Topik 2 BIG DATA

2.1 Fenomena Big Data

1. Karakteristik Big Data

2. Pemanfaatan Big Data

2.2 Siklus Hidup Data

Tahap 1: Define a research question

Tahap 2: Collect & Organize Data

Tahap 3: Wrangle Data

Tahap 4: Explore & Visualize Data

Tahap 5: Analyze & Interpret Data

Tahap 6: Communicate with Data

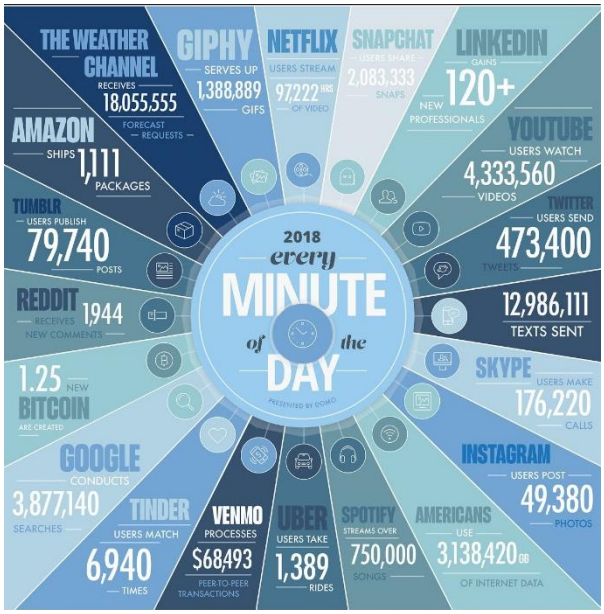
2.1 Fenomena Big Data

Fenomena Big data terjadi dengan berbagai kondisi yang ada sekarang yaitu bahwa

- ✓ Setiap orang dapat menghasilkan data yang jumlahnya sangat besar.
- ✓ Sebelum era web 2.0, penghasil data di internet hanyalah perusahaan, organisasi, atau orang-orang yang memiliki keahlian khusus. Kebanyakan orang hanyalah menjadi pengguna data yang dihasilkan dari web yang sifatnya statik. Era ini disebut sebagai era web 1.0.
- ✓ Pada era web 2.0, pengguna internet dapat berkontribusi pada konten suatu website, dan dapat berinteraksi dan berkolaborasi satu sama lain dengan menggunakan media sosial.
- ✓ Data yang dihasilkan oleh setiap pengguna internet ini sangatlah besar dari berbagai macam aktivitasnya.
- ✓ Datanya pun memiliki berbagai macam bentuk seperti teks, panggilan telpon, email, video, foto, musik, dan sebagainya.

Mari kita lihat infografis berikut yang menunjukkan berapa jumlah data yang dihasilkan di internet dalam setiap menitnya berdasarkan hasil penelitian DOMO, sebuah cloud based operating system.

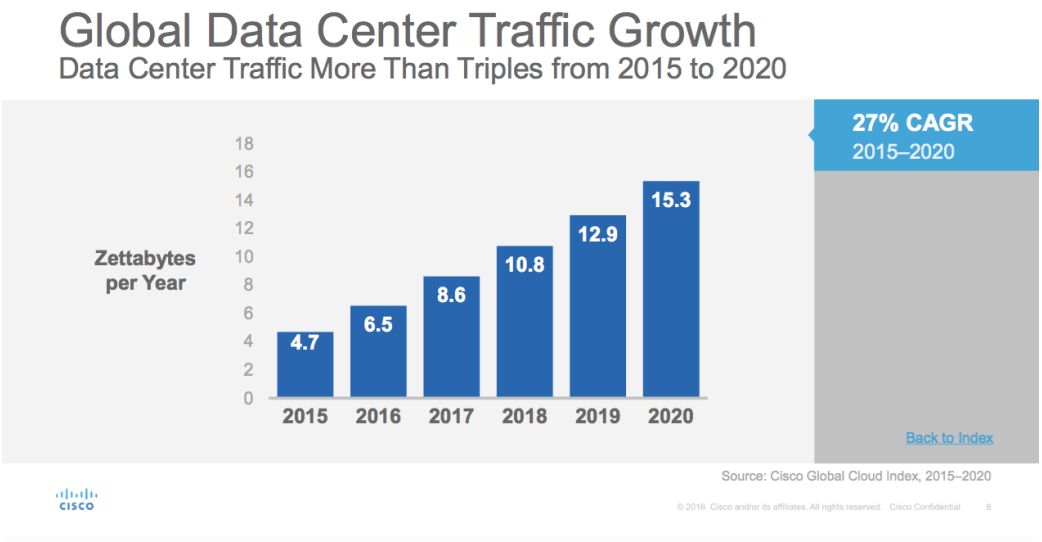
Dari infografis yang menggambarkan kondisi di tahun 2018 ini kita melihat bahwa dalam 1 menit:



- ✓ Pengguna youtube menonton lebih dari 4 juta video
- ✓ Pengguna twitter mengirimkan lebih dari 473 ribu tweets
- ✓ Pengguna Instagram mem-posting lebih dari 49 ribu foto
- ✓ Spotify memutarkan lebih dari 750 ribu lagu
- ✓ 3,8 juta pencarian dilakukan melalui Google
- ✓ Pengguna Netflix memutar lebih dari 97ribu jam video
- ✓ Hampir 13 juta teks dikirimkan

Hingga pada tahun 2020 diperkirakan setiap orang di bumi menghasilkan 1,7 megabyte data per detik.

Selanjutnya, berdasarkan data dari cisco global cloud index, didapatkan informasi betapa pertumbuhan traffic data internet terjadi sedemikian massive.



Traffic data pada tahun 2020 tiga kali lipat lebih besar dibandingkan dengan traffic data pada tahun 2015, yaitu sebesar 15,3 zettabytes pertahun.

Seberapa besarkah 1 zettabyte itu?

Dapat dilihat pada gambar bahwa 1 zettabyte adalah sebesar 1000 pangkat 7 bytes atau 10 pangkat 21 bytes.

Unit	Value	Size
bit (b)	0 or 1	1/8 of a byte
byte (B)	8 bits	1 byte
kilobyte (KB)	1000^1 bytes	1,000 bytes
megabyte (MB)	1000^2 bytes	1,000,000 bytes
gigabyte (GB)	1000^3 bytes	1,000,000,000 bytes
terabyte (TB)	1000^4 bytes	1,000,000,000,000 bytes
petabyte (PB)	1000^5 bytes	1,000,000,000,000,000 bytes
exabyte (EB)	1000^6 bytes	1,000,000,000,000,000,000 bytes
zettabyte (ZB)	1000^7 bytes	1,000,000,000,000,000,000,000 bytes
yottabyte (YB)	1000^8 bytes	1,000,000,000,000,000,000,000,000 bytes

Kalau kita memiliki laptop dengan kapasitas 1 terabyte, berarti dibutuhkan 1 milyar laptop untuk menyimpan 1 zettabytes.

Sungguh angka yang sangat besar bukan?

Jumlah data yang massive inilah yang sering disebut sebagai **BIG DATA**

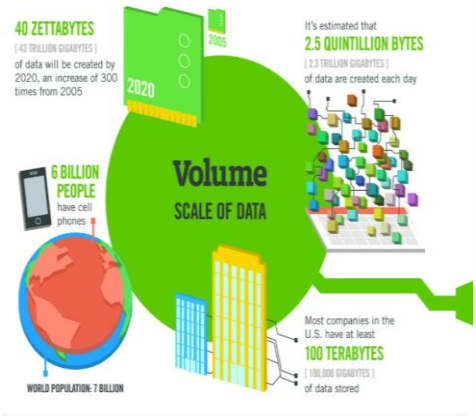
1. Karakteristik Big Data

Big data merupakan istilah yang seringkali digunakan untuk data yang jumlahnya sangat besar dan juga kompleks, yang pengelolaan dan pemrosesannya tidak dapat dilakukan oleh

system atau teknik manajemen data yang konvensional. Bagaimana kita bisa mengenali big data? Ada cara yang mudah untuk mengenali Big Data.

- ✓ Volume, Variety, dan Velocity adalah tiga dimensi utama yang mengkarakterisasi big data ditambah satu tambahan terakhir yaitu Veracity
- ✓ Ada satu kalimat untuk mendeskripsikan BigData dengan tiga karakter tersebut yaitu: data dengan jumlah yang besar, yang memiliki berbagai macam bentuk, dan harus diproses dengan cepat.

a. Volume



Volume mengacu pada besarnya data yang dihasilkan setiap detik, menit, jam, dan hari di dunia digital kita.

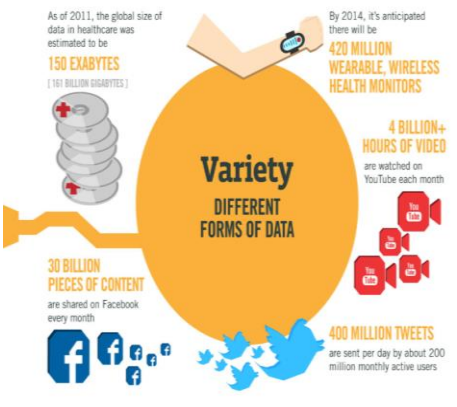
Data yang dimaksud tidak hanya dihasilkan oleh orang, tetapi juga mesin dan benda-benda lainnya.

Big data merepresentasikan data yang besar. Jika dulu data dihasilkan oleh karyawan maka saat ini data diciptakan oleh mesin yang merekam aktifitas manusia dari transaksi perbankan hingga aktifitas di media sosial.

Karena sekarang kita berada pada zamannya IoE atau *Internet of Everything*.

Nilai dari data bisa jadi tidak dapat ditentukan, seperti data twitter, data klik atau clickstream pada halaman website atau aplikasi mobile, ataupun peralatan yang menggunakan sensor. Beberapa organisasi mungkin ada yang mengelola puluhan terabytes. Adapula yang mengelola data hingga ratusan petabytes. Banyak faktor yang mendukung meningkatnya volume data secara pesat, diantaranya adalah hampir semua transaksi bisnis melibatkan data, meningkatnya jumlah unstructured data yang mengalir dari media sosial, dan meningkatnya jumlah data yang dihasilkan dari mesin serta perangkat mobile.

b. Variety



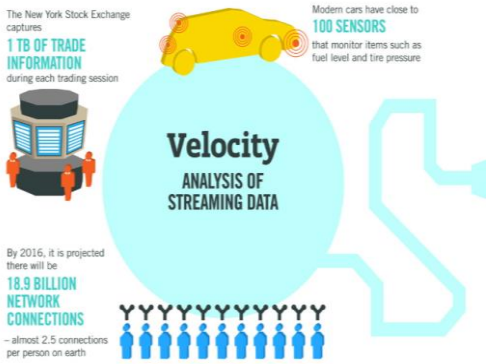
Variety mengacu pada meningkatnya ragam bentuk data seperti teks, gambar, suara, dan data geospasial.

Misal Rumah sakit mengelola berbagai bentuk data, baik data terstruktur, semi terstruktur, atau tidak terstruktur. Data terstruktur yang dikelola bisa jadi berbentuk data excel. Data semi-terstruktur misalnya log files dalam aplikasi manajemen rumah sakit, dan data tidak terstruktur misalnya foto x-ray.

Secara umum ada dua jenis data yaitu *Structured data* dan *Unstructured data*.

- ✓ Structured data adalah data-data yang sudah rapi. Contoh data *structured* adalah transaksi di ATM yang menghasilkan data seperti jumlah uang yang diambil, waktu penarikan dan sebagainya telah tersimpan dalam database pihak Bank.
- ✓ *Unstructured data*, biasanya berupa text document, email, video, audio, dan data dari sosial media dan mungkin juga berisi data seperti tanggal, angka, dan fakta lainnya. Contohnya adalah komentar komentar di media sosial, status di facebook, atau foto foto yang di upload di Instagram.

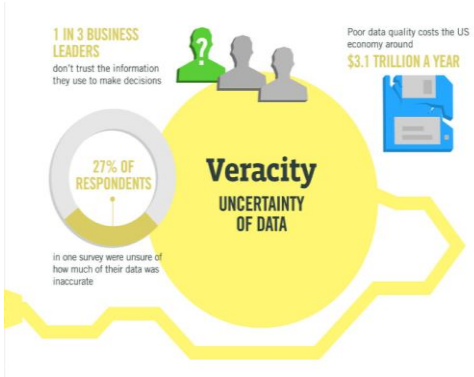
c. Velocity



Velocity mengacu pada cepatnya data yang dihasilkan dan kecepatan data bergerak dari satu titik ke titik yang lain.

Sebuah mobil dengan teknologi terkini memiliki setidaknya 100 sensors yang mampu memonitor banyak hal seperti tingkat bahan bakar ataupun tekanan udara di ban.

d. Veracity



Veracity juga mengacu kepada kualitas data.

Karena data berasal dari begitu macam sumber yang berbeda, sangatlah sulit untuk menghubungkan, mencocokan, membersihkan, dan mentransformasi data lintas system. Organisasi perlu mendefinisikan hubungan antar data tersebut, jika tidak, maka data akan sulit untuk dikendalikan dan dimanfaatkan.

Varicity adalah karakterisitik dalam data yang menyulitkan kita memvalidasi kebenaran data karena semakin luas dan beragam hingga sulit dikontrol. Dengan teknologi Big data maka akan memungkinkan kita untuk bekerja melalui hasil analisis secara cepat dan mudah.

2. Pemanfaatan Big Data

> *Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. (Gartner)*

Big data dapat membantu dalam berbagai macam aktivitas bisnis, diantaranya product development, predictive maintenance, customer experience, machine learning, operational efficiency, dan drive innovation.

a. *product development*

- ✓ *Mengantisipasi kebutuhan customer*
- ✓ *Memprediksi produk dan layanan baru*

Pemanfaatan big data pada product development, diantaranya dilakukan oleh perusahaan seperti Netflix dan Procter & Gamble atau P & G untuk mengantisipasi kebutuhan dari customernya. Mereka membuat model untuk memprediksi produk dan layanan baru dengan mengklasifikasikan atribut penting dari produk dan layanan yang pernah atau sedang disediakan. Mereka lalu membuat relasi antara atribut-atribut tadi dengan kesuksesan komersilnya. Selanjutnya P&G menggunakan data dan analitik dari media social, tes pasar, dan peluncuran produk awal di toko untuk merencanakan, memproduksi dan meluncurkan produk barunya.

b. *Predictive Maintenance*

- ✓ Menghemat biaya pemeliharaan mesin
- ✓ Memaksimalkan waktu kerja komponen dan peralatan

Pemanfaatan big data berikutnya adalah pada predictive maintenance atau untuk memprediksi pemeliharaan peralatan. Faktor-faktor yang bisa memprediksikan kegagalan mekanis bisa jadi tersembunyi dalam data yang terstruktur, misalnya tahun pembuatan, dan model dari peralatan. Ataupun tersembunyi dari data yang tidak terstruktur yang mewadahi jutaan log, data sensor, pesan error, dan suhu mesin. Dengan menganalisis indikasi-indikasi yang berpotensi akan bermasalah sebelum terjadi kerusakan, organisasi dapat melakukan pemeliharaan dengan lebih hemat dan memaksimalkan waktu kerja komponen dan peralatan.

c. *Customer Experience*

- ✓ Menawarkan produk secara personal
- ✓ Meningkatkan loyalitas pengguna
- ✓ Menangani permasalahan secara proaktif

Kompetisi untuk menarik customer saat ini sedang sangat marak. Saat ini perusahaan dapat lebih jelas melihat pengalaman pengguna. Big Data memungkinkan perusahaan untuk mengumpulkan data dari media social, website, panggilan telepon, dan berbagai sumber lainnya untuk meningkatkan pengalaman interaksi penggunanya dan memaksimalkan manfaat yang disampaikan kepada customer. Diantara yang bisa dilakukan adalah dengan menawarkan produk secara personal, meningkatkan hubungan baik dengan customer untuk meningkatkan loyalitas, dan menangani masalah yang muncul secara proaktif.

d. Machine Learning

Machine Learning sedang menjadi topik perbincangan yang hangat sekarang ini. Dan keberadaan data, khususnya big data adalah yang sangat berperan dalam keberadaan machine learning. Kita sekarang bisa melatih mesin untuk belajar dari big data yang tersedia.

e. Operational Efficiency

- ✓ Menekan biaya operasional
- ✓ Mengantisipasi kebutuhan ke depan
- ✓ Meningkatkan pengambilan keputusan yang sejalan dengan kebutuhan pasar

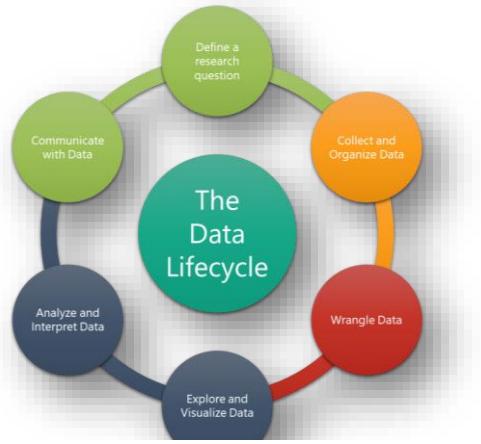
Efisiensi operasional mungkin bukan hal yang baru, tapi di area inilah big data memiliki peranan yang besar. Dengan big data, perusahaan bisa menganalisis dan menilai produksi, umpan balik dari *customer*, dan faktor-faktor lainnya untuk menekan biaya operasional dan mengantisipasi kebutuhan di masa depan.

Big data juga dapat digunakan untuk meningkatkan pengambilan keputusan yang sejalan dengan kebutuhan pasar.

f. Drive Innovation

Big data dapat membantu perusahaan untuk melakukan inovasi dengan mempelajari hubungan saling kebergantungan antar manusia, institusi, entitas, dan proses. Selanjutnya, menentukan cara baru untuk menggunakan insight yang didapat untuk mengembangkan pertimbangan dalam pengambilan keputusan terkait keuangan dan perencanaan.

2.2 Siklus Hidup Data

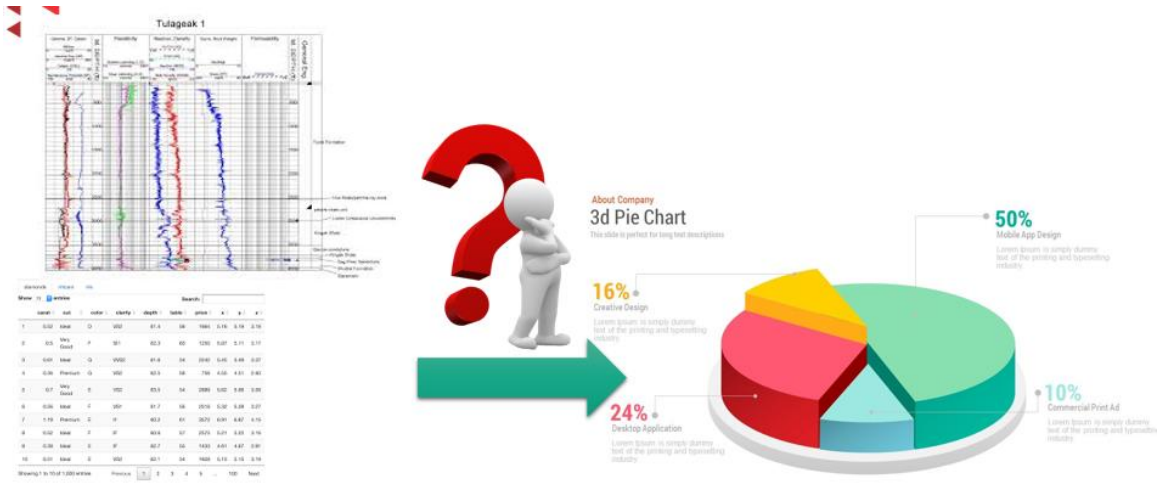


Pada pembahasan sebelumnya kita mengetahui begitu banyak manfaat yang didapat ketika kita mampu memahami data, berkerja dengan data dan berkomunikasi dengan data. Ketika bekerja dengan data kita akan melewati tahap-tahap yang disebut dengan siklus hidup data. Namun, sebelum mempelajari bagaimana siklus hidup data, mari kita simak dulu ilustrasinya

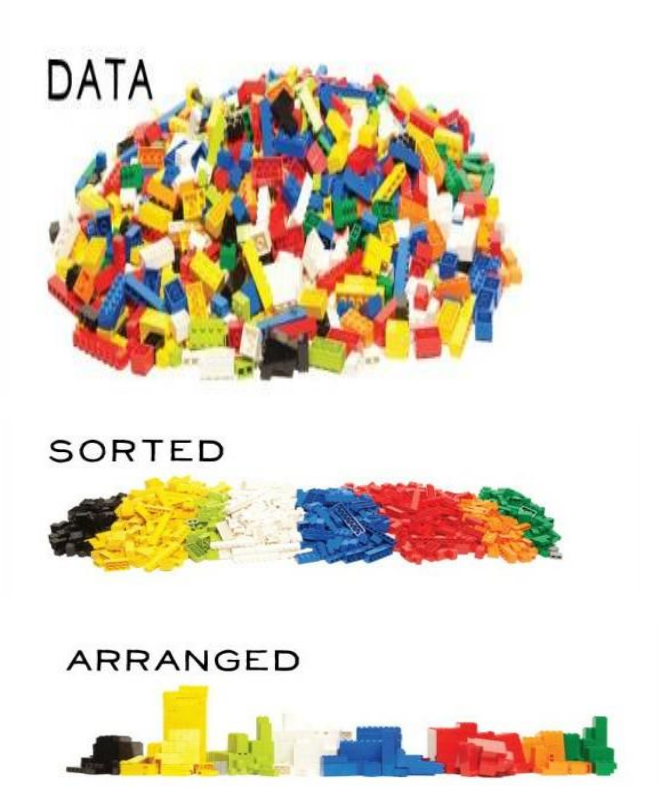
Berkerja dengan data seperti proses membuat kue. Sebuah kue tidak akan menjadi kue hingga anda menggunakan seluruh bahannya sesuai takarannya secara presisi. Setelah semua bahan terkumpul, anda harus mencampur bahan bahan tersebut dengan urutan yang sesuai. Ketika urutan pencampuran bahan tadi tidak sesuai, bisa jadi hasil yang didapatkan tidak sesuai dengan harapan anda. Setelah bahan tercampur dengan baik dan menjadi adonan, adonan tadi dituangkan ke dalam wadah yang tepat. Kemudian adonan dipanggang dengan temperature yang pas, untuk jangka waktu tertentu. Akhirnya, dikeluarkan dari oven. Jadilah kue dengan tampilan yang standar. Selanjutnya kue didinginkan, lalu kemudian diberi hiasan yang cantik.

Tentunya tampilan ini lebih menarik dibandingkan dengan kondisi sebelumnya bukan?

Proses ini mirip sekali dengan ketika bekerja dengan data. Ketika berkerja dengan data, kadang kita memiliki seluruh bahan yang kita butuhkan, dan kita tahu seperti apa bentuk yang akan dihasilkan pada akhirnya. Masalahnya adalah kita seringkali tidak memiliki resep yang tepat yang mengantarkan kita dari kondisi awal ke kondisi akhirnya.



Representasi lain dari proses ini adalah seperti bermain lego. Anda akan sangat mengintimidasi audience anda jika anda menyajikan setumpukan data yang tidak terorganisasi dan terstuktur seperti pada gambar.



- ✓ Ketika anda selesai mengelompokan data, anda akan memulai mencari cara bagaimana menyusun setiap bagian ini menjadi suatu yang menarik.
- ✓ Ketika anda selesai mengelompokan data, anda akan memulai mencari cara bagaimana menyusun setiap bagian ini menjadi suatu yang menarik.
- ✓ Naik ke level berikutnya, seperti memberikan lapisan gula pada kue, atau menghias makanan, menyajikan data secara indah dapat menarik perhatian dan membuat audience anda mau menggunakan informasi yang anda hasilkan.

Ketika bekerja dengan data, akan lebih baik jika kita memiliki langkah-langkah yang jelas. Langkah-langkah ini kita sebut dengan *data life cycle* atau siklus hidup data. Siklus hidup data dapat didefinisikan dalam enam langkah, yaitu mendefinisikan pertanyaan riset, mengumpulkan dan mengorganisasikan data, membersihkan atau menyiapkan data, mengeksplorasi dan memvisualisasikan data, menganalisis dan menginterpretasi data, lalu

diakhiri dengan mengomunikasikan data. Setiap langkah dalam siklus hidup data ini sama pentingnya.

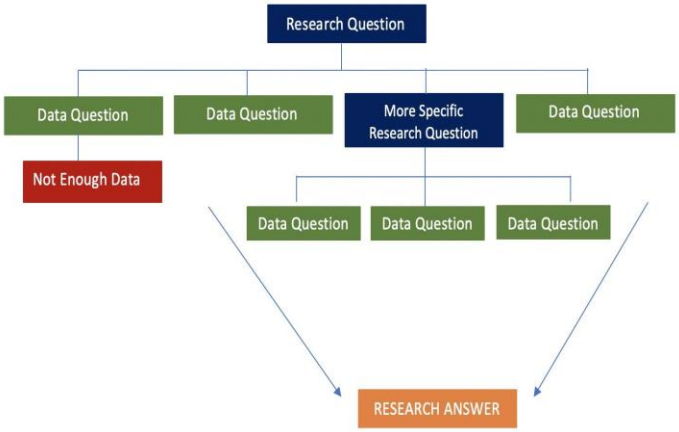
Mari kita bahas setiap langkahnya.

1. *Define a Research Question*

Siklus hidup data dimulai dengan mendefinisikan pertanyaan riset. Maksud dari tahap pertama ini adalah, ketika kita akan memulai bekerja dengan data kita perlu definisikan terlebih dahulu apa sih yang ingin kita ketahui?

Nah hal ini biasanya kita buat dalam bentuk pertanyaan riset.

Semakin jelas dan spesifik pertanyaan riset, akan mempermudah kita dalam menentukan data apa yang kita butuhkan untuk mendapatkan jawaban dari pertanyaan tadi.



2. *Collect & Organize Data*

Selanjutnya, ketika kita sudah mendefinisikan pertanyaan risetnya, langkah yang kita lakukan adalah mengumpulkan data lalu mengorganisasikannya. Perlu diingat ketika melakukan pengumpulan data, kita harus menentukan sumber data yang tepat.

Rambu-rambunya:

a) Pastikan bahwa kita memiliki hak akses terhadap data yang akan kita kumpulkan.
b) Apabila data yang kita kumpulkan mengandung informasi sensitive misalnya terkait data individu, pastikan informasi sensitifnya tidak sampai diketahui oleh pihak lain. Kita harus menjamin kerahasiaan dari data yang kita kelola.
c) Akan lebih aman jika kita menggunakan data yang memang sudah dipublikasikan oleh organisasi atau individu yang memiliki datanya. Misalnya data kesehatan yang dikelola oleh WHO atau world health organization, yang memang sengaja dipublikasikan dalam website resminya sebagai bahan riset bagi peneliti bidang kesehatan di seluruh dunia.

d) Mengorganisasikan data berdasarkan topik atau kategorinya.

Hal ini akan memudahkan kita ketika bekerja di langkah berikutnya.

3. *Wrangle Data*

Wrangle Data adalah membersihkan data. Yaitu membersihkan data yang tidak valid, tidak wajar, atau mungkin ada nilai yang kosong.

Misalkan kita memiliki data mahasiswa Angkatan 2020. Pada data tersebut ada variable tanggal lahir yang terdiri atas tanggal, bulan, dan tahun. Ketika ada mahasiswa yang memiliki tahun kelahiran 2020, tentunya nilai tersebut sangat tidak wajar. Untuk itu kita harus memperbaiki data yang ada. Cara perbaikannya bisa bermacam-macam bergantung tujuan risetnya. Jika risetnya ada kaitan dengan usia mahasiswa maka data ini akan menjadi outlier atau pencilan. Untuk mengantisipasinya, data ini bisa kita hapus atau kita ubah berdasarkan informasi pendukung lainnya, jika ada. Jika risetnya tidak ada kaitan dengan usia mahasiswa, maka data ini bisa kita biarkan apa adanya.

4. *Explore & Visualize Data*

Setelah data sudah bersih, kita dapat melakukan eksplorasi dan visualisasi data untuk mencari tahu hubungan antar variable. Yang dimaksud variable adalah karakteristik dari suatu data, misalkan pada data mahasiswa ada variable nama, alamat, asal sekolah, dan sebagainya.Eksplorasi data dilakukan dengan memeriksa variable-variable yang ada, dimulai dari variable target, satu per satu.Selanjutnya dilakukan perbandingan antar pasangan variable yang ada untuk mencari tahu relasi antara dua variable.

Terakhir, kita lakukan eksplorasi untuk mencari tahu relasi antara tiga variable atau lebih. Visualisasi data yang baik dapat membantu kita dalam mengidentifikasi relasi antar variable-variabel tersebut.

5. *Analyze & Interpret Data*

Dari tahapan eksplorasi data kita bisa mendapatkan pola umum dari data yang kita miliki.

Akan tetapi hal itu masih belum cukup untuk digunakan dalam membuat suatu kesimpulan yang spesifik. Langkah berikutnya yang perlu kita lakukan adalah melakukan eksplorasi yang lebih detail, untuk mendapatkan pola dan relasi yang lebih jelas. Lalu melakukan uji statistic dan membangun model statistic. Dengan melakukan semua hal ini, kita bisa mendapatkan klaim yang lebih kuat untuk kemudian memperluas klaimnya ke populasi yang lebih umum.

Di tahap ini harapannya kita sudah mendapatkan jawaban atas pertanyaan riset yang kita definisikan di awal.

6. *Communicate with Data*

Pada tahap terakhir, kita akan focus ke strategi bagaimana menulis, menjelaskan, dan mempresentasikan hal-hal yang sudah kita temukan ke berbagai audiens.

Kita bisa mempertajam visualisasi data yang sudah kita dapatkan pada tahap sebelumnya agar lebih mudah dipahami oleh audiens.

Akan lebih baik lagi jika kita bisa membuat visualisasi yang interaktif.