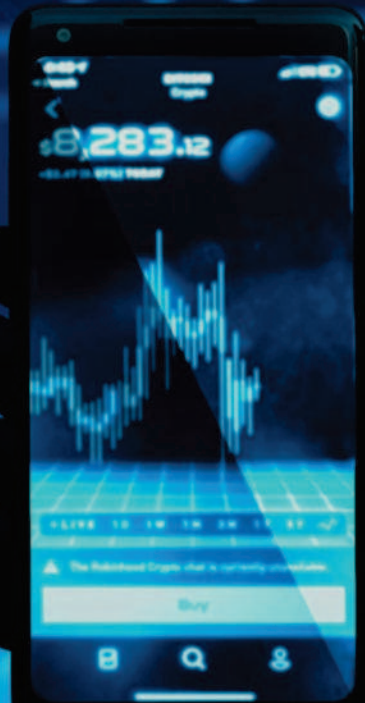


Dr. Ir. Wahyudi,S.T.,M.T.



PENGANTAR SAINS DATA :

Graph Database



eureka
media altara

Anggota IKAPI
No. 225/JTE/2021



0858 5343 1992



eurekamediaaksara@gmail.com



Jl. Banjaran RT.20 RW.10

Bojongsari - Purbalingga 53362

ISBN 978-623-120-664-0



9 786231 206640

PENGANTAR SAINS DATA: GRAPH DATABASE

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



eureka
media aksara

PENERBIT CV.EUREKA MEDIA AKSARA

PENGANTAR SAINS DATA: GRAPH DATABASE

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

Desain Sampul : Ardyan Arya Hayuwaskita

Tata Letak : Husnun Nur Afifah

ISBN : 978-623-120-664-0

Diterbitkan oleh : **EUREKA MEDIA AKSARA, MEI 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

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: Graph Database**". 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. Buku ini merupakan seri ketiga 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, April 2024

Tim Penulis

DAFTAR ISI

KATA PENGANTAR.....	iii
DAFTAR ISI.....	iv
DAFTAR GAMBAR.....	v
BAB 1 PENGANTAR NoSQL.....	1
A. Pengantar ke NoSQL.....	5
B. Studi Kasus: Penyakit Apa Itu?.....	26
C. Ringkasan.....	68
BAB 2 PENGANTAR BASIS DATA GRAF.....	70
A. Berkenalan dengan Data Terhubung dan Basis Data Graf.....	70
B. Memperkenalkan Neo4j: Basis Data Graf	81
C. Contoh Data yang Terhubung: Mesin Rekomendasi Resep	94
D. Ringkasan.....	116

DAFTAR GAMBAR

Gambar 1. 1.	Database NoSQL dan NewSQL.....	4
Gambar 1. 2.	Teorema CAP: ketika mempartisi database Anda, Anda harus memilih antara ketersediaan dan konsistensi	8
Gambar 1. 3.	Teorema CAP: jika node terputus, anda dapat memilih untuk tetap tetap tersedia, tetapi datanya bisa menjadi tidak konsisten	9
Gambar 1. 4.	Teorema CAP: jika node terputus, anda dapat memilih untuk tetap konsisten dengan menghentikan akses ke basis data sampai koneksi dipulihkan kembali	10
Gambar 1. 5.	Sharding: setiap pecahan dapat berfungsi sebagai basis data mandiri, tetapi mereka juga bekerja bersama secara keseluruhan. Contoh ini mewakili dua node, masing-masing berisi empat pecahan: dua pecahan utama dan dua replika. Kegagalan satu node akan dicadangkan oleh node lainnya.....	13
Gambar 1. 6.	ACID versus BASE: basis data relasional tradisional versus sebagian besar basis data NoSQL. Nama-nama tersebut berasal dari konsep kimia skala pH. Nilai pH di bawah 7 adalah asam; lebih tinggi dari 7 adalah basa. Pada skala ini, rata-rata air permukaan berfluktuasi antara 6,5 dan 8,5.....	14
Gambar 1. 7.	Basis data relasional berusaha untuk melakukan normalisasi (memastikan setiap data hanya disimpan sekali). Setiap tabel memiliki pengenal unik (primary key) yang digunakan untuk memodelkan hubungan	

	antara entitas (tabel), oleh karena itu disebut relasional.....	16
Gambar 1. 8.	Tata letak basis data berorientasi baris. Setiap entitas (orang) diwakili oleh satu baris, yang tersebar di beberapa kolom.....	17
Gambar 1. 9.	Pencarian berorientasi baris: dari atas ke bawah dan untuk setiap entri, semua kolom dimasukkan ke dalam memori	18
Gambar 1. 10.	Basis data berorientasi kolom menyimpan setiap kolom secara terpisah dengan baris terkait angka-angka. Setiap entitas (orang) dibagi menjadi beberapa tabel	18
Gambar 1. 11.	Penyimpanan nilai kunci menyimpan segala sesuatu sebagai kunci dan sebuah nilai.....	20
Gambar 1. 12.	Struktur bertingkat nilai kunci.....	21
Gambar 1. 13.	Penyimpanan dokumen menyimpan dokumen secara keseluruhan, sedangkan RDMS memotong artikel dan menyimpannya dalam beberapa tabel. Contoh ini diambil dari situs web Guardian.	23
Gambar 1. 14.	Contoh data graf dengan empat tipe entitas (orang, hobi, perusahaan, dan perabot) dan relasi mereka tanpa edge atau simpul tambahan informasi	25
Gambar 1. 15.	22 database teratas yang diperingkat berdasarkan popularitas menurut DB-Engines.com pada bulan Agustus 2023	26
Gambar 1. 16.	Awan Kata	28
Gambar 1. 17.	Langkah 1 dalam proses sains data: menetapkan tujuan penelitian.....	30

Gambar 1. 18.	Proses sains data langkah 2: pengambilan data. Dalam hal ini tidak ada data internal; semua data akan diambil dari Wikipedia.....	31
Gambar 1. 19.	Proses sains data langkah 3: persiapan data	33
Gambar 1. 20.	Halaman Daftar penyakit di Wikipedia, titik awal pengambilan data Anda.....	35
Gambar 1. 21.	Membuat indeks Elasticsearch dengan Python-Elasticsearch.....	37
Gambar 1. 22.	Tautan pada halaman Wikipedia Daftar penyakit. Ini memiliki lebih banyak tautan lebih banyak dari yang anda perlukan.	39
Gambar 1. 23.	Daftar pertama penyakit di Wikipedia, "daftar penyakit (0-9)"	41
Gambar 1. 24.	Contoh penumpukan URL Elasticsearch.....	43
Gambar 1. 25.	Dokumen penyakit Pemetaan jenis melalui URL Elasticsearch	44
Gambar 1. 26.	Proses sains data langkah 4: eksplorasi data	45
Gambar 1. 27.	Pencarian pertama Lupus dengan 34 hasil.....	48
Gambar 1. 28.	Upaya pencarian kedua lupus dengan enam hasil dan lupus berada di tiga teratas	49
Gambar 1. 29.	Pencarian ketiga lupus: dengan gejala yang cukup untuk menentukan bahwa itu pasti lupus	51
Gambar 1. 30.	Transposisi karakter yang berdekatan adalah salah satu operasi dalam jarak Damerau-Levenshtein. Tiga operasi lainnya adalah penyisipan, penghapusan, dan substitusi.	52
Gambar 1. 31.	Output hit dari kueri yang difilter dengan filter "diabetes" pada nama penyakit	55
Gambar 1. 32.	Agregasi istilah penting diabetes, lima kata kunci pertama	56

Gambar 1. 33.	Proses sains data langkah 3: persiapan data. Pembersihan data untuk teks dapat berupa penyaringan kata; transformasi data dapat berupa pengurangan huruf kecil.	59
Gambar 1. 34.	Filter token shingle untuk menghasilkan bigram	60
Gambar 1. 35.	Penganalisis khusus dengan tokenisasi standar dan filter token shingle untuk menghasilkan bigrams.....	61
Gambar 1. 36.	Awan kata unigram pada kata kunci diabetes yang tidak berbobot dari Elasticsearch	67
Gambar 2. 1.	Contoh data terhubung sederhana: dua entitas atau node (User1, User2), masing-masing dengan properti (nama depan, nama belakang), yang dihubungkan oleh sebuah hubungan (knows).....	72
Gambar 2. 2.	Contoh data terhubung yang lebih rumit di mana dua entitas lain telah disertakan (Negara1 dan Negara2) dan dua hubungan baru ("Has_been_in" dan "Is_born_in").....	73
Gambar 2. 3.	Pada intinya, sebuah graf terdiri dari simpul (juga dikenal sebagai simpul) dan sisi (yang menghubungkan simpul-simpul tersebut), seperti yang diketahui dari definisi matematis graf. Kumpulan objek-objek ini merepresentasikan graf.....	75
Gambar 2. 4.	Gambar ini mengilustrasikan posisi basis data graf pada ruang dua dimensi di mana satu dimensi mewakili ukuran data yang ditangani, dan dimensi lainnya mewakili kompleksitas dalam hal keterhubungan data. Ketika database relasional tidak dapat lagi	

	menangani kompleksitas kumpulan data karena keterhubungannya, tetapi bukan karena ukurannya, maka database graf dapat menjadi pilihan terbaik.	77
Gambar 2. 5.	Pencarian rekursif versi 1: semua data dalam satu tabel.....	80
Gambar 2. 6.	Pencarian rekursif versi 2: menggunakan tabel relasi orang tua-anak.....	81
Gambar 2. 7.	Antarmuka dengan kueri yang telah diselesaikan dari studi kasus bab ini.....	84
Gambar 2. 8	Sebuah contoh sederhana dari graf sosial sederhana dengan dua user dan satu relasi	85
Gambar 2. 9	Contoh data terhubung yang lebih rumit dengan beberapa node yang saling terhubung dari berbagai kategori	87
Gambar 2. 10	Graf yang digambar pada gambar 2.9 sekarang telah dibuat di antarmuka web Neo4j. Node tidak diwakili oleh labelnya tetapi oleh namanya. Kita dapat menyimpulkan dari graf tersebut bahwa kita kehilangan label Hobby dengan nama Traveling. Alasannya adalah karena kita lupa menyertakan node ini dan hubungan yang sesuai dalam pernyataan create.	90
Gambar 2. 11	Hasil dari pertanyaan 1: Negara mana saja yang pernah dikunjungi Annelies? Kita dapat melihat tiga negara yang pernah dikunjungi Annelies dengan menggunakan presentasi baris dari Neo4j. Penjelajahan tersebut hanya membutuhkan waktu 97 milidetik.....	91
Gambar 2. 12	Yang telah di mana? Pembuatan kueri dijelaskan.....	92

Gambar 2. 13	Hasil dari pertanyaan 2: Siapa yang pernah ke mana? Hasil dari penjelajahan kita sekarang ditampilkan dalam representasi graf Neo4j. Sekarang kita dapat melihat bahwa Paul, selain Annelies, juga pernah berkunjung ke Kamboja.	92
Gambar 2. 14	Gambaran proses sains data yang diterapkan pada model rekomendasi data terkoneksi	96
Gambar 2. 15	10 bahan teratas yang paling banyak muncul dalam resep.....	106
Gambar 2. 16	10 hidangan teratas yang dapat dibuat dengan keragaman bahan terbaik	107
Gambar 2. 17	Bahan pembuat Spaghetti Bolognese	108
Gambar 2. 18	Seorang user yaitu Ragnar menyukai beberapa hidangan.....	111
Gambar 2. 19	Keluaran rekomendasi resep; 20 hidangan teratas yang mungkin disukai pengguna.....	113
Gambar 2. 20	Saling keterkaitan antara hidangan pilihan pengguna dan 10 hidangan teratas yang direkomendasikan melalui sub-pilihan bahan yang tumpang tindih	115

BAB

1

PENGANTAR

NoSQL

Bab ini mencakup:

1. Memahami database NoSQL dan mengapa database ini digunakan saat ini
2. Mengidentifikasi perbedaan antara NoSQL dan database relasional
3. Mendefinisikan prinsip ACID dan bagaimana hubungannya dengan prinsip BASE NoSQL
4. Mempelajari mengapa teorema CAP penting untuk pengaturan basis data multi-node
5. Menerapkan proses sains data ke dalam proyek dengan basis data NoSQL Elasticsearch

Bab ini dibagi menjadi dua bagian: bagian awal yang bersifat teoritis dan bagian akhir yang bersifat praktis.

1. Pada bagian pertama bab ini, kita akan membahas basis data NoSQL secara umum dan menjawab pertanyaan-pertanyaan berikut: Mengapa mereka ada? Mengapa tidak sampai saat ini? Jenis apa saja yang ada dan mengapa Anda harus peduli?
2. Pada bagian kedua, kita akan membahas masalah nyata-diagnosa penyakit dan pembuatan profil-menggunakan data yang tersedia secara gratis, Python, dan database NoSQL.

BAB 2

PENGANTAR BASIS DATA GRAF

Di satu sisi kita memproduksi data dalam skala besar, mendorong orang-orang seperti Google, Amazon, dan Facebook untuk menemukan cara-cara cerdas untuk mengatasinya, di sisi lain kita dihadapkan pada data yang semakin saling terhubung. Graf dan jaringan telah merasuk ke dalam kehidupan kita. Dengan menyajikan beberapa contoh yang memotivasi, kami berharap dapat mengajarkan pembaca bagaimana mengenali masalah graf ketika masalah tersebut muncul dengan sendirinya. Dalam bab ini kita akan melihat bagaimana memanfaatkan koneksi-koneksi tersebut dengan menggunakan basis data graf, dan mendemonstrasikan bagaimana menggunakan Neo4j, sebuah basis data graf yang populer.

A. Berkenalan dengan Data Terhubung dan Basis Data Graf

Mari kita mulai dengan membiasakan diri dengan konsep data terhubung dan representasinya sebagai data graf.

1. Data terhubung-Seperti namanya, data terhubung dicirikan oleh fakta bahwa data yang ada memiliki hubungan yang membuatnya terhubung.