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

PENGANTAR SAINS DATA

big data dan visualisasi 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.

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 978-623-151-504-9
9 786231 515049

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

KATA PENGANTAR

Alhamdulillahirobbil'alamin. Penulis bersyukur kehadirat 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

DAFTAR GAMBAR

DAFTAR TABEL

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/

Cyton. (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