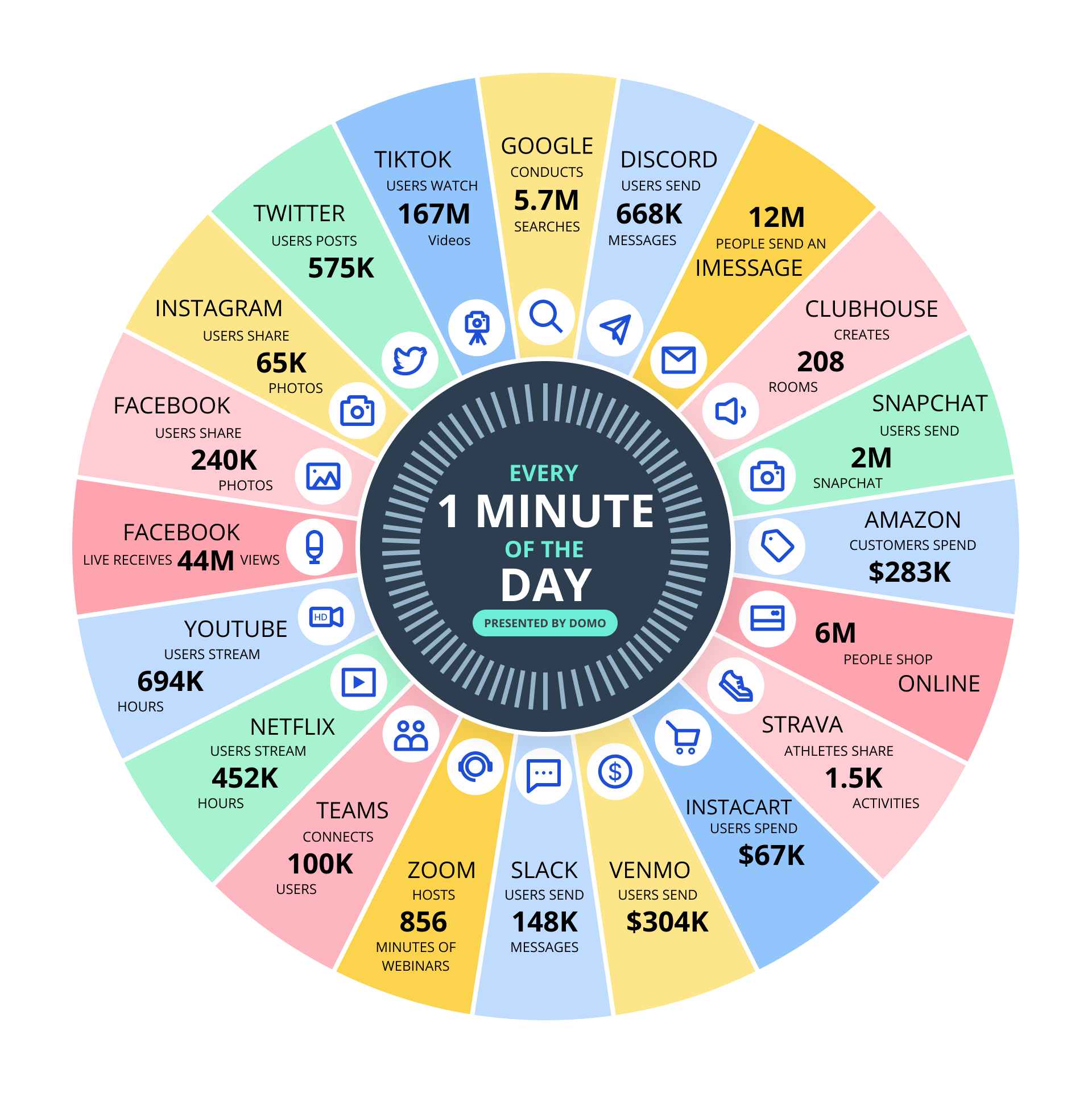
**Pengenalan Analisis Data**

Di kelas ini, akan tersaji berbagai materi mengenai analisis data yang luar biasa menarik dan atraktif, mulai dari siklus pengolahan data, penerapan dasar-dasar statistik, pertimbangan dalam pengolahan data, berbagai teknik dalam *data wrangling*, *exploration* dan *visualization*, hingga cara membuat dashboard.

*First of all*, kita perlu mengetahui terlebih dahulu apa itu data? Data merupakan sekumpulan fakta yang dapat direpresentasikan dalam berbagai bentuk seperti angka, gambar, video, teks, hasil pengukuran, dll.

Tahukah Anda bahwa sebagai generasi kekinian, kita telah memproduksi banyak data setiap harinya? Ketika kita mencari sesuatu secara online, mendengarkan musik atau bahkan berbelanja melalui *e-commerce*, semua kegiatan tersebut akan menghasilkan banyak data. Gambar di bawah ini menunjukkan banyaknya data yang diproduksi tiap menit dari berbagai aplikasi yang sering kita gunakan [1].



Selain menghasilkan banyak data, tanpa sadar kita juga sering memanfaatkan data untuk mengambil sebuah keputusan. Sebagai contoh, sebelum membeli sebuah barang secara online terdapat banyak hal yang perlu dipertimbangkan, seperti rating, testimoni, serta reviu dari orang lain terkait toko dan barang tersebut. Hal ini dilakukan untuk memastikan barang tersebut sesuai dengan ekspektasi kita.

Kegiatan di atas merupakan bentuk sederhana dari pemanfaatan analisis data dalam kehidupan sehari-hari. Oleh karena itu, analisis data dapat diartikan sebagai sebuah proses mengumpulkan, mentransformasi, dan menata data untuk menarik kesimpulan, membuat prediksi, serta memberi pertimbangan yang tepat dalam mengambil keputusan.

Sederhananya, proses analisis data akan mengubah data menjadi sebuah *insight* (wawasan) yang selanjutnya akan digunakan untuk menyelesaikan sebuah masalah. Nah, orang yang melakukan proses analisis data ini sering disebut sebagai *data analyst*.

Anda bisa membayangkan seorang data analyst sebagai detektif yang berusaha mencari sebuah solusi dari suatu permasalahan menggunakan data. Untuk memperoleh solusi yang paling efektif, ia harus mampu menggali insight dari data melalui sebuah proses yang disebut analisis data. Proses ini tentunya membutuhkan beberapa keterampilan seperti berikut [2].

* **Analytical Skills**  
  Seorang data analyst yang andal harus memiliki *analytical skills*atau keterampilan analitikal yang baik. Keterampilan ini berhubungan dengan kualitas dan karakteristik yang berhubungan dengan penyelesaian masalah menggunakan fakta. Berikut merupakan beberapa poin penting dalam analytical skill.
  + Memiliki rasa ingin tahu yang tinggi untuk belajar hal baru serta mencari pengalaman dan tantangan yang baru.
  + Memiliki kemampuan untuk memahami konteks atau kondisi di balik suatu kejadian.
  + Memiliki pola pikir teknis sehingga mampu menguraikan suatu hal menjadi langkah-langkah yg lebih kecil dan mengerjakan tiap bagiannya dengan cara yang teratur dan logis.
  + Mampu mengolah atau mengatur informasi secara baik dengan menerapkan *data design*.
  + Mampu menerapkan *data strategy* untuk manajemen orang, proses, dan tools yang digunakan dalam analisis data.
* **Analytical Thinking**  
  Selain analytical skills, seorang data analyst yang andal juga harus mampu menerapkan *analytical thinking*(pemikiran analisis) dengan baik. Hal ini melibatkan pengidentifikasian dan pendefinisian sebuah masalah, kemudian menyelesaikannya berdasarkan data dengan cara yang terorganisasi dan berurutan. Secara umum, terdapat enam aspek kunci dari analytical thinking antara lain sebagai berikut.
  + Merepresentasikan sebuah informasi dalam bentuk grafik atau visual. Hal ini akan membantu seorang data analyst untuk memahami dan menyampaikan sebuah informasi secara efektif.
  + Selalu menerapkan pola pikir strategis untuk menjaga fokus dengan membuat perencanaan matang. Perencanaan yang matang tersebut akan membantu seorang data analyst dalam memahami data dan mencapai sebuah tujuan atau *goals* yang ingin dicapai.
  + Menerapkan pola pikir yang berorientasi pada masalah untuk mengidentifikasi, mendeskripsikan, dan menyelesaikan sebuah *business problem*.
  + Mengidentifikasi korelasi antara dua data atau lebih. Hal ini akan membantu data analyst dalam menarik sebuah kesimpulan atau insight dari data.
  + Menerapkan *big-picture thinking* (pemikiran gambaran besar) dalam proses analisis. Hal ini berarti seorang data analyst perlu melihat gambaran besar dari sebuah data dan masalah untuk menilai probabilitas dan oportunitas dalam menemukan ide baru atau inovasi yang menarik.
  + Selain melihat gambaran besar, seorang data analyst juga harus mampu menerapkan *detail-oriented thinking* (pemikiran yang berorientasi pada detail) dalam proses analisis. Ia akan membantu data analyst untuk mengidentifikasi tiap aspek yang dibutuhkan untuk mewujudkan ide baru dan inovasi yang ingin dibuat.

**Pentingnya Data dalam Dunia Bisnis**

Pada materi sebelumnya, kita telah melihat banyaknya data yang dihasilkan tiap menit melalui berbagai platform online. Data tersebut mengandung semua informasi berkaitan tentang kita dan semua aktivitas yang dilakukan pada platform tersebut. Setelah mengetahui hal ini, mungkin mulai muncul pertanyaan berikut dalam benak Anda.

* “Kenapa perusahaan tersebut mengumpulkan data sebanyak itu?”
* “Apa yang akan mereka lakukan dengan data tersebut?”

Di zaman yang serba canggih seperti saat ini, terdapat banyak inovasi dalam pemanfaatan teknologi dan informasi terutama pada bidang data. Data merupakan salah satu resource penting bagi perusahaan dari berbagai industri mulai dari e-commerce, entertainment, manufacturing, healthcare, marketing, finance, tech, dll.

Berbagai perusahaan tersebut membutuhkan data untuk mengidentifikasi peluang, tren, mengoptimalkan proses yang ada, meningkatkan layanan pelanggan, dan masih banyak lagi penerapannya di industri. Hal inilah yang mengakibatkan data dijuluki sebagai “The new oil” karena dianggap sebagai sumber daya paling penting saat ini.

Selain itu, *insigh*t (wawasan) yang diperoleh dari data dapat digunakan sebagai pendukung dalam pengambilan keputusan. Praktek seperti ini mendorong terbentuknya sebuah kultur baru dalam dunia bisnis yang disebut *data-driven decision making*.

**Data-Driven Decision Making**

*By definition*, data-driven decision making merupakan sebuah kultur, *best practice*, serta proses dalam penggunaan data dan hasil analisis untuk memandu seluruh pengambilan keputusan bisnis yang tentunya sejalan dengan objektif perusahaan. Umumnya perusahaan atau organisasi yang menerapkan kultur ini memiliki beberapa karakteristik seperti berikut.

* Memiliki data yang tersentralisasi dan terorganisir dengan baik.
* Memiliki kebijakan dan ketentuan terkait tata kelola data.
* Memiliki tingkat aksesibilitas data yang tinggi.
* Memiliki tool analisis yang mudah digunakan.

Tahukah Anda, banyak perusahaan besar yang mulai menerapkan kultur data-driven decision making karena dinilai memiliki banyak manfaat. Berikut merupakan beberapa manfaat penerapan data-driven decision making [3].

* Manfaat utama dalam data-driven decision making ialah akuntabilitas dan konsistensi dalam setiap pengambilan keputusan. Hal ini akan mempermudah perusahaan dalam membuat sebuah *benchmark* yang jelas untuk mengevaluasi setiap keputusan dan strategi bisnis yang dibuat sehingga mampu mendorong perkembangan perusahaan secara kontinu.
* Pengambilan keputusan berdasarkan data akan mengarahkan perusahaan untuk berinovasi dan memperoleh peluang bisnis.
* Data-driven decision making mampu meningkatkan efisiensi dengan menyediakan berbagai informasi yang relevan dalam mengambil sebuah keputusan.

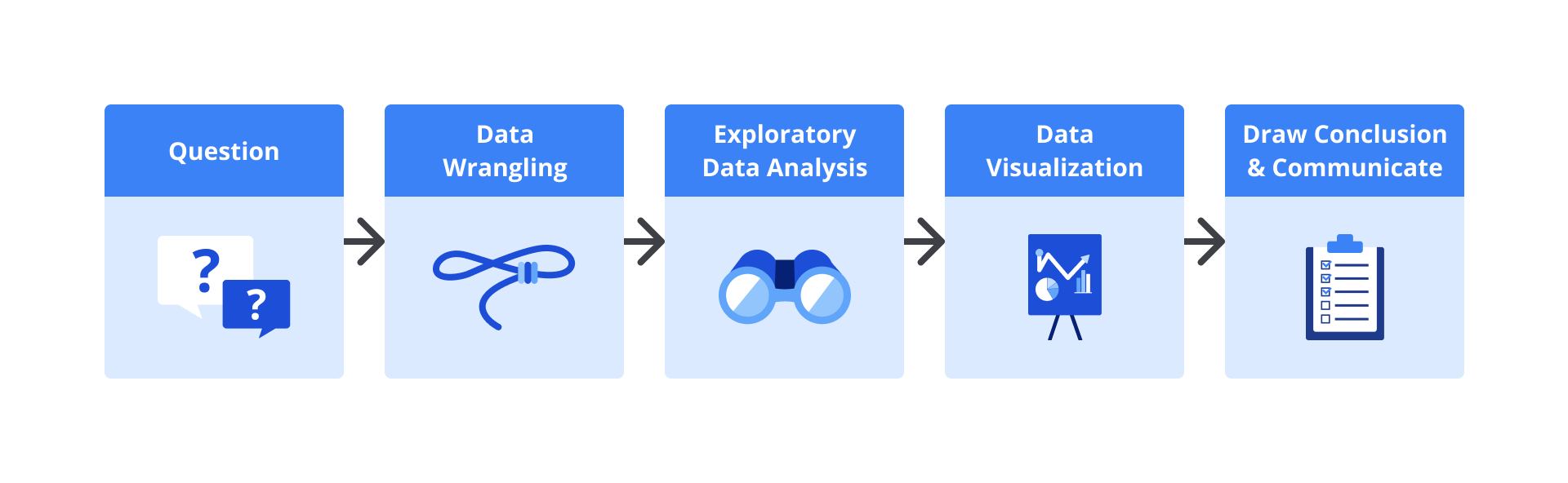
Oke, pada tahap ini, Anda telah memahami pentingnya data-driven decision making untuk perkembangan bisnis. Pertanyaan selanjutnya adalah bagaimana cara menerapkan hal tersebut? Terdapat beberapa poin yang harus diperhatikan ketika ingin menerapkan kultur data-driven decision making yaitu seperti berikut.

* **Mengumpulkan data sebanyak mungkin**  
  Pada kultur data-driven decision making, segala hal yang kita lakukan sangat bergantung pada data. Oleh karena itu, kita sangat disarankan untuk mengumpulkan data sebanyak mungkin. Namun, perlu diingat bahwa selain kuantitas, kita juga perlu memastikan kualitas dari data yang dikumpulkan.
* **Menentukan tujuan yang dapat diukur**  
  Setelah mengumpulkan data, tahap selanjutnya adalah menentukan tujuan yang ingin dicapai. Tujuan tersebut harus dapat diukur untuk mempermudah proses evaluasi dan improvement.
* **Memastikan aksesibilitas data**  
  Idealnya, kultur data-driven decision making diterapkan pada seluruh komponen dalam sebuah organisasi. Oleh karena itu, kita perlu memastikan bahwa setiap *department* memiliki akses dan visibilitas data yang cukup.
* **Menentukan prioritas**  
  Untuk memastikan setiap keputusan yang dibuat berhasil dikerjakan, kita perlu menentukan prioritas terlebih dahulu. Hal ini akan membantu kita dalam membuat keputusan yang paling berdampak terhadap perkembangan perusahaan. Selain itu, hal ini juga akan membantu perusahaan dalam menggunakan sumber daya seefisien mungkin.

**Mengenal Siklus Analisis Data**

Setelah memahami semua pembahasan dalam materi Pengenalan Analisis Data dan Pentingnya Analisis Data dalam Dunia Bisnis, kini mungkin Anda mulai mulai bertanya, “Bagaimana kiat-kiat atau cara untuk melakukan analisis data?”.

Nah, kali ini kita akan fokus pada tahapan analisis data yang umum dijumpai di industri. Jika Anda sudah siap, yuk, simak penjelasan berikut!



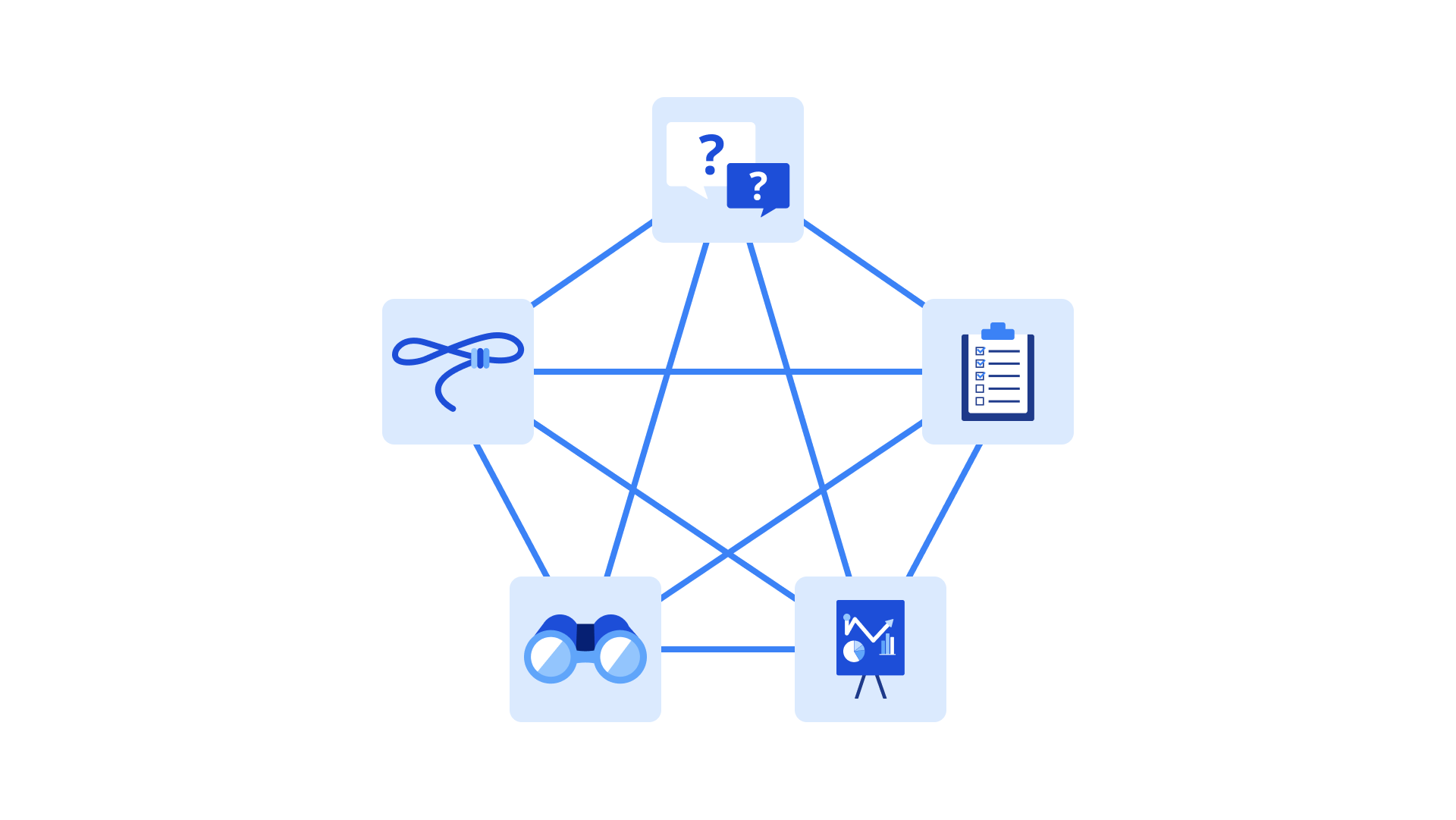
Seperti yang telah kita bahas bersama, analisis data merupakan sebuah proses mengumpulkan, mentransformasi, dan menata data untuk menarik kesimpulan, membuat prediksi, serta memberi pertimbangan yang tepat dalam mengambil keputusan. Sebagai suatu proses, analisis data tentunya memiliki beberapa tahapan seperti berikut.

1. **Mendefinisikan pertanyaan**  
   Pada awal proyek analisis data, kita perlu mendefinisikan berbagai pertanyaan analisis terkait data ataupun permasalahan yang ingin diselesaikan. Pertanyaan tersebut akan membantu kita dalam memahami masalah bisnis yang dihadapi serta tujuan yang ingin dicapai. Bukan hanya itu, dengan adanya pertanyaan tersebut, kita akan merasa terbantu karena bisa lebih fokus pada bagian data yang relevan dengan tujuan.  
     
   Pada proses analisis data, terkadang kita diberikan data terlebih dahulu lalu mulai mendefinisikan pertanyaan analisis berdasarkan data tersebut. Namun, tidak jarang kita juga berangkat dari pertanyaan analisis terlebih dahulu lalu mencari data yang sesuai untuk menjawab pertanyaan tersebut.  
     
   Dalam mendefinisikan pertanyaan analisis data, seringkali digunakan konsep 4W, yaitu ***What* (Apa), *Why* (Mengapa), *When* (Kapan), dan *Where* (Dimana)**. Keempat pertanyaan ini membantu dalam merinci dan merumuskan pertanyaan analisis data dengan lebih spesifik dan terfokus. Berikut adalah penjelasan singkat tentang masing-masing W:

* *What* (Apa): Pertanyaan ini membantu mengidentifikasi variabel atau fenomena yang sedang diamati. Apa yang ingin Anda ketahui atau pelajari dari data yang ada?  
  **Contoh:**Apa hubungan antara tingkat kepuasan pelanggan dengan penjualan produk?
* *Why* (Mengapa): Fokus pada alasan atau penyebab di balik suatu peristiwa atau pola dalam data. Mengapa suatu kejadian atau tren terjadi?  
  **Contoh:**Mengapa terjadi penurunan penjualan pada bulan tertentu?
* *When* (Kapan): Menentukan waktu atau periode tertentu yang menjadi fokus analisis. Kapan suatu peristiwa terjadi atau pola tertentu muncul?  
  **Contoh:**Kapan waktu paling efektif untuk meluncurkan kampanye pemasaran?
* *Where* (Dimana): Menentukan lokasi atau wilayah geografis yang relevan dengan analisis. Di mana peristiwa atau pola ini terjadi?  
  **Contoh:** Dimana lokasi geografis dengan tingkat penggunaan produk tertinggi.

1. **Data wrangling**  
   Setelah kita memiliki pertanyaan analisis yang baik, tahap selanjutnya adalah data wrangling. Pada tahap ini, Anda memulai dengan mengumpulkan data yang dibutuhkan untuk menjawab pertanyaan analisis yang telah dibuat sebelumnya. Pada tahap ini, data yang diperlukan untuk menjawab pertanyaan analisis dikumpulkan dari berbagai sumber. Data bisa berasal dari database internal, file eksternal, API, atau sumber data lainnya.  
     
   Kemudian Anda perlu menilai kualitas dan struktur dari data tersebut. Hal ini dilakukan untuk mengidentifikasi masalah yang terdapat dalam data serta membuat strategi pembersihan data yang sesuai. Proses pembersihan data umumnya dilakukan dengan memodifikasi, mengganti, atau menghilangkan data yang bermasalah sehingga diperoleh dataset yang berkualitas dan terstruktur dengan baik.  
     
   *FYI*, pembersihan data merupakan salah satu proses yang paling menghabiskan waktu dalam proyek analisis data. Oleh karena itu, kita sangat disarankan untuk menyediakan waktu yang cukup untuk melakukan pembersihan data.
2. **Exploratory Data Analysis**  
   Tahap berikutnya adalah Exploratory Data Analysis atau sering disingkat EDA. Pada tahap inilah Anda akan melakukan eksplorasi terhadap data yang telah dibersihkan untuk memperoleh *insight* dan menjawab pertanyaan analisis.  
     
   Proses eksplorasi umumnya dilakukan dengan memanfaatkan berbagai teknik *descriptive statistics* (penggunaan konsep statistik untuk mendeskripsikan data). Hal ini bertujuan untuk menemukan pola, hubungan, serta membangun intuisi terkait data yang diolah.  
     
   Selain menggunakan descriptive statistic, terkadang kita juga menggunakan teknik machine learning atau *inferential statistic* untuk memprediksi data di masa depan. Namun, kedua teknik tersebut di luar cakupan pembahasan kita pada kelas ini.
3. **Data visualization**  
   Hasil yang diperoleh dari EDA perlu divisualisasikan melalui tahap data visualization. Pada tahap ini, kita akan menerapkan berbagai teknik dan jenis visualisasi data yang tepat untuk mengomunikasikan temuan atau insight dari hasil analisis secara efektif.
4. **Draw conclusion & communicate**  
   Pada akhir proyek analisis data, kita perlu membuat conclusion atau kesimpulan dari hasil analisis. Kesimpulan yang dibuat haruslah menjawab semua pertanyaan dan tujuan yang telah didefinisikan di awal.  
     
   Selain membuat kesimpulan, kita juga perlu membagikan temuan atau insight dari hasil analisis. Terdapat banyak cara dalam mengomunikasikan hasil analisis, seperti menggunakan report, slide deck, dan atau presentation.

Nah, itulah beberapa tahapan yang umum dijumpai dalam proyek analisis data. Di industri, tahapan tersebut tidak selamanya berbentuk liner. Terkadang setelah melakukan EDA, Anda akan menemukan masalah baru dalam data sehingga perlu membersihkan kembali data tersebut. Selain itu, terkadang kita juga menemukan sebuah pola yang menarik data sehingga kita perlu mengubah pertanyaan analisis yang telah kita definisikan. Hal inilah yang membuat tahapan dalam analisis data lebih cocok disebut sebagai sebuah siklus karena terdapat banyak iteratif proses di dalamnya.



**Pengenalan Tool Pengolahan Data**

Pada materi sebelumnya, kita telah mengenal konsep dasar dalam analisis data, manfaat analisis data, serta siklus analisis data yang umum dijumpai di industri. Nah, pada bagian ini, kita akan berkenalan dengan beberapa tools pengolahan data yang sering digunakan oleh para praktisi data.

Sesuai dengan nama kelas ini, seluruh tool yang akan kita bahas tersedia dalam bahasa pemrograman Python. Oleh karena itu, bagi yang belum familier dengan bahasa pemrograman Python, Anda bisa mengambil kelas [Memulai Pemrograman dengan Python](https://www.dicoding.com/academies/86) terlebih dahulu.

Sebagai salah satu bahasa pemrograman kekinian, Python dapat digunakan untuk mengerjakan banyak tugas salah satunya adalah mengolah dan memvisualisasikan data. Hal inilah yang membuat Python menjadi salah satu bahasa pemrograman paling populer di kalangan praktisi data.

Untuk mendukung proses pengolahan dan analisis data, Python memiliki banyak *libraries* dan *tools* yang bersifat *open-source* dan dapat digunakan secara bebas. Berikut merupakan tiga libraries yang paling sering digunakan untuk mengolah data.

* Library NumPy
* Library pandas
* Library SciPy

Nah, pada materi kali ini kita akan fokus membahas ketiga library tersebut. Apakah Anda sudah siap? Yuk, simak penjelasan berikut!

**Pengenalan Library NumPy**

NumPy (singkatan dari *numerical Python*) merupakan salah satu library Python yang sangat *powerful* untuk membuat dan mengolah *multi-dimensional arrays* (sering juga disebut sebagai matriks atau tensor). NumPy dibuat dengan tujuan untuk mempercepat proses komputasi numerik dalam bahasa pemrograman Python.

Untuk mendukung tujuan tersebut, NumPy hadir dengan banyak fungsi bawaan yang memungkinkan kita untuk melakukan berbagai macam operasi matematis mulai dari yang sederhana hingga kompleks, seperti *basic linear algebra*, *random simulation*, *Fourier transforms*, dll. Hal inilah yang membuat NumPy menjadi salah satu library populer untuk melakukan *scientific computing* dalam bahasa pemrograman Python. Selain itu, NumPy juga banyak digunakan sebagai komponen dasar pada banyak library populer lain, seperti pandas, SciPy, dll.

Nah, sebelum mulai menggunakan NumPy, Anda perlu ingat bahwa ia **bukan** merupakan library bawaan Python sehingga perlu diinstal terlebih dahulu. Anda dapat menginstal NumPy menggunakan *package manager* seperti **pip** ataupun **conda**.

**Pengenalan Library Pandas**

Selain NumPy, library lain yang sering digunakan adalah pandas. Ia merupakan library Python yang spesifik digunakan untuk memanipulasi dan menganalisis data. Library inilah yang akan kita gunakan sebagai tool utama dalam menganalisis data.

Untuk memenuhi fungsinya, pandas menyediakan *data structure* bernama DataFrame. Ia merupakan sebuah data structure berbentuk tabel yang terdiri dari *header*, *index*, *column*, & *row*. Data structure inilah yang akan kita gunakan untuk mengolah dan menganalisis data.

Selain itu, pandas juga menyediakan sekumpulan *function* serta *method* untuk memanipulasi dan menganalisis data secara lebih efektif. Hal inilah yang membuat pandas disebut sebagai salah satu library populer untuk mengolah data dan sering digunakan oleh para praktisi data.

Sama halnya dengan NumPy, pandas juga **bukan** library bawaan Python sehingga perlu diinstal terlebih dahulu. Anda dapat menginstal pandas dengan menjalankan salah satu perintah berikut.

**Pengenalan Library SciPy**

Library penting selanjutnya adalah SciPy (singkatan dari Scientific Python). Ia merupakan library Python yang khusus digunakan untuk kebutuhan komputasi saintifik. Library ini memanfaatkan NumPy sebagai komponen dasar dalam setiap perhitungan komputasi yang akan dilakukan.

Sebagai library yang sering digunakan untuk komputasi saintifik, SciPy menyediakan banyak fungsi yang dapat digunakan untuk menyelesaikan operasi matematis kompleks yang dapat digunakan dalam dunia science dan engineering. Selain itu, SciPy juga menyediakan banyak operasi statistik sehingga dapat membantu kita dalam mengidentifikasi parameter statistik dari suatu data.

SciPy juga **bukan**merupakan library bawaan Python sehingga perlu diinstal terlebih dahulu. Anda dapat menginstal SciPy dengan menjalankan salah satu dari perintah berikut.

**Pengenalan Tool Visualisasi Data**

Pada materi sebelumnya, kita berkenalan dengan berbagai tools atau library Python yang sering digunakan oleh para praktisi di industri dalam mengolah data. Namun, tools tersebut masih belum cukup untuk menangani seluruh kebutuhan kita dalam proses analisis data.

Pada proses analisis data, selain mengolah data, kita juga perlu merepresentasikan data tersebut ke dalam bentuk visual yang menarik dan mudah dipahami. Nah, pada materi kali ini, kita akan berkenalan dengan berbagai tools visualisasi data yang ada dalam bahasa pemrograman Python.

Sebagai salah satu bahasa pemrograman paling populer di kalangan praktisi data, Python memiliki banyak *libraries* dan *tools* yang bersifat*open-source* untuk mendukung pembuatan visualisasi data yang baik dan menarik. Berikut merupakan dua library Python yang paling sering digunakan oleh praktisi data di industri.

* Library matplotlib
* Library seaborn

Jika Anda masih asing dengan dua library tersebut, simak pembahasannya di bawah ini.

**Pengenalan Library Matplotlib**

Matplotlib merupakan sebuah library Python yang *reliable* dan komprehensif untuk mendukung kebutuhan pembuatan visualisasi data. Matplotlib menyediakan banyak *object* dan *method* yang memungkinkan kita membuat visualisasi data statis, *animated*, dan interaktif. Anda dapat melihat berbagai contoh visualisasi data yang dibuat menggunakan library ini pada tautan berikut: [examples gallery](https://matplotlib.org/stable/gallery/index.html).

Matplotlib memiliki sebuah*submodule* bernama **matplotlib.pyplot** yang sering dipanggil atau disingkat sebagai **plt**. Submodule ini menyediakan sebuah fungsi untuk membuat sebuah *figure* yang merupakan sebuah area atau kanvas kosong untuk membuat visualisasi data. Selain itu, plt juga menyediakan banyak fungsi untuk mendefinisikan berbagai jenis visualisasi data beserta dekorasinya, seperti label, title, dll.

Sebagai library visualisasi data yang komprehensif, matplotlib juga memungkinkan kita untuk memodifikasi berbagai tampilan visual sehingga dapat disesuaikan dengan kebutuhan. Tidak hanya itu, ia juga memungkinkan kita untuk menyimpan hasil visualisasi data yang telah dibuat ke dalam berbagai format penyimpanan, seperti png, jpeg, pdf, dll.

Sama halnya dengan berbagai library yang telah kita bahas sebelumnya, matplotlib **bukan**merupakan library bawaan Python sehingga perlu diinstal terlebih dahulu. Anda dapat menginstal matplotlib dengan menjalankan salah satu dari perintah berikut.

**Pengenalan Seaborn**

Selain matplotlib, library lain yang sering digunakan untuk visualisasi data adalah seaborn. Seaborn merupakan library Python yang spesifik digunakan untuk membuat visualisasi data yang atraktif dan informatif. Library ini memanfaatkan matplotlib untuk menampilkan grafik visualisasi data.

Sebagai salah satu library visualisasi data yang andal, seaborn menyediakan berbagai bentuk visual untuk merepresentasikan data secara lebih menarik dan informatif. Anda dapat melihat berbagai contoh visualisasi data yang dibuat menggunakan library ini pada tautan berikut: [example gallery](https://seaborn.pydata.org/examples/index.html).

Pada proses pembuatan visualisasi data, umumnya kita akan memadukan seaborn dan matplotlib. Hal ini dilakukan untuk menghasilkan bentuk visual yang lebih menarik, informatif, dan interaktif.