis, kami melihat bahwa mereka membayar bunga rata-rata 13,32% untuk pinjaman	102
Gambar 3.1	Dataset obat-obatan apotek yang dibuka di Excel: 10 baris pertama dari stok pertama data ditingkatkan dengan variabel lightsensitivitas.....	107
Gambar 3.2	Contoh interaktif dc.js di situs web resminya	111
Gambar 3.3	Memulai server HTTP Python sederhana.....	114
Gambar 3.4	Input tabel obat yang ditampilkan di browser: lima baris pertama	120
Gambar 3.5	Data yang difilter berdasarkan nama obat Grazax 75.000 SQ-T	121
Gambar 3.6	Data yang disaring berdasarkan nama obat Grazax 75.000 SQ-T dan diurutkan berdasarkan hari	122
Gambar 3. 7	Tabel MapReduced dengan obat sebagai kelompok dan jumlah baris data sebagai nilai.....	123

Gambar 3.8	Tabel MapReduced dengan stok rata-rata per obat.....	127
Gambar 3.9	Grafik dc.js: jumlah stok obat selama tahun 2015.....	131
Gambar 3.10	interaksi bagan garis dan bagan baris dc.js	132
Gambar 3.11	dasbor dc.js yang sepenuhnya interaktif untuk obat-obatan dan stoknya di apotek rumah sakit	134

DAFTAR TABEL

Tabel 1-1	Masalah klasifikasi: Dapatkah situs web dipercaya atau tidak?.....	39
Tabel 1-2	Contoh penghitungan jarak hamming	43
Tabel 1-3	Menggabungkan informasi dari kolom yang berbeda ke dalam kolom film. Ini juga merupakan cara kerja DNA bekerja: semua informasi dalam sebuah string yang panjang.....	44
Tabel 1-4	Kutipan dari basis data klien dan film yang disewa pelanggan	46
Tabel 1-5	Pelanggan yang paling mirip dengan pelanggan 27	55
Tabel 1-6	Film dari pelanggan 2 dapat digunakan sebagai saran untuk pelanggan 27	56
Tabel 2-1	Listing perintah sistem file Hadoop yang umum	75

BAB 1 | MENANGANI DATA BESAR PADA SATU KOMPUTER

Bab ini mencakup

- Bekerja dengan kumpulan data yang besar pada satu komputer
- Bekerja dengan *library* Python yang sesuai untuk kumpulan data yang lebih besar
- Memahami pentingnya memilih algoritma dan struktur data yang tepat
- Memahami bagaimana Anda dapat mengadaptasi algoritma untuk bekerja di dalam basis data

Bagaimana jika Anda memiliki begitu banyak data yang tampaknya melebihi kemampuan Anda, dan teknik Anda tampaknya tidak lagi memadai? Apa yang akan Anda lakukan, menyerah atau beradaptasi?

Untungnya Anda memilih untuk beradaptasi, karena Anda masih membaca. Bab ini memperkenalkan Anda pada teknik dan *tool* untuk menangani kumpulan data yang lebih besar yang masih dapat dikelola oleh satu komputer jika Anda menggunakan teknik yang tepat.

BAB

2

LANGKAH PERTAMA DALAM BIG DATA

Bab ini mencakup

- Mengambil langkah pertama dengan dua aplikasi big data: Hadoop dan Spark
- Menggunakan Python untuk menulis pekerjaan big data
- Membangun dasbor interaktif yang terhubung ke data yang disimpan dalam database big data

Selama dua bab terakhir, kami terus meningkatkan ukuran data. Pada bab 3 buku seri 1, kita telah bekerja dengan kumpulan data yang dapat masuk ke dalam memori utama komputer. Bab 1 buku seri 2, memperkenalkan teknik untuk menangani kumpulan data yang terlalu besar untuk dimasukkan ke dalam memori tetapi masih dapat diproses pada satu komputer. Pada bab ini Anda akan belajar untuk bekerja dengan teknologi yang dapat menangani data yang begitu besar sehingga satu node (komputer) tidak lagi memadai. Bahkan mungkin tidak muat di seratus komputer. Nah, ini merupakan sebuah tantangan, bukan?

Kita akan tetap sedekat mungkin dengan cara kerja dari bab-bab sebelumnya; fokusnya adalah memberi Anda

BAB

3

VISUALISASI DATA KEPADA END USER

Bab ini mencakup

- Mempertimbangkan opsi untuk visualisasi data bagi pengguna akhir (end user) Anda
- Menyiapkan aplikasi Crossfilter MapReduce dasar
- Membuat dasbor dengan dc.js
- Bekerja dengan alat pengembangan dasbor

Anda akan segera menyadari bahwa bab ini tentu saja berbeda dengan bab lainnya karena fokusnya terletak pada langkah 6 dari proses sains data. Lebih khusus lagi, apa yang ingin kita lakukan di sini adalah membuat aplikasi kecil sains data. Oleh karena itu, kita tidak akan mengikuti langkah-langkah proses sains data di sini. Data yang digunakan dalam studi kasus ini hanya sebagian yang nyata tetapi berfungsi sebagai data yang mengalir dari tahap persiapan data atau pemodelan data. Nikmati perjalanannya.

Seringkali, data scientist harus menyampaikan wawasan baru mereka kepada pengguna akhir. Hasilnya dapat dikomunikasi-kan dalam beberapa cara:

DAFTAR PUSTAKA

- Apache Spark. (2023). Apache Spark for Python. <https://spark.apache.org/>
- Cielen , Meysman, Ali. (2016). Introducing Data Science. Manning Publications Co. Shelter Island, New York
- Cielen. (2016). Overview of the data science process. <http://freecontent.manning.com/overview-of-the-data-science-process/>
- Cython. (2023). C-Extensions For Python. <https://cython.org/>
- Mike Bostock. (2017). diagram Sankey. <https://bost.ocks.org/mike/>
- Pydata. (2023). NumExpr: Fast numerical expression evaluator for NumPy. <https://github.com/pydata/numexpr>
- Wikipedia. (2023). Time complexity. https://en.wikipedia.org/wiki/Time_complexity