**Mempersiapkan Data untuk Eksplorasi**

**Eksplorasi Data**

1. **Memahami tipe dan struktur data**: Kita semua menghasilkan banyak data dalam kehidupan kita sehari-hari. Di bagian materi ini, Anda akan melihat bagaimana kita menghasilkan data dan bagaimana seorang analis memutuskan data mana yang akan dikumpulkan untuk dianalisis. Anda juga akan belajar tentang data terstruktur dan tidak terstruktur, tipe data, dan format data selagi Anda mulai memikirkan cara menyiapkan data untuk eksplorasi.
2. **Memahami bias, kredibilitas, privasi, etika, dan akses**: Ketika analis data bekerja dengan data, mereka selalu memeriksa apakah data itu kredibel dan tidak bias. Di bagian materi ini, Anda akan belajar bagaimana mengidentifikasi berbagai jenis bias dalam data dan bagaimana memastikan kredibilitas data Anda. Anda juga akan mengeksplorasi open data dan hubungan antara etika data dan privasi data serta pentingnya hubungan tersebut.
3. **Database**: Di mana data berada: Saat Anda menganalisis data, Anda akan mengakses banyak data dari sebuah database. Di situlah data berada. Di bagian materi ini, Anda akan mempelajari semua hal tentang database, termasuk cara mengaksesnya dan mengekstrak, memfilter, dan mengurutkan data yang ada di dalamnya. Anda juga akan memeriksa metadata untuk menemukan berbagai tipe data dan bagaimana para analis menggunakannya.
4. **Mengatur dan melindungi data Anda**: Keterampilan organisasi/pengaturan yang baik adalah hal penting dari hampir semua pekerjaan, dan analitik data tidak berbeda. Di bagian materi ini, Anda akan mempelajari praktik terbaik untuk mengatur data dan menjaganya agar tetap aman. Anda juga akan mempelajari bagaimana para analis menggunakan konvensi penamaan file untuk membantu mereka menjaga pekerjaan mereka tetap teratur.

**Mengumpulkan data-Memilih Data Yang Tepat**

* **Bagaimana data akan dikumpulkan**

Putuskan apakah Anda akan mengumpulkan data menggunakan sumber daya Anda sendiri atau menerimanya (dan mungkin membelinya) dari pihak lain. Data yang Anda kumpulkan sendiri disebut data pihak pertama.

* **Sumber data**

Jika Anda tidak mengumpulkan data menggunakan sumber daya sendiri, Anda mungkin mendapatkan data dari pihak kedua atau penyedia data pihak ketiga. Data pihak kedua dikumpulkan langsung oleh kelompok lain dan kemudian dijual. Data pihak ketiga dijual oleh penyedia yang tidak mengumpulkan data itu sendiri. Data pihak ketiga mungkin berasal dari sejumlah sumber yang berbeda.

* **Memecahkan masalah bisnis Anda**

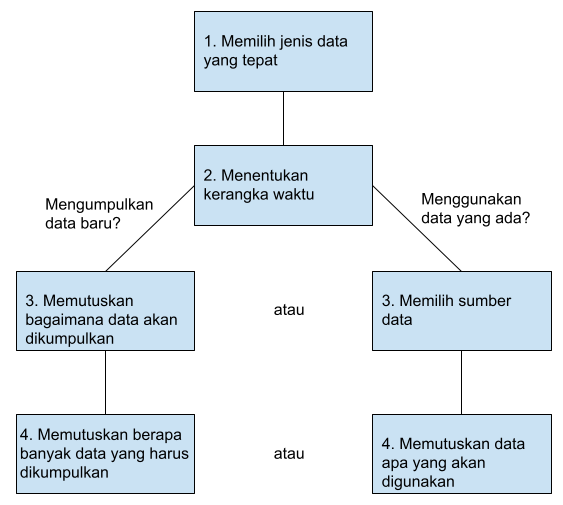
Dataset dapat menampilkan banyak informasi menarik. Tetapi pastikan untuk memilih data yang benar-benar dapat membantu memecahkan masalah Anda. Misalnya, jika Anda menganalisis tren dari waktu ke waktu, pastikan Anda menggunakan data dengan urutan waktu - dengan kata lain, data yang menyertakan tanggal.

* **Berapa banyak data yang harus dikumpulkan**

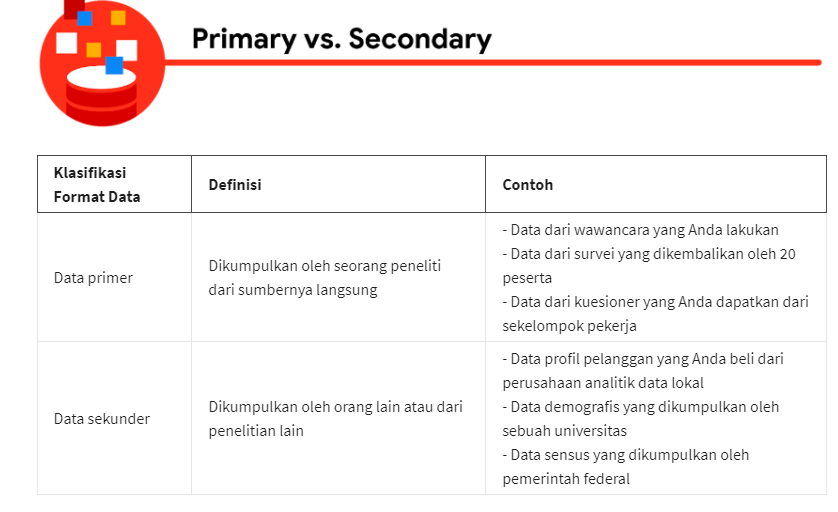
Jika Anda mengumpulkan data Anda sendiri, buat keputusan ukuran sampel yang beralasan. Sampel acak dari data yang ada mungkin bisa digunakan untuk beberapa proyek tertentu. Proyek lain mungkin memerlukan pengumpulan data yang lebih strategis untuk memenuhi kriteria tertentu. Setiap proyek memiliki kebutuhannya masing-masing.

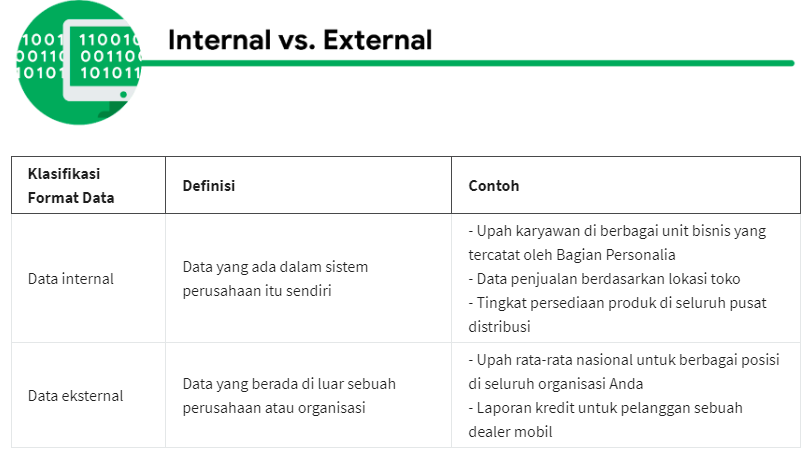
* **Jangka waktu**

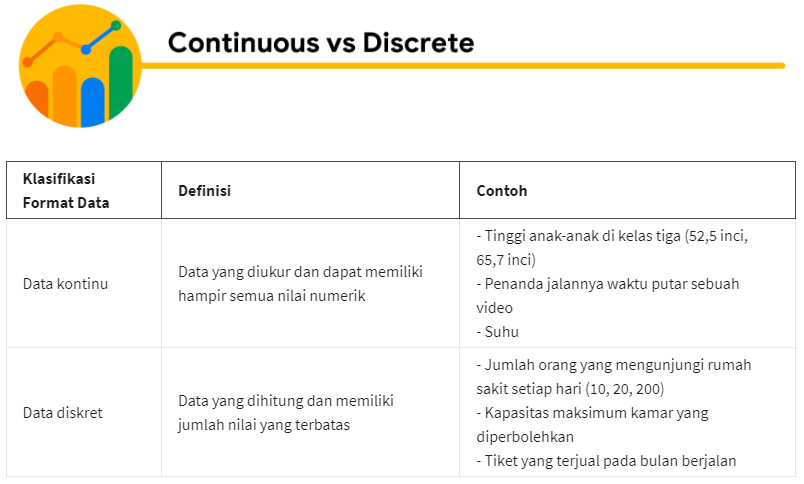
Jika Anda mengumpulkan data Anda sendiri, putuskan berapa lama Anda perlu mengumpulkannya, terutama jika Anda melacak tren dalam jangka waktu yang lama. Jika Anda membutuhkan jawaban segera, Anda mungkin tidak punya waktu untuk mengumpulkan data baru. Dalam hal ini, Anda perlu menggunakan data historis yang sudah ada.

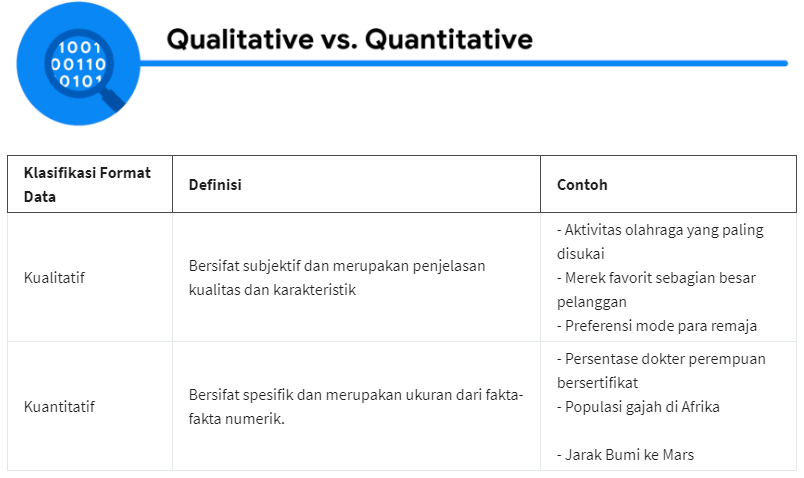


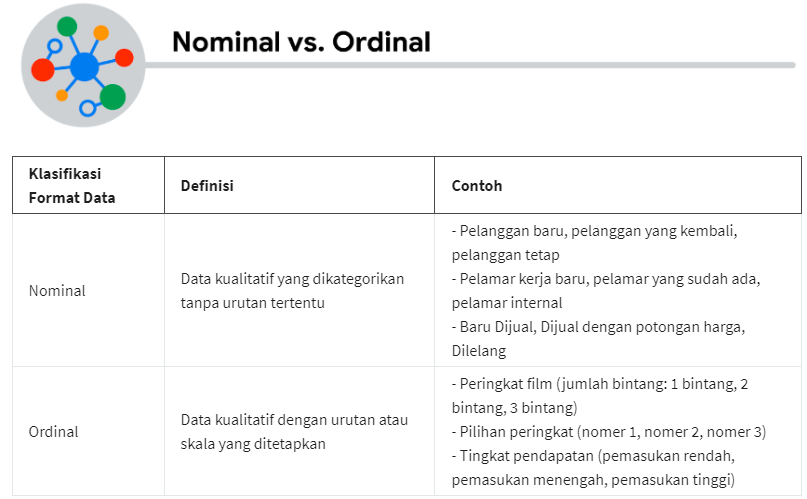
**Format data dalam praktik**

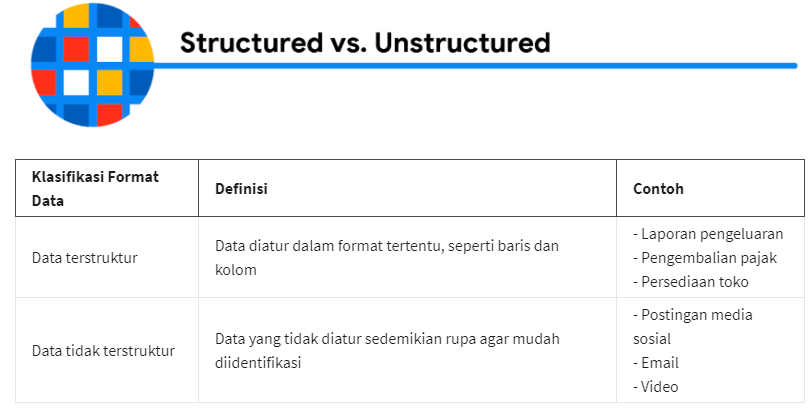








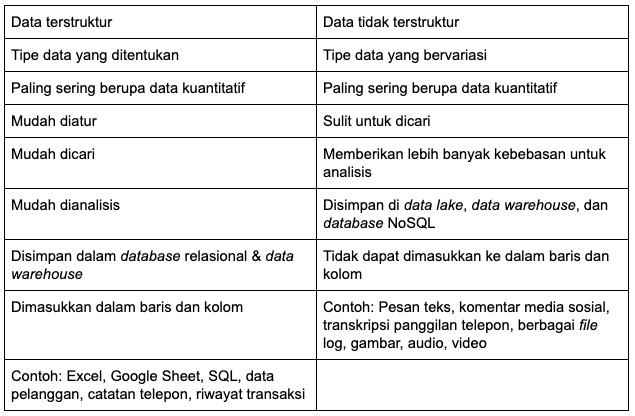




**Struktur data**

Data ada di mana-mana dan dapat disimpan dalam banyak cara. Dua kategori data secara umum adalah:

* **Data terstruktur:** Data diatur dalam format tertentu, seperti baris dan kolom**.**
* **Data tidak terstruktur:** Tidak diatur dengan cara yang mudah diidentifikasi.

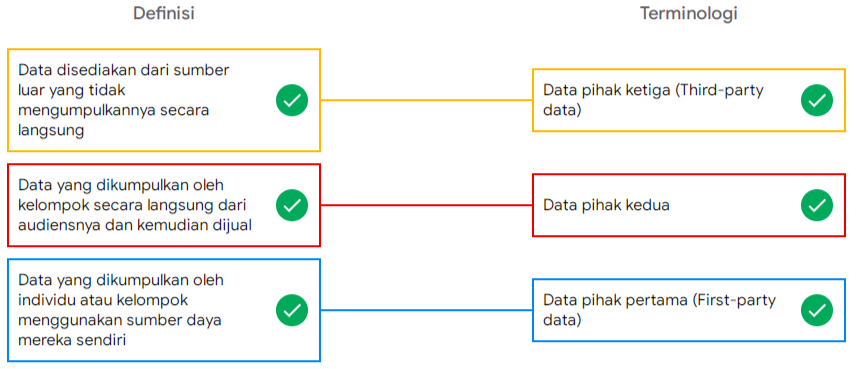


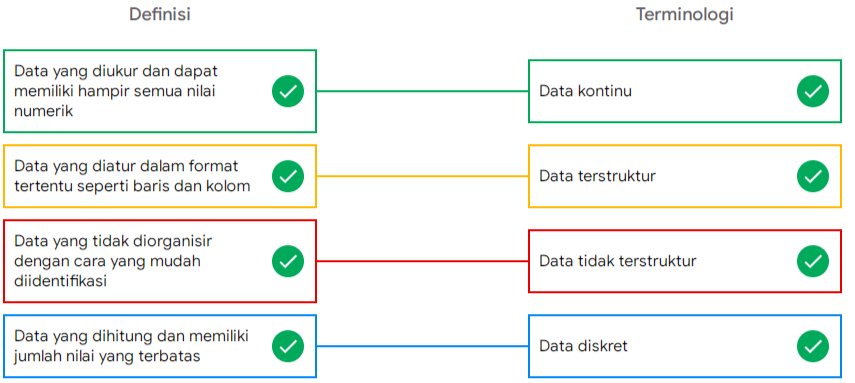
**Data terstruktur**

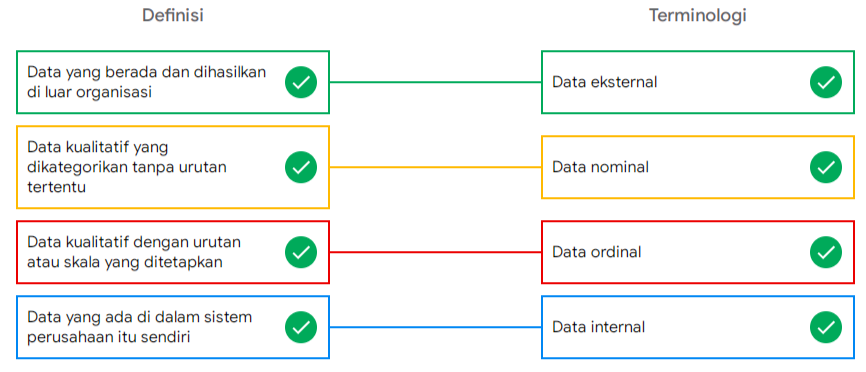
Seperti yang telah dijelaskan sebelumnya, data terstruktur diatur dalam format tertentu. Sehingga membuatnya lebih mudah disimpan dan dibuatkan query untuk memenuhi berbagai kebutuhan bisnis. Jika data itu diekspor, strukturnya akan mengikuti ke mana data itu pergi.

**Data tidak terstruktur**

Data yang tidak terstruktur Data yang tidak diatur sedemikian rupa agar mudah diidentifikasi. Dan jumlah data tidak terstruktur jauh lebih banyak daripada jumlah data terstruktur di dunia. File video dan audio, file teks, konten media sosial, citra satelit, presentasi, file PDF, hasil survei terbuka, dan situs web semuanya memenuhi kualifikasi sebagai jenis data tidak terstruktur.





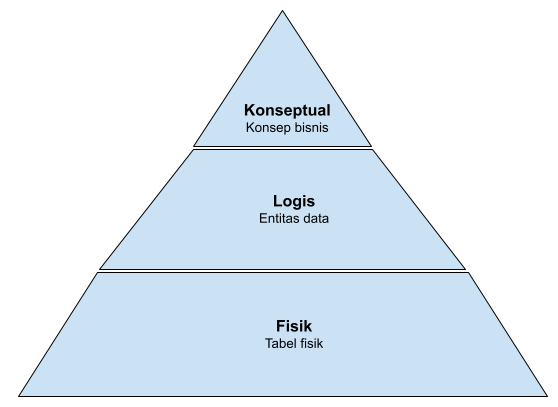


**Apa itu pemodelan data?**

Pemodelan data adalah proses pembuatan diagram yang secara visual menggambarkan bagaimana data diatur dan terstruktur. Representasi visual ini disebut model data. Anda dapat menganggap pemodelan data sebagai cetak biru sebuah rumah. Satu saat, akan ada tukang listrik, tukang kayu, dan tukang ledeng yang menggunakan cetak biru itu. Masing-masing tukang ini memiliki kepentingan yang berbeda dengan cetak biru itu, tetapi mereka semua membutuhkannya untuk memahami struktur keseluruhan rumah. Model data juga serupa; pengguna yang berbeda mungkin memiliki kebutuhan data yang berbeda, tetapi model data memberi mereka pemahaman tentang struktur secara keseluruhan.

**Tingkat pemodelan data**

Setiap tingkat pemodelan data memiliki tingkat detail yang berbeda.



* **Pemodelan data konseptual** memberikan pandangan tentang struktur data dari atas, seperti bagaimana data berinteraksi di seluruh organisasi. Misalnya, model data konseptual dapat digunakan untuk mendefinisikan persyaratan bisnis untuk sebuah database baru. Model data konseptual tidak berisi detail teknis.
* **Pemodelan data logis** berfokus pada detail teknis sebuah database seperti relasi, atribut, dan entitas. Misalnya, suatu model data logis mendefinisikan bagaimana catatan individual diidentifikasi secara unik dalam sebuah database. Tapi itu pemodelan data logis tidak menguraikan nama sebenarnya dari tabel database. Itu adalah tugas model data fisik.
* **Pemodelan data fisik** menggambarkan bagaimana sebuah database beroperasi. Model data fisik mendefinisikan semua entitas dan atribut yang digunakan; termasuk misalnya, nama tabel, nama kolom, dan tipe data untuk database itu.

**Teknik pemodelan data**

Ada banyak pendekatan untuk mengembangkan model data, tetapi dua metode umum adalah diagram Entity Relationship Diagram (ERD) dan Unified Modeling Language (UML). ERD adalah cara visual untuk memahami relasi antara entitas dalam model data. Diagram UML adalah diagram yang sangat rinci yang menggambarkan struktur sistem dengan menunjukkan entitas sistem, atribut, operasi, dan relasi mereka. Sebagai analis data junior, Anda perlu memahami bahwa ada teknik pemodelan data yang berbeda, tetapi dalam praktiknya, Anda mungkin akan mengikuti teknik yang ada di organisasi Anda .

**Analisis data dan pemodelan data**

Pemodelan data dapat membantu Anda mengeksplorasi garis secara detail dan bagaimana kaitannya dengan seluruh sistem informasi sebuah organisasi. Pemodelan data terkadang memerlukan analisis data untuk memahami bagaimana data disatukan; dengan begitu, Anda mengetahui cara memetakan data. Dan akhirnya, model data memudahkan semua orang di organisasi Anda untuk memahami dan berkolaborasi bersama Anda dalam data.

**Memahami logika Boolean**

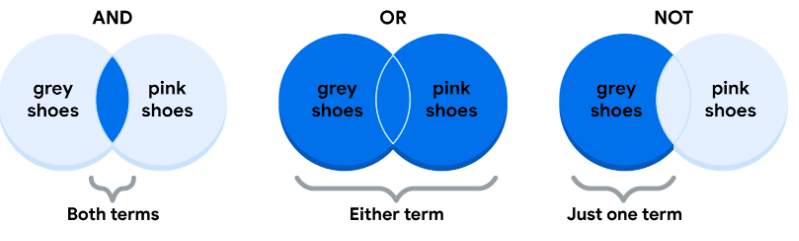
Kondisi ini dibuat dengan operator Boolean, termasuk AND, OR, dan NOT. Operator ini mirip dengan operator matematika dan dapat digunakan untuk membuat pernyataan logis yang menyaring hasil Anda. Analis data menggunakan pernyataan Boolean untuk melakukan berbagai tugas analisis data, seperti membuat query untuk pencarian dan memeriksa kondisi saat menulis kode pemrograman.

**Contoh logika Boolean**

Bayangkan Anda berbelanja sepatu, dan sedang mempertimbangkan pilihan tertentu:

* Anda hanya akan membeli sepatu berwarna merah muda dan abu-abu.
* Anda akan membeli sepatu jika warnanya merah muda atau abu-abu saja, atau jika berwarna merah muda dan abu-abu.
* Anda akan membeli sepatu jika warnanya abu-abu, tetapi tidak jadi membeli jika berwarna merah muda.

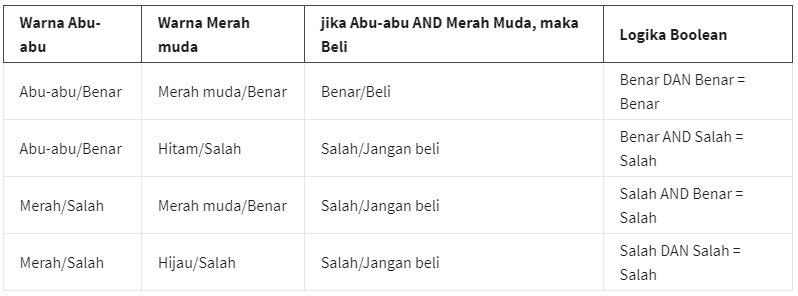
Di bawah ini adalah diagram Venn yang menggambarkan preferensi ini. AND adalah pusat diagram Venn, di mana dua kondisi tumpang tindih. OR menerima semua kondisi. NOT adalah bagian diagram Venn yang tidak berisi bagian yang dikecualikan.



**Operator AND**

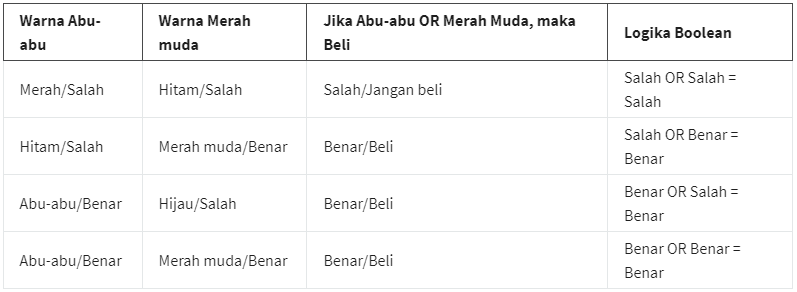
Kondisi Anda adalah "Jika warna sepatu memiliki kombinasi abu-abu dan merah muda, Anda akan membelinya." Pernyataan Boolean akan mengurai logika pernyataan itu untuk menyaring hasil yang memenuhi kondisi Anda berdasarkan kedua warna itu. Pernyataannya akan mengatakan "IF (Warna = "Abu-abu") AND (Warna = "Merah muda") maka Anda membelinya.” Operator AND atau “DAN” memungkinkan Anda menggabungkan beberapa kondisi.

Di bawah ini adalah tabel sederhana yang menguraikan logika Boolean yang dijalankan dalam pernyataan ini. Di kolom Warna Abu-abu, ada dua pasang sepatu yang memenuhi warna yang diinginkan. Dan di kolom Warna Merah muda ada dua pasang sepatu yang juga memenuhi kondisi itu. Namun di kolom jika Abu-abu AND Merah muda hanya ada satu pasang sepatu yang memenuhi kedua kondisi tersebut. Jadi, menurut logika Boolean dari pernyataan itu, hanya ada satu pasangan yang ditandai benar. Dengan kata lain, hanya ada sepasang sepatu yang bisa Anda beli.



**Operator OR**

Operator OR atau “ATAU” memungkinkan Anda mendapatkan hasil jika salah satu dari dua kondisi Anda terpenuhi. Kondisi Anda adalah "Jika sepatu berwarna abu-abu atau merah muda, Anda akan membelinya." Pernyataan Boolean adalah "IF (Warna="Abu-abu") OR (Warna="Merah muda") Anda akan membelinya.” Perhatikan bahwa setiap sepatu yang memenuhi kondisi Warna abu-abu atau Warna merah muda ditandai sebagai benar oleh logika Boolean. Menurut tabel kebenaran di bawah ini, ada tiga pasang sepatu yang bisa Anda beli.



**Operator NOT**

Akhirnya, operator NOT atau “BUKAN” memungkinkan Anda menyaring jawaban yang tidak memenuhi kondisi tertentu. Kondisi Anda adalah "Anda akan membeli sepatu abu-abu kecuali sepatu yang ada warna merah mudanya." Pernyataan Boolean Anda adalah "IF (Warna="Abu-abu") AND (Warna=NOT "Merah muda") Anda akan membelinya." Sekarang, semua sepatu abu-abu yang tidak ada warna merah mudanya ditandai benar oleh logika Boolean untuk kondisi NOT atau BUKAN Merah muda. Sepatu merah muda ditandai salah oleh logika Boolean untuk kondisi BUKAN Merah muda. Hanya satu pasang sepatu yang dikecualikan dalam tabel kebenaran di bawah ini.



**Keampuhan penggunaan berbagai kondisi sekaligus**

Untuk analis data, kekuatan sebenarnya dari logika Boolean berasal dari kemampuannya menggabungkan beberapa kondisi dalam satu pernyataan. Misalnya, jika Anda ingin menyaring hasil sepatu yang berwarna abu-abu atau merah muda, dan tahan air, Anda dapat membuat pernyataan Boolean seperti: “IF ((Warna = ”Abu-abu”) OR (Warna = “Merah muda”)) AND (Tahan air=”Benar”).” Perhatikan bahwa Anda menggunakan tanda kurung untuk mengelompokkan kondisi-kondisi Anda.

**Mentransformasi data**

Transformasi data adalah proses mengubah format, struktur, atau nilai data. Sebagai analis data, kemungkinan besar Anda perlu melakukan transformasi data di saat-saat tertentu untuk memudahkan Anda menganalisanya.

Transformasi data biasanya melakukan:

* Menambahkan, menyalin, atau mereplikasi data
* Menghapus field atau record
* Standarisasi nama variabel
* Mengganti nama, memindahkan, atau menggabungkan kolom di sebuah database
* Menggabungkan satu set data dengan set data lainnya
* Menyimpan file dalam format yang berbeda. Misalnya, menyimpan spreadsheet sebagai file berbentuk comma separated values (CSV).

**Mengapa mengubah data?**

Tujuan untuk transformasi data di antaranya:

* **Organisasi data**: data terorganisir lebih baik lebih mudah digunakan
* **Kompatibilitas data:** aplikasi atau sistem yang berbeda dapat menggunakan data yang sama
* **Migrasi data**: data dengan format yang sesuai dapat dipindahkan dari satu sistem ke sistem lainnya
* **Penggabungan data**: data dengan pola pengaturan yang sama dapat saling bergabung
* **Peningkatan data:** data dapat ditampilkan dengan field yang lebih rinci
* **Perbandingan data:** data yang memiliki spesifikasi sama kemudian dapat dibandingkan

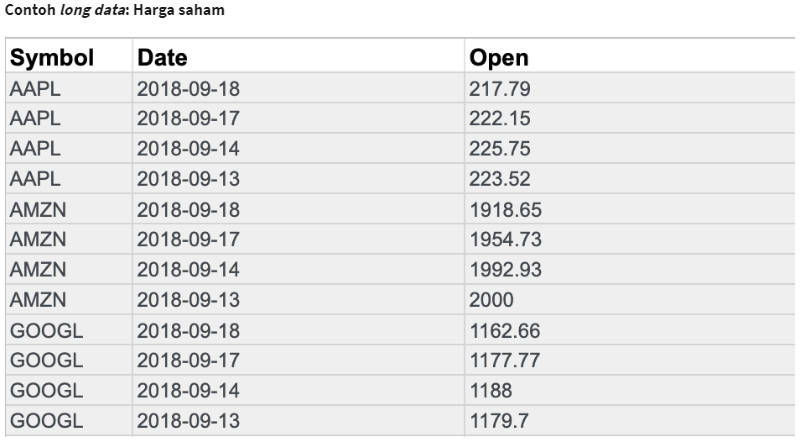
**Contoh transformasi data: penggabungan data**

Mario adalah tukang ledeng yang memiliki perusahaan pipa ledeng. Setelah bertahun-tahun menggeluti bisnis ini, ia membeli perusahaan pipa ledeng lain. Mario ingin menggabungkan informasi pelanggan dari perusahaannya yang baru diakuisisi dengan miliknya sendiri, tetapi perusahaan itu menggunakan database yang berbeda. Jadi, Mario perlu membuat datanya kompatibel. Untuk melakukan ini, ia harus mengubah format data perusahaan yang diakuisisi itu. Kemudian, ia harus menghapus baris-baris duplikat yang berisi pelanggan yang sama-sama dimiliki kedua perusahaan itu. Ketika data kompatibel dan bersatu, perusahaan pipa Mario akan memiliki gabungan database pelanggan yang lengkap.

**Contoh transformasi data: organisasi data (dari long ke wide)**

Agar Anda lebih mudah membuat grafik, Anda mungkin perlu mengubah long data ke wide data. Pertimbangkan contoh transformasi data harga saham (dikumpulkan sebagai long data) ke wide data berikut ini.

**`** adalah data di mana setiap baris berisi satu titik data untuk sebuah item tertentu. Pada contoh long data di bawah ini, harga saham individu (titik data) untuk Apple (AAPL), Amazon (AMZN), dan Google (GOOGL) (item tertentu) telah dikumpulkan pada tanggal yang diminta.



**Wide data** adalah data di mana setiap baris berisi beberapa titik data untuk item tertentu yang diidentifikasi dalam kolom.

