**Teknologi Penyimpanan Data**

**Berkenalan dengan Basis Data**

Menurut**Amazon Web Service (AWS)** basis data adalah koleksi data sistematis yang disimpan secara elektronik. Ini dapat berisi semua jenis data, seperti kata, angka, gambar, video, dan file. Anda dapat menggunakan perangkat lunak (*software*) yang disebut Database Management System (DBMS) atau sistem manajemen basis data.

Ada pula menurut sumber lain, yaitu **Oracle** menyatakan bahwa basis data adalah kumpulan terorganisasi dari informasi terstruktur, atau data, biasanya disimpan secara elektronik dalam sistem komputer. Basis data biasanya dikendalikan oleh Database Management System (DBMS) atau sistem manajemen basis data. Nah, data dan DBMS beserta aplikasi yang terkait dengannya disebut sebagai sistem basis data atau juga bisa disebut basis data saja.

**Tipe-Tipe Basis Data**

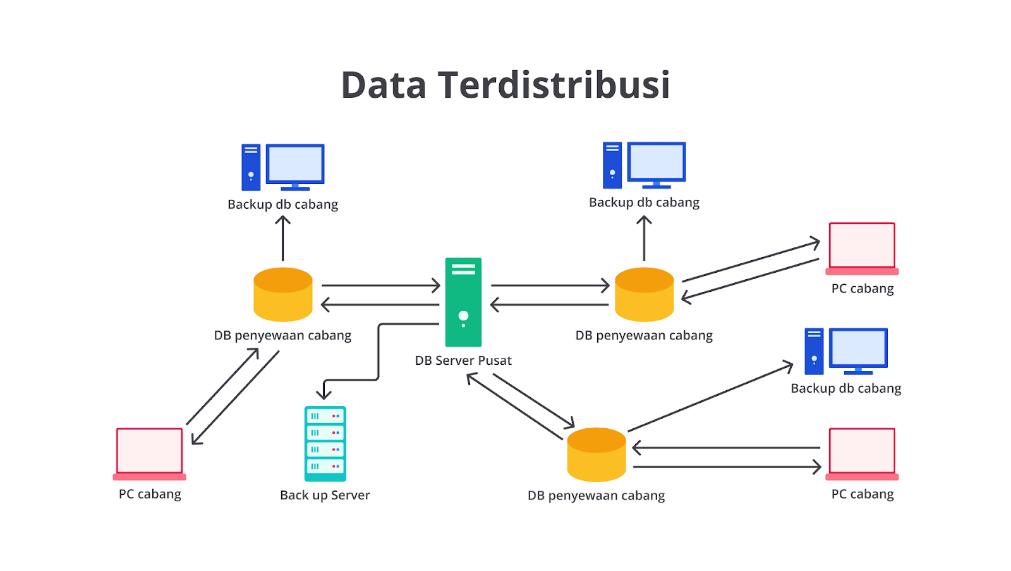
Tentu ada banyak jenis basis data yang dapat diaplikasikan, tetapi pada nyatanya basis data yang terbaik untuk organisasi bergantung pada bagaimana organisasi bermaksud menggunakan data tersebut. Berikut merupakan beberapa tipe-tipe basis data.

Basis Data Relasional (*Database Relational*)  
Dalam sistem manajemen basis data relasional (RDBMS), hubungan setiap data memiliki sifat yang relasional serta tersimpan dalam bentuk kolom dan baris. SQL (*Structured Query Language*) adalah bahasa pemrograman yang digunakan untuk permintaan RDBMS. Beberapa produk yang biasa digunakan adalah DB Browser SQLite, MySQL, Oracle, dll. Penjelasan Basis Data Relasional akan dibahas lebih lanjut pada kelas selanjutnya.

Basis Data Orientasi Objek (OOD - *Object-oriented Databases*)

*Object oriented database*adalah sebuah model basis data yang menyimpan informasi dalam bentuk objek. Beberapa produk yang biasa digunakan adalah IBM Db2, InterSystems IRIS, Google Cloud Storage for Firebase, dll.

Basis Data Terdistribusi (*Distributed Databases*)



Basis data terdistribusi adalah kumpulan data yang digunakan bersama-sama dan memiliki hubungan logis fisik tersebar di jaringan komputer. Konsep ini mendasarkan pada prinsip bahwa data dibuat, disediakan, dan dikelola secara terpusat.

Gudang Data (Data Warehouse)

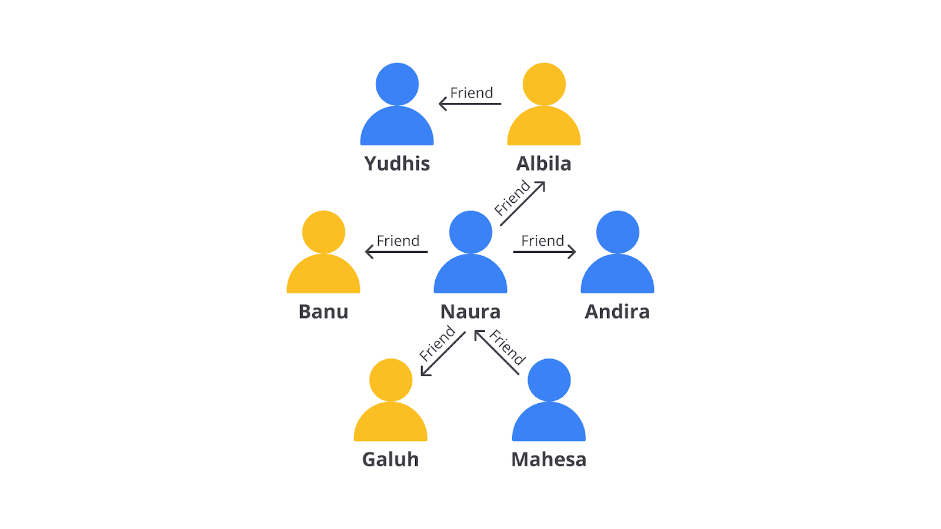


Sebuah sistem atau platform yang dirancang untuk mengumpulkan, menyimpan, menganalisis, dan mengelola data dari berbagai sumber yang berbeda untuk tujuan analisis bisnis dan pengambilan keputusan. Beberapa *tools*data warehouse di antaranya Amazon Redshift, Snowflake, Google BigQuery, dll.

Basis Data NoSQL

NoSQL atau basis data non-relasional, memungkinkan data tidak terstruktur dan semi terstruktur untuk disimpan dan dimanipulasi (berbeda dengan basis data relasional, yang menentukan semua data yang dimasukkan ke dalam basis data harus disusun). Beberapa contoh basis data NoSQL antara lain MongoDB, Cassandra, dan Couchbase.

Basis Data Grafik (Graph Databases)



Basis data grafik dibuat khusus untuk menyimpan dan menavigasi hubungan. Basis data grafik menggunakan node untuk menyimpan entitas data dan edge untuk menyimpan hubungan antar entitas. Gambar di atas merupakan analogi dari penerapan basis data grafik pada jejaring sosial. Mengingat orang (node) dan hubungan mereka (edge), Anda dapat mengetahui siapa "teman dari teman" dari orang tertentu—misalnya, teman dari teman Howard.

**Basis Data Everywhere**

*“Berbisnis dengan hati nurani sangat bermanfaat” -*

*Ingvar Kamprad (CEO IKEA)*

Berlatar belakang dari *quotes*di atas, tentu setiap hal apa pun di muka bumi terdapat manfaat yang dapat digunakan oleh setiap insan. Berikut merupakan manfaat dari adanya basis data yang digunakan oleh industri untuk berbagai kasus penggunaan.

1. **Deteksi penipuan**  
   Basis data grafik membantu manajemen identitas dan mendeteksi penipuan. Ketika digabungkan dengan algoritma *machine learning,* ia dapat menemukan pola dan mendeteksi aktivitas penipuan secara otomatis.
2. **Manajemen dokumen**  
   Basis data NoSQL juga dapat menyimpan dan mengelola dokumen, seperti artikel dan kontrak. Basis data NoSQL juga memungkinkan organisasi untuk melakukan kueri dan mengindeks dokumen.
3. **Gaming dan hiburan**  
   Banyak perusahaan *gaming* dan hiburan menggunakan basis data secara luas untuk memberikan pengalaman bermain yang lebih baik, seperti akses masuk simultan untuk jutaan pengguna.

**Data, Basis Data, Dataset**

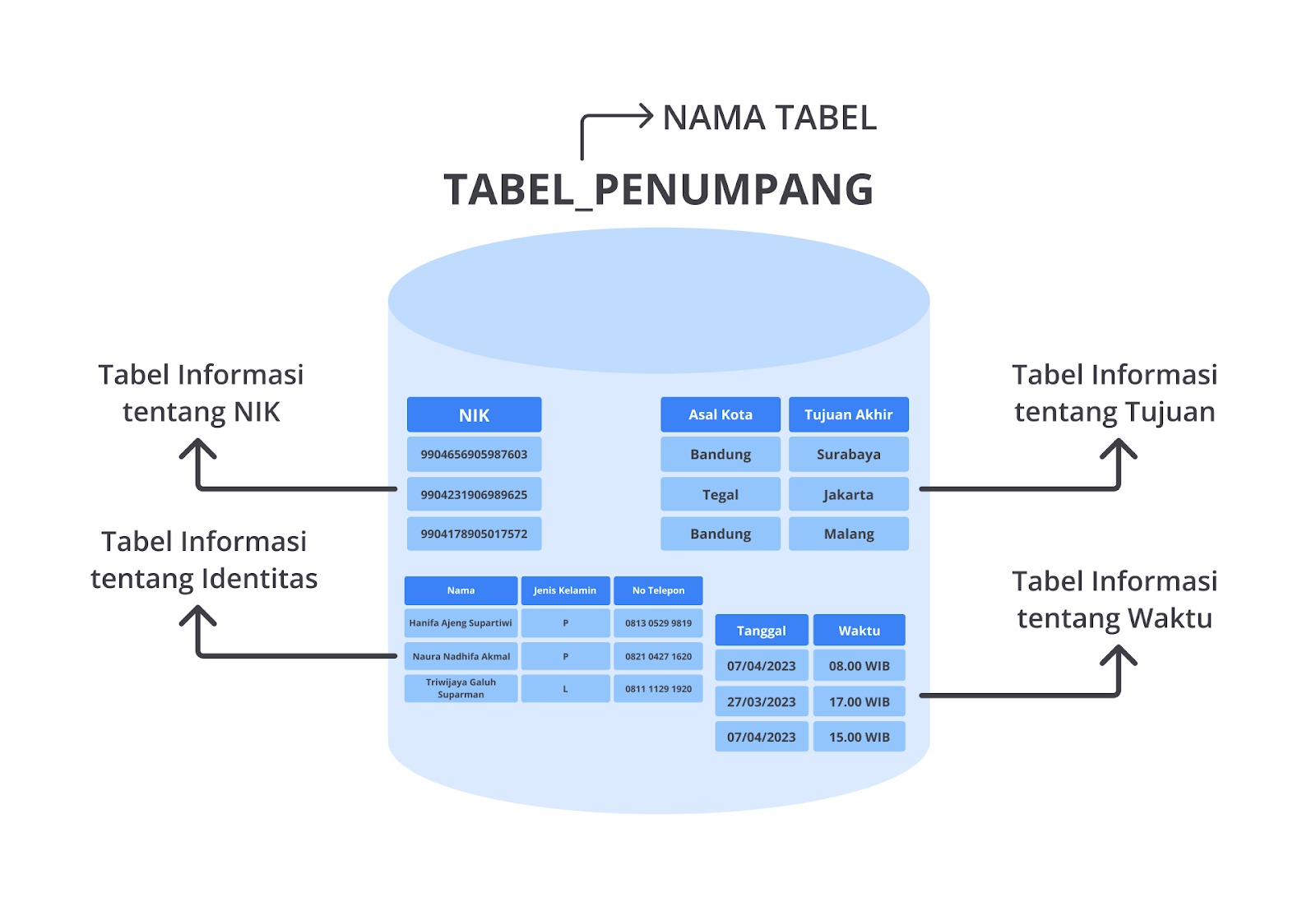
Mungkin terlintas di benak Anda “*Duh… bingung…*Apa perbedaan antara data, basis data, dan dataset?”

Nah, di bawah ini merupakan perbandingan antara data, basis data, dan dataset yang perlu Anda ketahui.

*First thing first,*di modul sebelumnya telah dijelaskan bahwa menurut Kamus Besar Bahasa Indonesia, data adalah informasi, fakta, dari keterangan yang benar dan nyata mengenai suatu objek atau kejadian yang bisa dicatat atau disimpan. Lalu, apa perbedaan antara basis data dan dataset?

|  |  |
| --- | --- |
| **Basis Data** | **Dataset** |
| Basis data adalah kumpulan data yang diatur dan disimpan dengan cara yang memungkinkan akses dan pengambilan yang mudah. Selain itu, ia juga dapat menyimpan berbagai macam tipe data, termasuk teks, nomor, gambar, dan tipe data lainnya. | Dataset adalah kumpulan data diatur dalam format tertentu, seperti spreadsheet, CSV, atau basis data. Umumnya, digunakan untuk keperluan spesifik. |
| Memiliki banyak kumpulan data dan dapat digunakan untuk aplikasi yang berbeda. | Dapat menjadi bagian dari data yang diambil dari basis data yang lebih besar. Bisa juga, dataset diambil dari sumber lain yang mendukung formatnya. |
| Contoh tools: BigQuery, MySQL Clients, DB Browser SQL, dll. | Contoh tools: Google Spreadsheet, Excel, Situs Public Datasets, dll. |

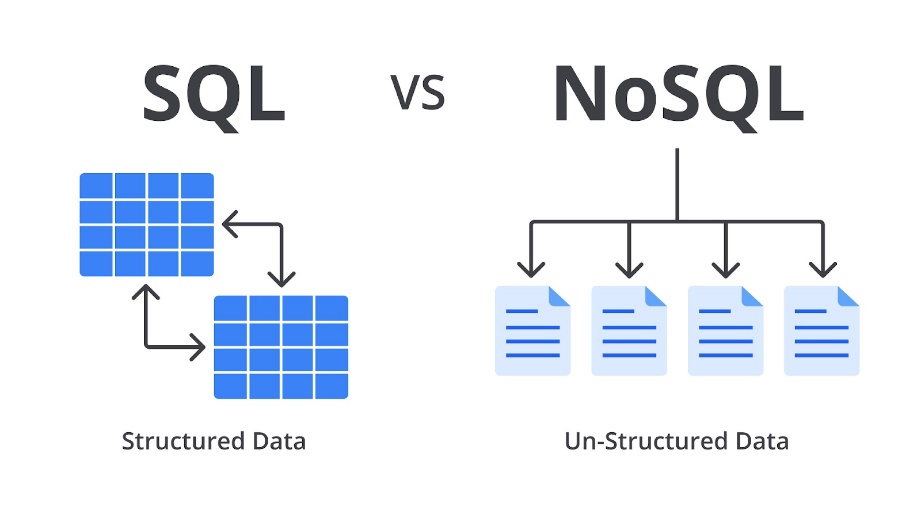
Agar Anda lebih memahami cara kerja antara basis data dan dataset, perhatikan gambar di bawah ini.



Sederhananya, gambar tabung di atas merupakan representasi dari sebuah basis data. Tabel yang terdapat di dalamnya adalah struktur dalam basis data yang diatur dengan bentuk kolom dan baris, serta basis data juga terdiri dari data-data yang saling terhubung seperti berikut.

Namun, apabila tabel-tabel tersebut dibungkus dengan menggunakan format tertentu, ia disebut sebuah dataset. Sederhananya adalah basis data terdiri dari dataset yang saling berhubungan satu sama lainnya.

**Rupa-Rupa Sistem Basis Data**



Perlu Anda ketahui bahwa SQL (Structured Query Language) dikembangkan oleh IBM pada tahun 1970 dengan tujuan untuk memanipulasi dan menarik data yang tersimpan pada IBM *database management system*(DBMS) yang disebut System R.

Seiring dengan perkembangan teknologi, di awal abad 21, muncul teknologi baru yang dikenal dengan NoSQL (No SQL atau Not Only SQL). Teknologi ini dipicu oleh kebutuhan-kebutuhan perusahaan Web 2.0 yang mengacu pada website dengan titik berat pada user, penggunaan yang mudah, dan sesuai dengan sistem yang digunakan oleh end user. Penasaran lebih untuk mengetahui kelanjutannya? Yuk simak penjelasannya di bawah ini.

**SQL (Structured Query Language)**

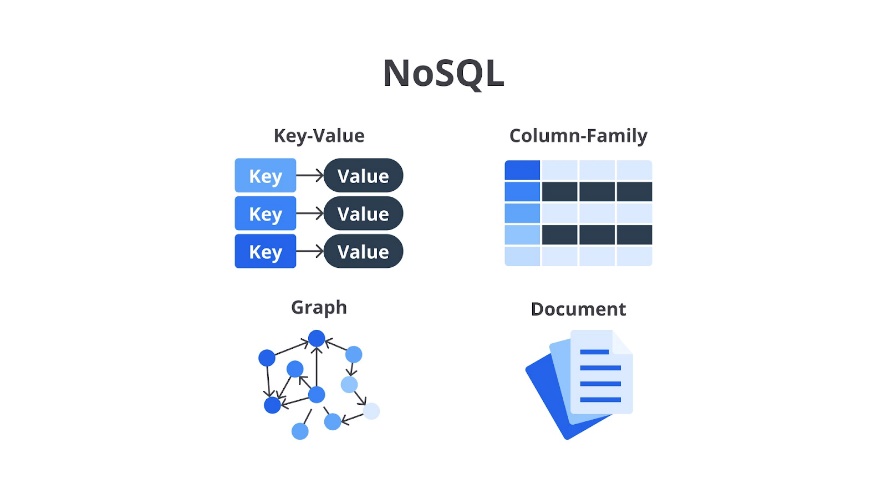
SQL adalah bahasa pemrograman untuk menyimpan dan memproses informasi dalam basis data relasional. Basis data relasional menyimpan informasi dalam bentuk tabel dengan baris dan kolom yang mewakili atribut data berbeda dan berbagai hubungan antara nilai data. Anda dapat menggunakan pernyataan SQL untuk menyimpan, memperbarui, menghapus, mencari, dan mengambil informasi dari basis data. Anda juga dapat menggunakan SQL untuk mempertahankan dan mengoptimalkan kinerja basis data.

Dalam penggunaan SQL ada beberapa jenis perintah yang digunakan diantaranya DDL (*Data Definition Language*) dan DML (*Data Manipulation Language*). Mari kita bahas satu per satu.

DDL merupakan subperintah pada SQL yang dimanfaatkan guna membangun kerangka basis data. Tabel di atas merupakan pernyataan *statements* yang utama dari DDL.

DML merupakan subperintah pada SQL yang dimanfaatkan dalam manipulasi basis data yang sudah dibuat. Pada dasarnya, dalam DML terdapat empat perintah berikut beserta fungsinya.

**NoSQL (Not Only SQL)**



NoSQL atau disebut juga sebagai “*not only SQL”*, “*non-SQL”*, adalah pendekatan untuk desain basis data yang memungkinkan penyimpanan dan kueri data di luar struktur tradisional yang ditemukan dalam basis data relasional. Keputusan untuk menggunakan basis data relasional dan non-relasional sebagian besar bersifat kontekstual dan bervariasi tergantung pada kasus penggunaan.

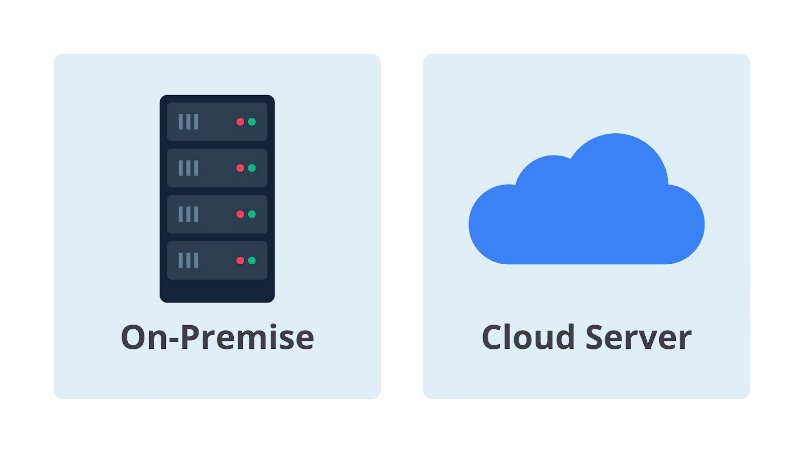
Pada dasarnya basis data NoSQL menyimpan data dalam satu struktur data seperti dokumen JSON. Karena desain basis data non-relasional ini tidak memerlukan skema, ia menawarkan skalabilitas cepat untuk mengelola kumpulan data besar dan tidak terstruktur.

Basis data NoSQL umumnya diklasifikasikan ke dalam empat kategori utama, di antaranya

1. **Document Databases**  
   Basis data ini menyimpan data sebagai dokumen semi-terstruktur, seperti JSON atau XML, dan dapat dikueri menggunakan bahasa kueri berorientasi dokumen.
2. **Key-value Stores**  
   Basis data ini menyimpan data sebagai *key-value pairs*dan dioptimalkan untuk operasi baca atau tulis yang sederhana dan cepat.
3. **Column-family stores**  
   Basis data ini menyimpan data sebagai keluarga kolom, yaitu kumpulan kolom yang diperlakukan sebagai satu kesatuan. Mereka dioptimalkan untuk kueri data dalam jumlah besar yang cepat dan efisien.
4. **Graph Databases**  
   Sempat dibahas pada materi sebelumnya, bahwa basis data ini menyimpan data sebagai *node*dan *edge*, serta dirancang untuk menangani hubungan kompleks antar data.

Basis data NoSQL sering digunakan dalam aplikasi yang membutuhkan banyak data untuk diproses dan dianalisis secara *real-time*, seperti analitik media sosial (*Social Media Analytics*), *e-commerce*, dan game. Mereka juga dapat digunakan untuk aplikasi lain, seperti sistem manajemen konten, manajemen dokumen, dan manajemen hubungan pelanggan.

**Database on Premise vs. Cloud**



*On premise*atau pusat data lokal adalah sekelompok server yang Anda miliki dan kendalikan secara pribadi, sedangkan *on cloud*melibatkan penyewaan sumber daya pusat data dari penyedia layanan pihak ketiga.

**Basis Data on Premise**

*On-premise*adalah jenis server berupa *software*yang dijalankan secara internal oleh tim IT perusahaan. Tim IT tersebut bertugas dalam menjalankan aplikasi server, memasang sistem operasi, dan meletakkan server di sebuah gedung. Dapat dibilang *on-premise*adalah sebuah vendor, yakni pemilik layanan *on-premise*memberikan lisensi penuh kepada tim tersebut untuk mengelola basis data di dalamnya. Jadi, *on-premise*bukanlah jaringan server secara jarak jauh atau *remote*atau secara sederhananya basis data *on-premise*disimpan di dalam memori internal komputer dari sebuah vendor.

**Basis Data on Cloud**

Berbeda dengan *on-premise, cloud server*alias *cloud computing*merupakan layanan infrastruktur berupa penyimpanan basis data, server, jaringan, dan *software*berbasis internet. Layanan ini tidak disediakan oleh tim IT perusahaan, tetapi melalui pihak ketiga yang merupakan penyedia layanan *cloud.* Pada layanan cloud server, basis data disimpan seluruhnya dalam server berbasis internet. Artinya, cloud server tidak memerlukan penyimpanan lokal atau offline terbatas. Salah satu contoh layanan basis data cloud server adalah Google Cloud databases. Perusahaan bisa memilih kapasitas cloud sesuai anggaran dan kebutuhan dalam penyimpanan data.

**Perangkat Pengolahan Data**

Pada modul pertama, Anda telah mempelajari sedikit banyak tentang data, tetapi belum menyelami terkait *tools*pengolahan data. Berikut merupakan beberapa *tools*yang umum digunakan.

**Excel dan Google Sheets**

**SPSS (*Statistical Package for the Social Sciences*)**

SPSS adalah program komputer yang dipakai untuk analisis statistika. SPSS digunakan untuk pengolahan dan menganalisis data yang memiliki kemampuan analisis statistik serta sistem manajemen data dengan lingkungan grafis.

**Perangkat Visualisasi Data**

Mari kita *recall*bahwa tujuan dari visualisasi data adalah untuk mempermudah komunikasi, mempercepat pengambilan keputusan, memperkaya perspektif tim, dan mengidentifikasi masalah. Maka dari itu, Anda akan berkenalan dengan Tableau, Metabase, dan Looker.

**Tableau**

[Direktorat Teknologi Informasi ITB](https://dti.itb.ac.id/tableau/) menuturkan bahwa Tableau adalah platform *Business Intelligence* untuk membantu dalam melihat dan memahami data. Platform ini merupakan hasil dari proyek ilmu komputer di Stanford yang didirikan pada tahun 2003. Tujuan dibuatnya Tableau adalah untuk meningkatkan aliran analisis dan membuat data lebih dapat diakses oleh orang-orang melalui visualisasi.

**Metabase**

Salah satu tools (*open source*) yang digunakan untuk visualisasi data adalah Metabase. Aplikasi ini menggunakan konsep *ask questions and display answers,*yakni memilih cara untuk mengajukan pertanyaan kemudian menjawab pertanyaan tersebut dengan penyajian data yang ada di metabase antara lain *Numbers, Trend, Progress bar, Gauge, Table, Line chart, Bar chart,*dll.

**Looker Studio**

*Last but not least, tools*visualisasi terakhir yang umum digunakan oleh *data analyst, data scientist, business analyst,*dan lain-lain. [Looker Studio](https://lookerstudio.google.com/) atau sebelumnya Google Data Studio adalah platform data yang menggunakan bahasa pemodelan sederhana yang disebut LookML yang memungkinkan tim data untuk menentukan hubungan di basis data mereka sehingga pengguna bisnis dapat menjelajahi, menyimpan, dan mengunduh data tanpa harus mengetahui SQL.

**Bahasa Pemrograman Data Science**

*“The limits of my language mean the limits of my world”*

*- Ludwig Wittgenstein -*

**Berkelana dengan Python**

Python adalah bahasa pemrograman yang banyak digunakan dalam aplikasi web, pengembangan perangkat lunak, ilmu data, dan *machine learning* (ML). Developer menggunakan Python karena efisien dan mudah dipelajari serta dapat dijalankan berbagai platform, juga merupakan *open source.*

Para data scientist menggunakan Python untuk mengerjakan proyek data science seperti berikut.

1. Memperbaiki dan menghapus data yang salah atau biasa disebut dengan pembersihan data.
2. Mengekstraksi dan memilih berbagai fitur data.
3. Pelabelan data (*data labeling*), yaitu menambahkan nama yang bermakna untuk data.
4. Melakukan proses analisis data.
5. Memvisualisasikan data dengan menggunakan diagram dan grafik seperti diagram garis, grafik batang, histogram, dan diagram lingkaran.
6. Mengembangkan model machine learning.

Tahukah Anda? *Mostly,*banyak praktisi machine learning menggunakan Python untuk mengembangkan model guna mengerjakan berbagai tugas seperti klasifikasi citra, teks, dan pengenalan wajah (*image processing*).

**Berkenalan dengan R**

Kemudahan dalam penggunaannya serta mudahnya akses dukungan komunitas, membuat R menjadi salah satu bahasa pemrograman paling populer di dunia. R memiliki aturan atau sintaks yang berbeda dengan bahasa pemrograman lain yang membuatnya memiliki ciri khusus tersendiri.

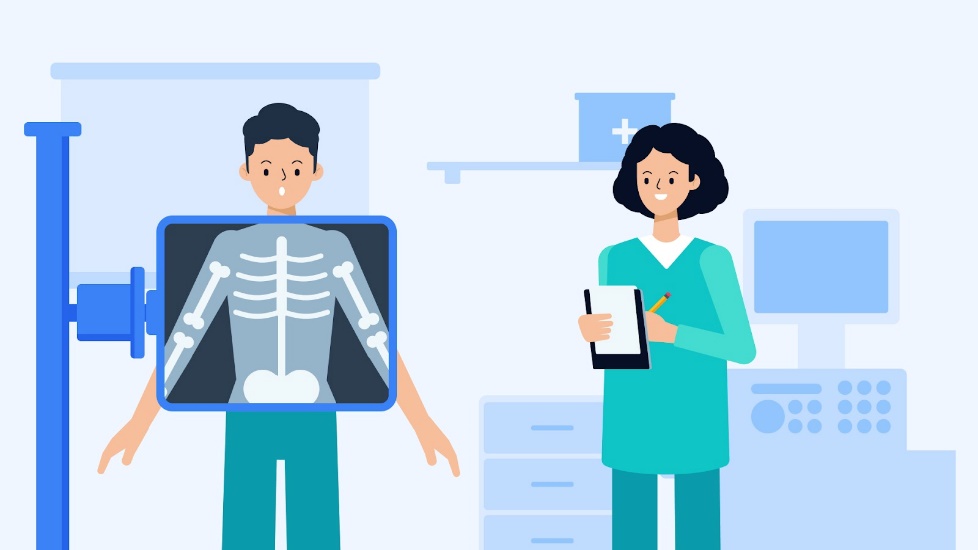
Berikut merupakan ciri dan fitur pada R.

1. Memiliki sifat yang *casesensitive.*Dalam proses input, huruf besar dan kecil sangat diperhatikan
2. Segala sesuatu akan dianggap sebagai objek. Konsep objek ini pada dasarnya sama dengan bahasa pemrograman berbasis objek lainnya, yaitu Python, Java, dan C++. Perbedaannya adalah bahasa pemrograman R relatif lebih sederhana dibandingkan bahasa pemrograman berbasis objek yang lain.
3. Mendukung proses loop, decision making, dan menyediakan berbagai jenis operator (aritmatika, logika, dll).
4. Menyediakan berbagai fungsi untuk keperluan visualisasi data. Visualisasi data pada R dapat menggunakan paket bawaan atau paket lain seperti ggplot2, ggvis, dll.

Dari kedua perbandingan bahasa pemrograman yang populer digunakan untuk data science, dapat disimpulkan bahwa perbedaan yang sangat terlihat dari kedua bahasa ini adalah R digunakan untuk analisis statistik, sedangkan Python lebih general. Python adalah bahasa multipurpose, sama seperti C++ dan Java. Penggunaannya pun cenderung lebih mudah dipelajari.

**Teknologi Machine Learning**

*Machine learning* adalah ilmu pengembangan algoritma yang memanfaatkan konsep matematis dan statistik dalam menjalankan tugas tertentu tanpa instruksi eksplisit. Pada prosesnya, *machine learning* akan berusaha mengenal pola yang terdapat dalam sebuah data serta menggunakannya untuk menghasilkan prediksi.



Gambar di atas merupakan salah satu contoh penerapan *machine learning* yang lain, yaitu para data scientist dapat melatih model untuk mendiagnosis kanker dari gambar sinar-x. Hal ini dilakukan dengan cara melatih model dengan ribuan hingga jutaan gambar yang dipindai dengan diagnosis yang sesuai. Setelah proses pelatihan selesai, model akan mengenali pola dalam gambar tersebut. Lalu, model akan menggunakan pola tersebut untuk membantu para praktisi medis dalam mendiagnosa penyakit kanker.

Lalu, apa hubungannya antara machine learning dan data science?

Dasarnya, *machine learning*adalah salah satu alat atau pendekatan yang digunakan dalam data science. Secara garis besar, kita dapat mengelompokkan *machine learning* ke dalam tiga kategori yaitu seperti berikut.

1. *Supervised learning*. Pendekatan *machine learning*yang menggunakan kumpulan data berlabel (*labeled datasets*). Kumpulan data ini dirancang untuk melatih atau “mengawasi” algoritma machine learning. Pada proses pelatihan inilah model belajar memahami pola yang terdapat dalam data dan labelnya. Pola ini selanjutnya akan digunakan untuk mengerjakan tugas yang spesifik seperti klasifikasi atau regresi. Klasifikasi digunakan untuk memprediksi kemungkinan sesuatu termasuk dalam kategori tertentu sedangkan regresi digunakan untuk memprediksi nilai numerik dari banyak kemungkinan nilai. Salah satu contoh penerapan *supervised learning*adalah *face recognition*yang terdapat pada *smartphone,*laptop, bahkan hingga pintu sekalipun.
2. *Unsupervised learning.*Jika supervised learning memanfaatkan data dan juga label untuk melatih model, unsupervised learning hanya membutuhkan data (tanpa label) untuk melatih model. Pada prosesnya model akan belajar untuk mengidentifikasi pola tersembunyi yang terdapat dalam data. Pola ini selanjutnya akan digunakan untuk mengerjakan tugas yang spesifik, seperti membagi data ke dalam beberapa klaster, mendeteksi anomali dalam data, dll. Salah satu contoh penerapan un*supervised learning*adalah adanya rekomendasi pada sebuah media sosial yang *relate*dengan *favorit*dan yang sering Anda buka pada laman tersebut.
3. Reinforcement learning. Salah satu penerapan reinforcement learning adalah pada algoritma AlphaGo. AlphaGo adalah program komputer yang digunakan untuk menjalankan board game. Teknologi yang digunakan pada board game tersebut dikembangkan oleh DeepMind Technologies yang diakuisisi oleh Google. Pada Oktober 2015, permainan tersebut mampu mengalahkan pemain manusia profesional pada papan berukuran 19x19. Hal tersebut diakibatkan oleh program AlphaGo yang semakin canggih. Board game tersebut menggunakan reinforcement learning karena metode tersebut mampu meniru perilaku yang diinginkan. Pada board game tersebut, AlphaGo telah mempelajari lebih dari 30 gerakan manusia yang dapat ditiru sehingga seiring berjalannya waktu AlphaGo dapat mengalahkan permainan manusia. Cukup menarik, bukan?

**Teknologi Pendukung GitHub**

*First thing first,*GitHub adalah platform hosting kode untuk versi kontrol dan kolaborasi. Hal ini memungkinkan Anda bekerja sama dengan orang lain dalam proyek yang berbeda. Mengapa menggunakan GitHub? GitHub adalah *software* yang memungkinkan kontrol proyek melalui alat yang dikenal sebagai Git. Ini memungkinkan pengguna untuk mendistribusikan kode, berkolaborasi dalam proyek, membantu penerapan proyek melalui CI/CD (Integrasi berkelanjutan/penerapan berkelanjutan), kembali ke versi proyek sebelumnya, dll.

Para data scientist membutuhkan GitHub untuk manajemen sumber kode. Hal itu menampung Git, sistem kontrol versi sumber terbuka (*open-source version control*) yang melacak perubahan dan permintaan proyek. Menggunakan GitHub, pengguna dapat mengkloning kode dari repositori pusat ke mesin lokal mereka, membuat perubahan, melakukan modifikasi, dan menggabungkannya kembali ke repository pusat.

Sebagai data scientist, Anda memerlukan GitHub untuk mengumpulkan data dari berbagai sumber dan memodifikasi atau mengimplementasikan perubahan ke dalam file proyek yang ada. Beberapa *developer*dan manajer lain dapat meninjau perubahan dan juga melihat modifikasi yang ada. Berikut merupakan beberapa praktik terbaik, meliputi:

* Menyimpan semua data, file proyek, dan model Anda di suatu tempat;
* Melacak perubahan dan versi proyek secara lokal menggunakan Git; dan
* Menyimpan model pembelajaran mesin dan melakukan analitik menggunakan kode atau alat lain seperti Tableau atau PowerBI. Ini akan memungkinkan pengguna untuk menerapkan proyek melalui metodologi CI/CD.