**Memproses Data dari Kotor ke Bersih**

**Pembersihan data adalah suatu keharusan**

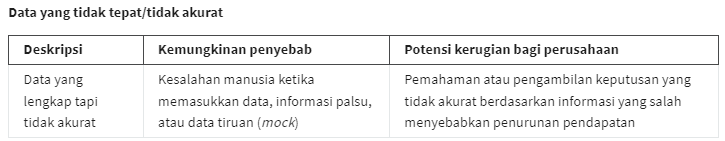
**Apakah data kotor itu?**

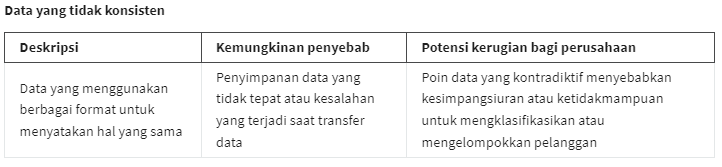
**Data kotor** adalah data yang tidak lengkap, tidak benar, atau tidak relevan terhadap masalah yang ingin Anda pecahkan.

**Data bersih** adalah data yang lengkap, benar, dan relevan dengan masalah yang ingin Anda pecahkan.

**Jenis-jenis data kotor**

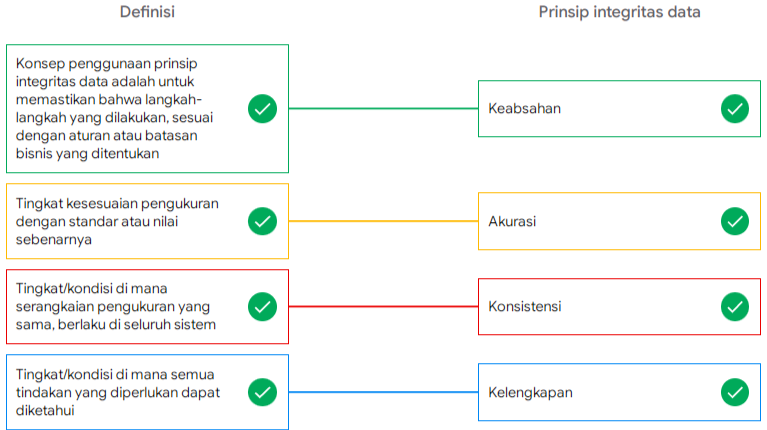






**Orang yang akan bekerja dengan seorang analis data**

* **Data engineer,** mengubah data ke dalam format yang tepat untuk dianalisis dan memberikannya infrastruktur yang dapat diandalkan. Ini berarti mereka mengembangkan, memelihara, dan menguji database, prosesor data serta sistem terkait.
* **Data warehousing specialist** atau spesialis penyimpanan data mengembangkan berbagai proses dan prosedur untuk penyimpanan dan pengaturan data yang efektif. Mereka memastikan bahwa data tersedia, aman, dan sudah dicadangkan guna mencegah terjadinya kehilangan data.





**Kesalahan umum pembersihan data**

* **Tidak memeriksa kesalahan ejaan**: Kesalahan ejaan bisa jadi sesederhana kesalahan mengetik atau memasukkan data. Seringnya kesalahan ejaan atau kesalahan gramatikal umum dapat dideteksi, tapi menjadi semakin sulit ketika itu berupa nama atau alamat. Misalnya, jika Anda bekerja menggunakan tabel spreadsheet data pelanggan, Anda mungkin akan menemukan seorang pelanggan bernama “John” yang namanya ditulis dengan tidak tepat sebagai “Jon” di beberapa tempat. Pemeriksa ejaan (spellcheck) pada spreadsheet tidak menandai hal tersebut, jadi jika Anda tidak memeriksa ulang dan menemukannya, analisis Anda akan mengandung kesalahan ini.
* **Lupa mendokumentasikan kesalahan**: Mendokumentasikan kesalahan bisa sangat menghemat waktu, karena hal tersebut membantu Anda menghindari kesalahan serupa di masa mendatang dengan menunjukkan cara mengatasinya. Misalnya, Anda menemukan sebuah kesalahan pada rumus (formula) di spreadsheet Anda. Anda menemukan beberapa tanggal di salah satu kolom belum diformat dengan benar. Jika perbaikan ini dicatat, maka Anda dapat menjadikannya sebagai referensi ketika kesalahan rumus serupa terjadi di waktu mendatang, sehingga penyelesaian masalah dapat segera dilakukan. Mendokumentasikan kesalahan juga membantu Anda melacak perubahan dalam pekerjaan, sehingga Anda dapat melacaknya kembali jika perbaikan tidak berhasil.
* **Tidak memeriksa nilai-nilai misfielded**: Sebuah nilai misfielded terjadi ketika nilai tersebut dimasukkan ke field yang salah. Nilai-nilai tersebut mungkin tampak sudah diformat dengan benar, yang akan membuatnya semakin sulit ditemukan jika Anda tidak teliti. Misalnya, Anda mempunyai sebuah dataset dengan sejumlah kolom untuk kota dan negara. Ini adalah jenis data yang sama, sehingga mudah tercampur. Tapi ketika Anda mencoba menemukan semua yang terkait dengan Spanyol di kolom negara, dan ternyata Spanyol tak sengaja dimasukkan dalam kolom kota, maka Anda tidak akan menemukan poin data tersebut. Memastikan data Anda dimasukkan dengan benar adalah kunci analisis yang akurat dan lengkap.
* **Mengabaikan nilai-nilai yang hilang**: Nilai-nilai yang hilang dalam dataset dapat menyebabkan kesalahan dan menjadikan kesimpulan Anda tidak akurat. Misalnya, jika Anda mencoba mendapatkan jumlah total penjualan tiga bulan terakhir, tetapi ada transaksi satu minggu yang hilang, maka perhitungan Anda menjadi tidak akurat. Sebagai praktik terbaik, cobalah untuk menjaga data Anda selalu bersih dengan mempertahankan kelengkapan dan konsistensinya.
* **Hanya melihat sebagian (subset) dari suatu data**: Penting halnya untuk mempertimbangkan semua data yang relevan ketika Anda membersihkan data. Hal ini membantu memastikan bahwa Anda memahami apa yang disampaikan oleh data tersebut, dan memperhatikan kesalahan yang mungkin terjadi. Misalnya, jika Anda sedang bekerja menggunakan data tentang pola migrasi burung dari berbagai sumber, tetapi Anda hanya membersihkan satu sumber data saja, Anda mungkin tidak menyadari adanya pengulangan sejumlah data. Hal ini akan menimbulkan masalah dalam analisis Anda nantinya. Jika Anda ingin menghindari kesalahan umum seperti duplikasi data, setiap field data harus Anda cermati satu per satu.
* **Teralihkan dari tujuan bisnis**: Ketika membersihkan data, Anda mungkin menemukan sesuatu yang baru dan menarik dalam dataset -- tetapi Anda tidak menginginkan penemuan itu mengalihkan perhatian dari tugas yang Anda punya. Misalnya, ketika mengolah data tentang cuaca untuk mengetahui nilai rata-rata hari hujan di kota Anda, Anda mungkin juga menemukan pola tentang hujan salju yang menarik. Itu adalah hal yang menarik, tapi tidak berkaitan dengan tugas yang harus Anda selesaikan sekarang. Keingintahuan adalah hal yang baik! Tetapi jangan sampai hal tersebut mengalihkan perhatian Anda dari tugas utama.
* **Tidak memperbaiki sumber kesalahan**: Memperbaiki kesalahan itu sendiri adalah hal yang penting. Tetapi jika kesalahan itu sebenarnya adalah bagian dari masalah yang lebih besar, Anda perlu menemukan sumber masalahnya. Jika tidak, Anda harus memperbaiki kesalahan yang sama berulang kali. Misalnya, bayangkan Anda memiliki sebuah spreadsheet tim yang melacak progres setiap orang. Tabel tersebut terus bermasalah karena orang-orang memasukkan berbagai nilai yang berbeda. Anda dapat terus memperbaiki semua masalah tersebut satu per satu, atau Anda dapat mengatur tabel untuk memperlancar proses entri data sehingga semua orang mempunyai pemahaman yang sama. Mengatasi sumber kesalahan dalam data akan menghemat banyak waktu Anda untuk jangka panjang.
* **Tidak menganalisis sistem sebelum pembersihan data**: Jika kita ingin membersihkan data dan menghindari kesalahan serupa di waktu mendatang, kita perlu memahami penyebab utama adanya data kotor. Bayangkan Anda adalah seorang mekanik otomotif. Anda akan menemukan penyebab masalah sebelum mulai memperbaiki mobilnya, bukan? Begitu pula dengan data. Pertama, Anda harus mencari tahu letak kesalahannya. Kesalahan mungkin terjadi ketika entri data, tidak menyiapkan pemeriksa ejaan, format yang kurang tepat, atau dari duplikasi. Kemudian, setelah mengetahui dari mana data buruk berasal, Anda dapat mengontrolnya dan menjaga kebersihan data Anda.
* **Tidak mencadangkan data sebelum pembersihan**: Ada baiknya bersikap proaktif dan membuat cadangan data sebelum Anda memulai pembersihan data. Jika program Anda bermasalah, atau jika perubahan yang Anda lakukan menyebabkan masalah pada dataset, Anda dapat selalu kembali ke versi yang telah disimpan sebelumnya dan memulihkannya. Prosedur sederhana seperti mencadangkan data dapat menghemat waktu kerja Anda -- dan yang terpenting, menghindari sakit kepala.
* **Tidak merencanakan pembersihan data dalam proses kerja Anda**: Semua hal baik memerlukan cukup waktu, begitu pula dengan pembersihan data. Hal tersebut harus selalu Anda ingat selama proses bekerja dan menentukan tenggat waktu. Jika Anda menyediakan waktu khusus untuk pembersihan data, Anda dapat memperkirakan batas waktu penyerahan hasil kerja yang lebih akurat kepada para pemangku kepentingan, dan membantu Anda menentukan kapan perlu melakukan permintaan penyesuaian batas waktu.

**Fitur Pembersihan Data di Spreadsheet**

* Pemformatan bersyarat adalah sebuah alat spreadsheet yang mengubah bagaimana sel ditampilkan saat nilai yang ada sesuai dengan kondisi tertentu. Alat ini dapat pula memberi tahu jika sebuah sel tidak memenuhi kondisi yang telah Anda tetapkan. Isyarat visual seperti ini sangat berguna untuk analis data, terutama ketika kita bekerja pada sebuah spreadsheet besar dengan banyak data.
* Remove duplicates” adalah suatu alat yang secara otomatis mencari dan menghilangkan duplikasi entri dari suatu spreadsheet.
* Split adalah alat yang membagi string teks di seputar karakter tertentu dan menempatkan setiap fragmennya ke dalam suatu sel baru dan terpisah. Split dapat membantu saat Anda memiliki lebih dari satu bagian data dalam suatu sel dan Anda ingin memisahkannya.
* Concatenate adalah fungsi yang menggabungkan beberapa string teks menjadi satu string tunggal.

**Mengoptimalkan proses pembersihan data**

**Syntax** adalah struktur yang telah ditentukan yang mencakup semua informasi yang diperlukan dan penempatannya yang tepat.

* COUNTIF adalah fungsi yang menghasilkan jumlah sel yang sesuai dengan nilai tertentu. Pada dasarnya, fungsi ini menghitung jumlah berapa kali sebuah nilai muncul dalam sebuah rentang sel.
* LEN adalah fungsi yang memberi tahu Anda panjang <i>string</i> teks dengan menghitung jumlah karakter yang dimilikinya. Fungsi ini berguna saat membersihkan data jika Anda memiliki informasi tertentu di <i>spreadsheet</i> yang mana Anda tahu harus memiliki panjang tertentu.
* LEFT adalah sebuah fungsi yang memberi Anda sejumlah karakter dari sisi kiri suatu string teks.
* RIGHT adalah sebuah fungsi yang memberi Anda sejumlah karakter dari sisi kanan suatu string teks.
* MID adalah sebuah fungsi yang memberi Anda sebuah segmen yang berasal dari bagian tengah suatu string teks.
* CONCATENATE, yang merupakan sebuah fungsi yang menggabungkan dua atau lebih string teks menjadi satu.
* TRIM adalah sebuah fungsi untuk menghapus spasi yang berada di awal, akhir, serta yang berulang di suatu data.
* FILTER berarti hanya menampilkan data yang memenuhi kriteria tertentu dan menyembunyikan yang lainnya.
* TABEL PIVOT mengurutkan, mengatur ulang, mengelompokkan, menghitung, menjumlah atau rata-rata data yang tersimpan di database. Dalam pembersihan data, tabel pivot digunakan untuk memberikan gambaran sekilas dan rapi tentang data.
* VLOOKUP adalah singkatan dari <i>vertical lookup</i> (pencarian vertikal). Ini adalah fungsi yang mencari nilai tertentu dalam suatu kolom untuk menunjukkan bagian informasi yang dimaksud.

