



JURUSAN TEKNOLOGI INFORMASI

Mata Kuliah Big Data
01. Pengantar Big Data



Topik

- Sumber Data
- Big Data
- 3V
 - Volume
 - Variety
 - Velocity



Topik-1: Sumber Data

1. Sumber Data

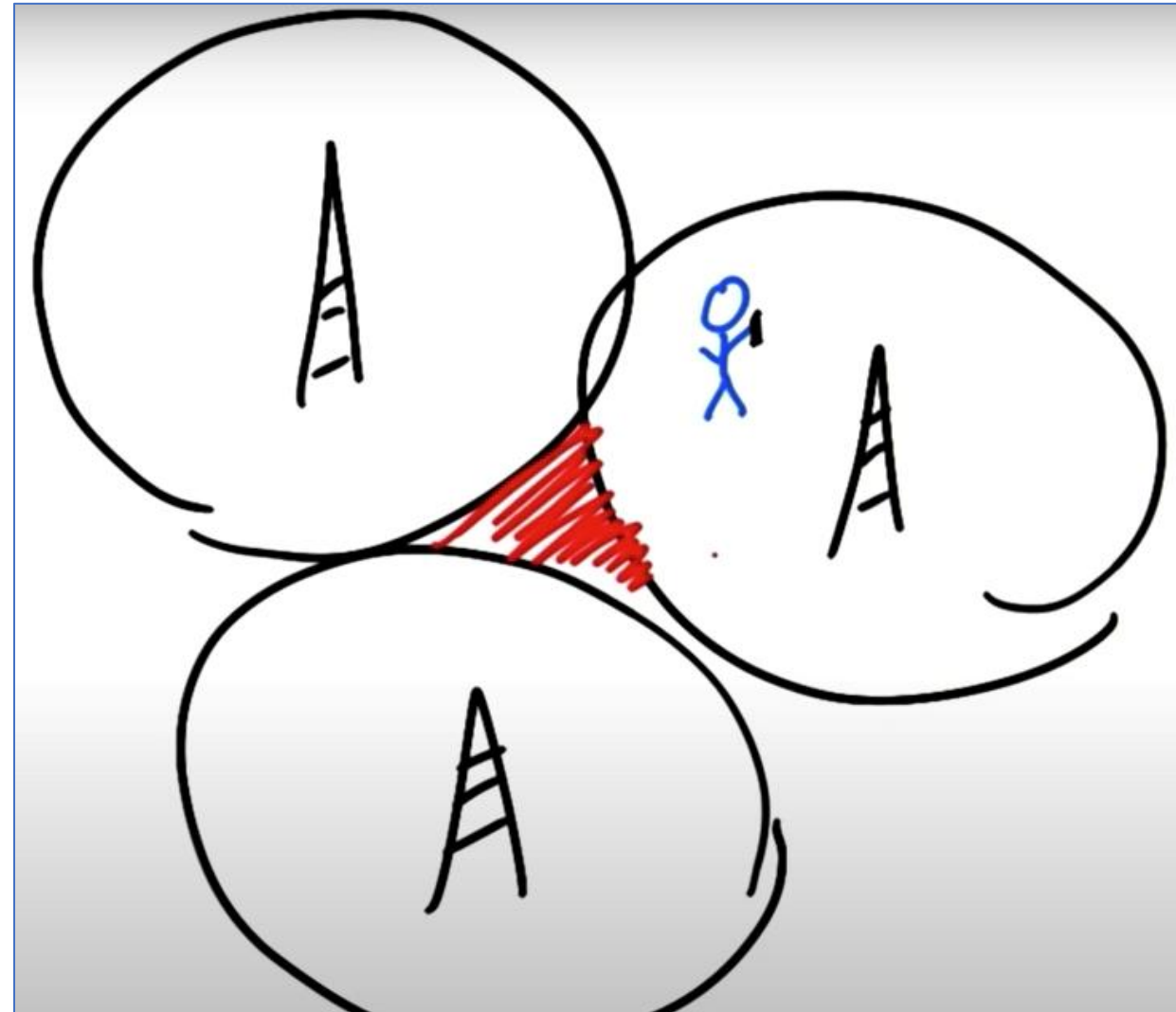


- Perusahaan dan organisasi telah menghasilkan data sejak lama.
- Namun dalam beberapa tahun belakangan, besar data yang dihasilkan meningkat secara eksponensial.
- IBM memperkirakan 90% data yang ada di dunia ini dihasilkan dalam waktu 2 tahun kebelakang!
- Data yang dihasilkan bisa berupa apa saja:
 - Data medis
 - Data retail
 - Data telekomunikasi
 - Data social media
 - Dan data-data yang lainnya, masih banyak lagi..

1. Sumber Data

Contoh-1

- Pada kasus telekomunikasi misalnya, ada sangat banyak data yang di-generate **setiap detik** oleh **setiap operator**!
- Ketika anda mengaktifkan ponsel, maka ponsel Anda akan terhubung ke suatu tower.
- Ketika Anda berpindah tempat/area, maka segala informasi terkait perpindahan tersebut akan dicatat dalam bentuk *log* oleh server telco/operator.
- Data tersebut bisa digunakan untuk berbagai macam keperluan:
 - Menganalisis area tidak tersentuh jangkauan sinyal
 - Melihat tower mana yang paling sibuk.
 - Melacak Anda berada di mana ketika melakukan panggilan darurat.
 - Dll.



1. Sumber Data

Contoh-2

- Pada situs layanan *streaming* seperti Netflix atau Amazon, data kunjungan Anda akan selalu dicatat.
 - Data tersebut memuat banyak hal yang disimpan dalam bentuk *log* dan sangat banyak jumlahnya

```
10.113.178.216 - - [03/Dec/2011:13:04:58 -0800] "GET /assets/img/home-logo.png HTTP/1.1" 200 3892
10.113.178.216 - - [03/Dec/2011:13:04:58 -0800] "GET /images/filmpics/0000/5695/THE_DUEL_-_PACKSHOT_3D_thumb.jpg HTTP/1.1" 200 3602
10.113.178.216 - - [03/Dec/2011:13:04:58 -0800] "GET /images/clientlogos/0000/0042/Chelsea_Films_Logo.jpg HTTP/1.1" 200 59191
10.113.178.216 - - [03/Dec/2011:13:04:58 -0800] "GET /images/filmpics/0000/5693/THE_DUEL_-_PACKSHOT_2D_thumb.jpg HTTP/1.1" 200 5150
10.113.178.216 - - [03/Dec/2011:13:06:11 -0800] "GET /assets/css/printstyles.css HTTP/1.1" 200 778954
10.113.178.216 - - [03/Dec/2011:13:06:11 -0800] "GET /images/filmpics/0000/1421/RagingPhoenix_2DSleeve.jpeg HTTP/1.1" 200 778954
10.113.178.216 - - [03/Dec/2011:13:06:11 -0800] "GET /images/filmpics/0000/2537/14blades_BD_2D.jpg HTTP/1.1" 200 144
10.113.178.216 - - [03/Dec/2011:13:06:11 -0800] "GET /downloadSingle.php?id=6475&fid=680 HTTP/1.1" 200 331
10.113.178.216 - - [03/Dec/2011:13:06:11 -0800] "GET /release-schedule/index.php?o=a&r=a&l=Go HTTP/1.1" 200 4599
10.113.178.216 - - [03/Dec/2011:13:11:26 -0800] "GET /images/filmpics/0000/1421/RagingPhoenix_2DSleeve.jpeg HTTP/1.1" 200 778954
10.113.178.216 - - [03/Dec/2011:13:11:26 -0800] "GET /downloadSingle.php?id=7083&fid=712 HTTP/1.1" 200 300982
10.113.178.216 - - [03/Dec/2011:13:12:58 -0800] "GET /images/filmediablock/618/16.jpg HTTP/1.1" 200 6990
10.113.178.216 - - [03/Dec/2011:13:12:58 -0800] "GET /displaytitle.php?id=101 HTTP/1.1" 200 4460
10.113.178.216 - - [03/Dec/2011:13:12:58 -0800] "GET /assets/css/printstyles.css HTTP/1.1" 200 540
10.113.178.216 - - [03/Dec/2011:13:12:58 -0800] "GET /assets/css/combined.css HTTP/1.1" 200 6112
10.113.178.216 - - [03/Dec/2011:13:12:59 -0800] "GET /assets/js/javascript_combined.js HTTP/1.1" 200 20404
10.113.178.216 - - [03/Dec/2011:13:12:59 -0800] "GET /assets/img/home-logo.png HTTP/1.1" 200 3892
10.113.178.216 - - [03/Dec/2011:13:12:58 -0800] "GET /images/filmpics/0000/3695/Pelican_Blood_2D_Pack.jpg HTTP/1.1" 200 444923
10.113.178.216 - - [03/Dec/2011:13:13:00 -0800] "GET /images/filmediablock/481/swpb_988.jpg HTTP/1.1" 200 67218
10.113.178.216 - - [03/Dec/2011:13:12:59 -0800] "GET /images/filmediablock/481/pb-0622.jpg HTTP/1.1" 200 132304
10.113.178.216 - - [03/Dec/2011:13:13:00 -0800] "GET /images/filmpics/0000/3999/pb-0622_thumb.jpg HTTP/1.1" 200 61483
```

Dari mana Anda

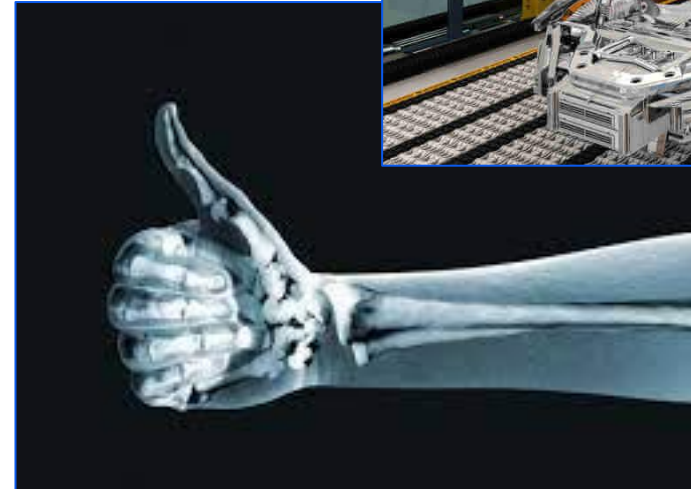
Berapa lama
Anda melihat
halaman tersebut

Halaman mana
yang Anda lihat

1. Sumber Data

Contoh-contoh Lain

- Data telepon dan website tadi hanyalah sedikit contoh dari banyak jenis data-data yang lain seperti:
- Data medis
 - Foto rongent/sinar-X
 - Rekam medis seluruh pasien pada suatu rumah sakit
 - Data BPJS/Asuransi
- Data riset
 - Data *similarities* pada penelitian tentang tumor.
 - Data telemetri pada eksperimen akselerator partikel.
 - Foto-foto dari teleskop luar angkasa.
- Data industri
 - Log mesin/robot/sistem/instrumen.
 - Citra satelit.
 - Log transportasi dan pengiriman barang.
 - Data sensor/IoT.



1. Sumber Data Permasalahan



- Dengan tersedianya data dalam jumlah yang sangat besar tersebut kemudian muncul pertanyaan:
- 1. Bagaimana **menyimpannya**?
- 2. Bagaimana **memprosesnya**?

Topik-2: Big Data

2. Big Data

- Dari contoh-contoh yang telah dikemukakan sebelumnya, banyak sekali kasus yang memang merupakan permasalahan yang membutuhkan teknologi Big Data.
- Namun ingat: **tidak semua** data.
 - Banyak juga kasus yang bisa diselesaikan dengan metode penyimpanan dan pemrosesan konvensional (basis data biasa).
- Sebelum memutuskan untuk membangun dan/atau menggunakan teknologi Big Data, pertimbangkan: **Apakah Anda memiliki data yang memang besar (*big*)?**
 - Atau jangan-jangan data yang Anda miliki belum masuk ke kategori "*big*"?
- Lalu seberapa besar, atau apa tolok ukur dari data yang bisa dikatakan "*big*"?

2. Big Data Contoh



- Manakah di bawah ini contoh data yang bisa dikatakan “big”?
- A. Data detail pembelian pada suatu toko Indomaret selama setahun.
- B. Semua data pemesanan pada seluruh toko Indomaret di Jawa Timur.
- C. Portofolio saham seorang trader.
- D. Data seluruh saham di Bursa Efek Jakarta dalam satu tahun terakhir.

2. Big Data Contoh



- Manakah di bawah ini contoh data yang bisa dikatakan “*big*”?
- A. Data detail pembelian pada suatu toko Indomaret selama setahun.
- **B. Semua data pemesanan pada seluruh toko Indomaret di Jawa Timur.**
- C. Portofolio saham seorang trader.
- **D. Data seluruh saham di Bursa Efek Jakarta dalam satu tahun terakhir.**

2. Big Data Definisi



- Definisi dari “*big data*” bisa sangat subjektif.
 - Sebagian menganggap data berukuran terabyte ke atas adalah “*big*”.
 - Namun banyak juga orang yang mengolah data yang ukurannya lebih kecil dengan menggunakan teknologi big data dengan hasil yang sangat memuaskan.
- Lalu apa definisi dari “*big*” data tersebut?
- Cloudera (salah satu perusahaan terkemuka dibidang Big Data) mendefinisikan: “**Suatu data dapat dikatakan sebagai *big data* apabila terlalu besar untuk bisa disimpan dan diolah dalam satu mesin (komputer).**”
- Dengan kata lain, terdapat **tantangan** yang menyebabkan data sulit untuk disimpan dan diolah dalam satu komputer.
- Tantangan apa saja itu?

2. Big Data Tantangan



- Manakah di bawah ini kira-kira, yang merupakan tantangan yang melatarbelakangi adanya teknologi Big Data?
- A. Dari sekian banyak data, kebanyakan tidak berguna.
- B. Data di-*generate* dengan sangat cepat.
- C. Data datang dari berbagai sumber dengan berbagai macam format/bentuk.

2. Big Data Tantangan



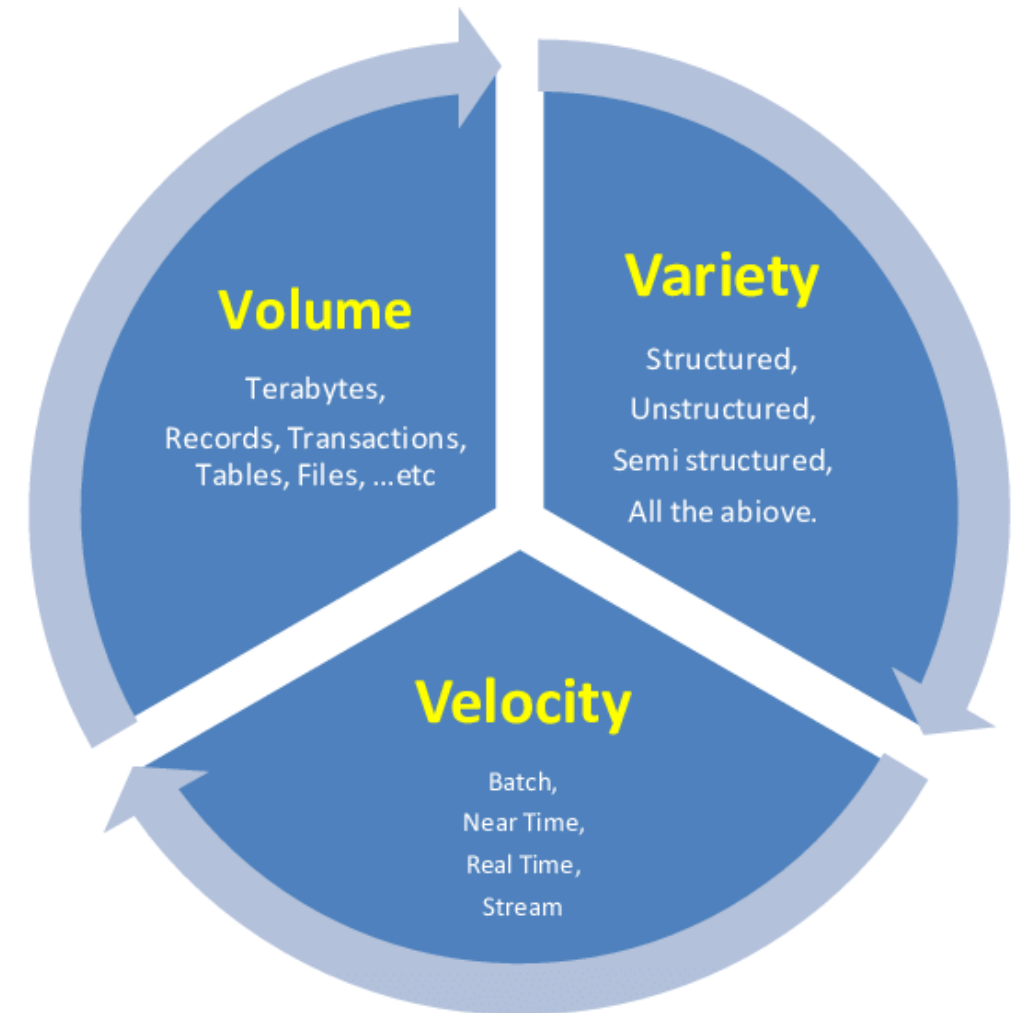
- Manakah di bawah ini kira-kira, yang merupakan tantangan yang melatarbelakangi adanya teknologi Big Data?
- A. Dari sekian banyak data, kebanyakan tidak berguna.
- **B. Data di-generate dengan sangat cepat.**
- **C. Data datang dari berbagai sumber dengan berbagai macam format/bentuk.**

Topik-3: 3V



3. 3V

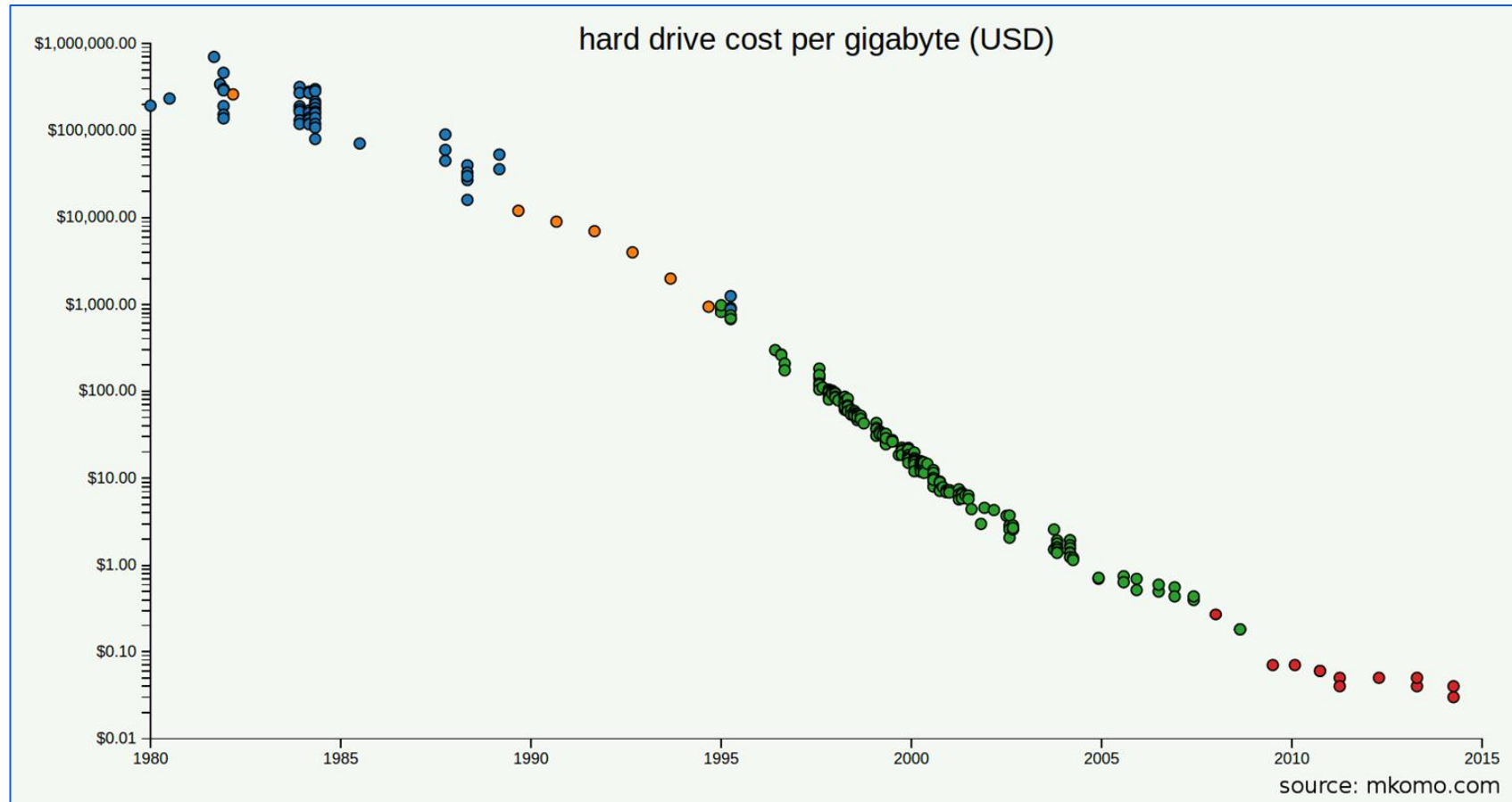
- Ketika membicarakan tentang Big Data, seringkali dibahas mengenai 3V yang merupakan tantangan Big Data.
- Ketiga “V” tersebut adalah:
 - *Volume* (Ukuran)
 - *Variety* (Keberagaman)
 - *Velocity* (Kecepatan)
- **Volume**
 - Data yang dihasilkan dalam ukuran yang sangat besar.
- **Veriety**
 - Data yang datang dari berbagai macam sumber dan beragam bentuk/format.
- **Velocity**
 - Data yang dihasilkan dalam tempo yang sangat cepat.



3. 3V Volume



- Biaya yang diperlukan untuk menyimpan data telah turun drastis sejak tahun 60an..



Sumber: <https://aiimpacts.org/costs-of-information-storage/>

3. 3V Volume

- Harga penyimpanan per GB:
 - 1980an → US \$ 100.000+ (Rp. 1,4 Milyar)
 - 2013 → US \$ 0,10 (Rp. 1.400an)
- Namun itu baru harga penyimpanannya saja. Untuk bisa menyimpan data dengan baik (*reliable*) dibutuhkan biaya tambahan:
 - Setidaknya untuk membeli PC (bagi pengguna rumahan)
 - Atau membeli SAN (Storage Area Network – bagi pengguna sekelas perusahaan)
- Harga SAN yang mahal membatasi jumlah data yang dapat disimpan oleh instansi/perusahaan. Akibatnya:
 - Hanya data penting dan kritis saja yang disimpan seperti data penjualan aktual.
- Namun belakangan diketahui bahwasannya data yang banyak tersebut, yang terlihat tidak berguna, ternyata dapat mendatangkan keuntungan tambahan yang besar!



tokopedia

Kategori

Cari dompet



Masuk

Daftar

Jaket P... Ipad Ai... Lampu Ta... Iphone... Face Shi... Kandang Kuc...



Synology Ds3617xs - Nas Server 12 Bays

HARGA

Rp 56.249.999

DESKRIPSI

Selamat Datang Ditoko Kami, Selamat Berbelanja...

DiskStation DS3617xs, Centralize data right on your desk

Meet the 12-bay desktop NAS that allows instant deployment with scalability up to 36 drives, delivering outstanding 2,358MB/s sequential throughput reading. DS3617xs is ... [Lihat Selengkapnya](#)

3. 3V Volume

- Menyimpannya saja sudah merupakan masalah tersendiri.
 - Data sangat banyak, padahal SAN mahal. Bagaimana jika penuh?
- Belum lagi: **memprosesnya**.
 - *Streaming* data berukuran terabita dari SAN melewati jaringan, dan mengumpulkannya pada satu tempat pemrosesan akan memakan waktu yang sangat (sangat) lama.
- Pada kasus seperti ini, terlihatlah kelebihan teknologi Big Data yang mampu:
 - Menyimpan data bervolume besar, dengan baik namun dengan harga yang lebih murah.
 - Membaca dan memproses bervolume besar dengan efisien dan cepat.
- Bagaimana bila storage penuh?
 - Cukup tambahkan komputer baru sebagai **data node** dengan harga yang tidak begitu mahal (*commodity hardware*).
- Bagaimana memprosesnya?
 - Big data mampu memproses data secara parallel dengan tanpa memindahkan data ke satu pusat pemrosesan, melalui teknik **MapReduce**.

3. 3V Variety



- Metode penyimpanan data konvensional sangat populer digunakan.
 - Basis data biasa: MySQL, SQL Server
 - Data warehouse: Oracle, IBM
- Namun terdapat kelemahan utama yaitu: Data harus disusun sedemikian rupa agar bisa cocok dimasukkan kedalam **struktur** tabel yang telah dirancang sebelumnya (*predefined*).
- Banyak data pada zaman sekarang yang bentuknya tidak terstruktur dan berbagai macam bentuk serta formatnya.
 - **Unstructured** data → Hasil scan, foto, dokumen, suara, video, dll.
 - **Semi-structured** data → Email, log, halaman web, XML, CSV, TSV, paket-paket TCP/IP, dlsb.
- Data semacam itu sangat sulit untuk disimpan dan direkonsiliasikan pada system penyimpanan data konvensional.

3. 3V Variety – Format Data

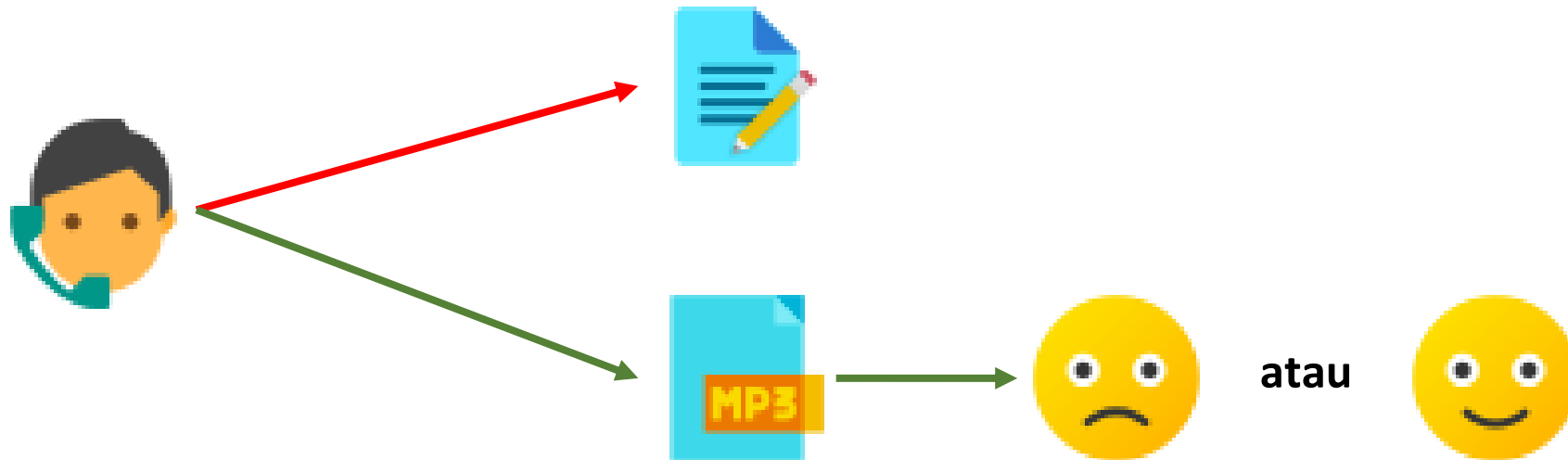
- Bank memiliki begitu banyak data seperti: Daftar transaksi pada kartu kredit maupu kredit Anda, scan dari cek, catatan sesi dengan *customer service* dan bahkan rekaman telepon dengannya.
 - Padahal data-data tersebut perlu disimpan dalam **format aslinya**.



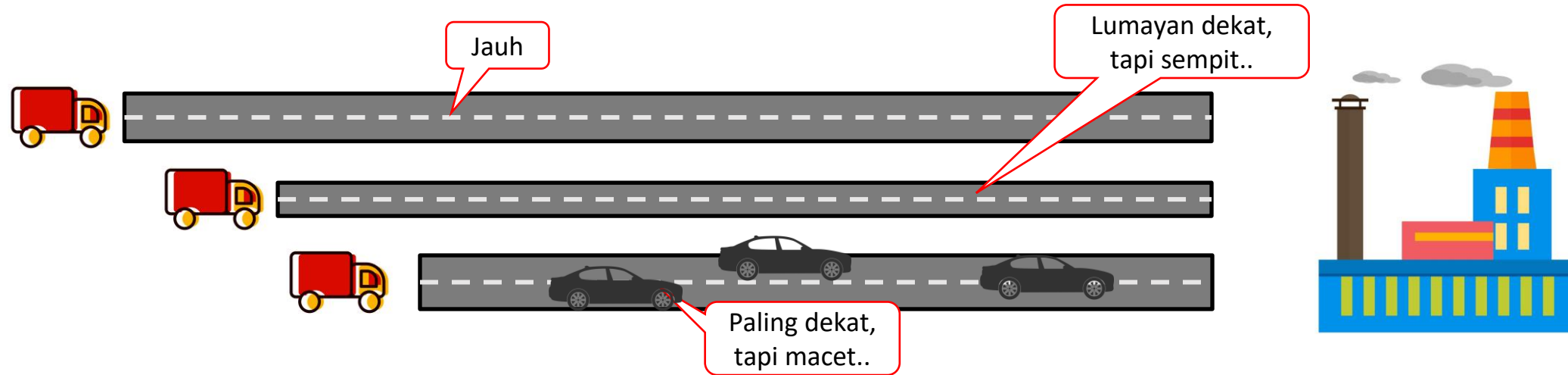
3. 3V

Variety – Mengapa menyimpan format data asli?

- Misalkan pada kasus rekaman sesi antara nasabah dengan customer service yang disimpan untuk menganalisis kepuasan pelanggan:
 - Mengapa tidak dikonversikan saja suaranya ke teks agar bisa menghemat penyimpanan?



3. 3V Variety – Studi Kasus



- Sistem koordinasi logistik adalah sistem yang lazim digunakan di industry.
 - Sistem ini mengarahkan kendaraan-kendaraan transport pabrik dengan **muatan** tertentu ke suatu **tujuan** tertentu melalui **route** tertentu.
- Ketika pabrik membutuhkan sebuah truk untuk membawa muatan, maka sistem konvensional akan memerintahkan kendaraan yang terdekat untuk segera Kembali ke pabrik.
- Namun, terdekat tidak selalu berarti terbaik.
 - Bisa saja walaupun dekat, tapi jalannya macet atau sempit.
 - Bisa saja dekat, tapi muatannya sedang penuh.

3. 3V

Variety – Studi Kasus

- Sistem konvensional, karena keterbatasan penyimpanan, hanya akan menyimpan data yang dianggap penting saja.
 - Karena itu yang direkomendasikan adalah truk terdekat saja.
- Padahal, data yang cenderung tidak dianggap penting bisa saja justru menjadi pengantar menuju jalan keluar yang lebih optimal.
- Pada kasus logistic tersebut, Anda mengusulkan untuk meng-upgrade sistemnya dengan mengimplementasikan Big Data. Dengan demikian Anda sekarang bisa menyimpan data apa saja.
- Kira-kira dari sekian banyak data berikut ini, manakah yang dapat membantu mengatasi permasalahan logistik tersebut?
 - A. Data GPS saat ini dari semua truk/kendaraan logistik pabrik.
 - B. Data rencana rute setiap truk/kendaraan logistik pabrik untuk hari ini.
 - C. Data trafik kendaraan realtime di jalan-jalan saat ini.
 - D. Data muatan semua truk/kendaraan logistik pabrik saat ini, termasuk volume dan beratnya.
 - E. Data tingkat efisiensi bahan bakar untuk semua truk/kendaraan logistik pabrik.

3. 3V Variety – Studi Kasus

- Kira-kira dari sekian banyak data berikut ini, manakah yang dapat membantu mengatasi permasalahan logistik tersebut?
 - A. Data GPS saat ini dari semua truk/kendaraan logistik pabrik.**
 - B. Data rencana rute setiap truk/kendaraan logistik pabrik untuk hari ini.**
 - C. Data trafik kendaraan realtime di jalan-jalan saat ini.**
 - D. Data muatan semua truk/kendaraan logistik pabrik saat ini, termasuk volume dan beratnya.**
 - E. Data tingkat efisiensi bahan bakar untuk semua truk/kendaraan logistik pabrik.**
- *Semua data dapat membantu kita mencari solusi yang lebih baik.*

3. 3V Velocity



- “V” yang ketiga adalah Velocity → Seberapa cepat data datang dan siap untuk diproses.
- Pada penjelasan sebelumnya: Data yang banyak bisa mendatangkan keuntungan bila disimpan dan diolah lebih lanjut (seperti data telepon tadi).
- Secepat apapun data yang datang, kita harus mampu menerima dan menyimpannya ke dalam storage.
 - Bahkan hingga walaupun *rate*-nya 5 TB/hari, kita harus mampu menyimpannya!
- Jika tidak bisa menyimpan semuanya, maka akhirnya sebagian data tersebut harus terbuang.
- Kita tidak ingin hal itu terjadi, mengingat akan ada potensi keuntungan yang terbuang setiap kali kita membuang data.

3. 3V

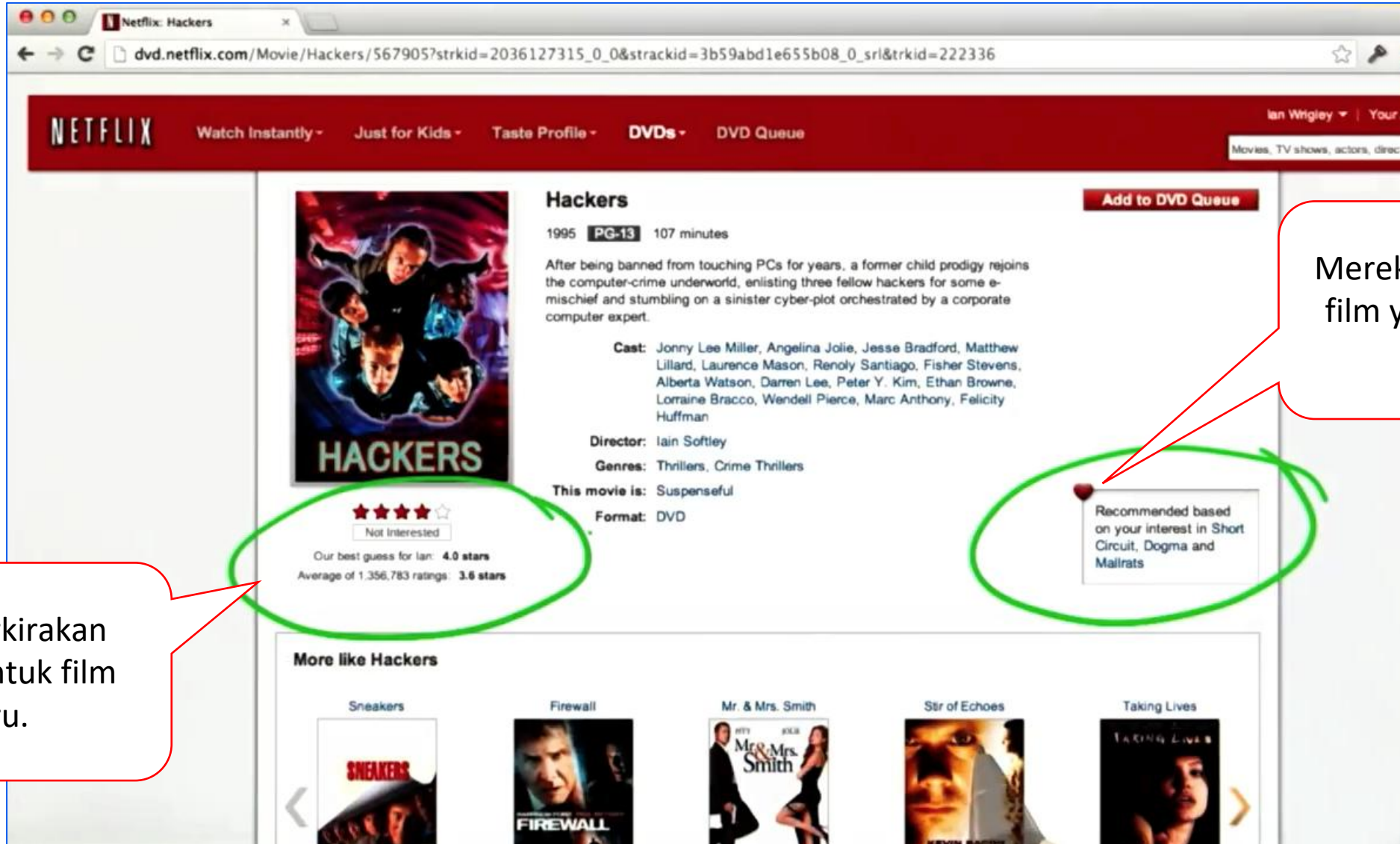
Velocity – Mengapa perlu menyimpan semua data?



- Dari kunjungan seorang pengguna ke suatu website jual beli online:
 - Jika ada data produk yang dilihat sebelumnya → Bisa menampilkan rekomendasi produk sejenis di kunjungan berikutnya.
 - Jika kita tahu selama 5 menit seorang pengguna memandangi produk tertentu → Bisa mengirimkan email berisi informasi ketika produk tersebut sedang diskon.
 - Jika ada data device yang digunakan saat *browsing* dan misalkan pengguna tersebut menggunakan tablet iPad model lama → Bisa menyarankan iPad model terbaru.
- Semua contoh diatas dapat meningkatkan pengalaman pengguna dalam berbelanja online yang pada akhirnya berkorelasi langsung dengan **profit!**
 - Tidak akan bisa dicapai jika data yang disimpan hanya riwayat pembelian aktual saja.

3.3V

Velocity – Mengapa perlu menyimpan semua data?



The screenshot shows the Netflix interface for the movie 'Hackers'. The browser address bar displays the URL: `dvd.netflix.com/Movie/Hackers/567905?strkid=2036127315_0_0&strackid=3b59abd1e655b08_0_srl&trkid=222336`. The Netflix navigation bar includes links for 'Watch Instantly', 'Just for Kids', 'Taste Profile', 'DVDs', and 'DVD Queue'. The user's name 'Ian Wrigley' and 'Your Account' link are visible on the right.

The movie 'Hackers' is featured with a poster showing four characters. Below the poster, the title 'Hackers' is displayed, followed by the year '1995', rating 'PG-13', and duration '107 minutes'. A description states: 'After being banned from touching PCs for years, a former child prodigy rejoins the computer-crime underworld, enlisting three fellow hackers for some e-mischief and stumbling on a sinister cyber-plot orchestrated by a corporate computer expert.'

The cast list includes: Jonny Lee Miller, Angelina Jolie, Jesse Bradford, Matthew Lillard, Laurence Mason, Renoly Santiago, Fisher Stevens, Alberta Watson, Darren Lee, Peter Y. Kim, Ethan Browne, Lorraine Bracco, Wendell Pierce, Marc Anthony, Felicity Huffman. The director is 'Iain Softley'. The genres are 'Thrillers, Crime Thrillers'. The format is 'DVD'.

Below the movie details, there is a star rating section. It shows five stars with the first four filled. Below the stars is a 'Not interested' button. The text reads: 'Our best guess for Ian: 4.0 stars' and 'Average of 1,356,783 ratings: 3.6 stars'. A green circle highlights this rating section, with a callout bubble stating: 'Memperkirakan rating, untuk film baru.'

To the right of the movie details, there is a red heart icon and a text box that says: 'Recommended based on your interest in Short Circuit, Dogma and Mailrats'. A green circle highlights this recommendation section, with a callout bubble stating: 'Merekomendasikan film yang mungkin disukai.'

At the bottom, there is a section titled 'More like Hackers' featuring a row of movie posters: 'Sneakers', 'Firewall', 'Mr. & Mrs. Smith', 'Str of Echoes', and 'Taking Lives'.

Pertanyaan?



Terima Kasih

Latihan

- Buatlah laporan yang berisi:
 - Contoh kasus yang membutuhkan teknologi “big” data.
 - Pengertian tentang Hadoop
 - Pengertian HDFS
 - Pengertian MapReduce
- Terdiri dari 1 atau 2 halaman saja (A4, dengan ukuran margin dan *font* standar)
 - Tidak boleh copas.
 - Jika menggunakan AI, harus dipahami terlebih dahulu, lalu tulis ulang intinya dengan menggunakan bahasa Anda sendiri.
- Kumpulkan di Google Classroom/LMS.