**Apa itu telaah data?**

Telaah data merujuk pada proses pengumpulan, pembersihan, eksplorasi, dan analisis data untuk mendapatkan pemahaman yang lebih yang terkandung dalam data tersebut. Tujuan dari telaah data adalah mendukung pengambilan keputusan berdasarkan bukti yang ditemukan dalam data.

**Mengapa perlu telaah data?**

1. Data dari masing-masing sumber belum tentu dapat langsung dipakai karena:

* maksud dan tujuan data berbeda-beda
* keadaan asal terpisah-pisah atau justru terintegrasi secara ketat.
* tingkat kekayaan (richness) berbeda-beda
* tingkat keandalan (reliability) berbeda-beda

1. Data understanding memberikan gambaran awal tentang:

* kekuatan data
* kekurangan dan batasan penggunaan data
* tingkat kesesuaian data dengan masalah bisnis yang akan dipecahkan
* ketersediaan data (terbuka/tertutup, biaya akses, dsb.)

**Bentuk Data**

Data dapat memiliki berbagai bentuk, diantaranya:

* Spreadsheet(excel, csv, dll)
* Database(SQL, Oracle, dll)
* Text file(txt, doc, pdf, dll)
* Multimedia documents(audio, video, gambar, dll)

**Sumber Data**

1. Sumber Internal

* Data yang dihasilkan oleh perusahaan sendiri
* Data yang dihasilkan oleh perusahaan lain yang terkait dengan perusahaan

1. Sumber Eksternal

* Repositori data publik
* Halaman web publik

**Daftar Sumber Data Daring**

* [Portal Satu Data Indonesia](https://data.go.id/)
* [Portal Data Jakarta](https://data.jakarta.go.id/)
* [Portal Data Bandung](http://data.bandung.go.id/)
* [Badan Pusat Statistik](https://www.bps.go.id/)
* [Badan Informasi Geospasial](https://tanahair.indonesia.go.id/)
* [UCI Machine Learning repository](https://archive.ics.uci.edu/ml/index.php)
* [Kaggle](https://www.kaggle.com/datasets)
* [World Bank Open Data](https://data.worldbank.org/)

**Tipe Data**

**Bedasarkan sifat**

* **Data dikotomi**, merupakan data yang bersifat pilah satu sama lain, misalnya suku, agama, jenis kelamin, pendidikan, dan lain sebagainya.
* **Data diskrit**, merupakan data yang proses pengumpulan datanya dijalankan dengan cara menghitung atau membilang. Seperti, jumlah anak, jumlah penduduk, jumlah kematian dan sebagainya.
* **Data kontinu**, merupakan data pengumpulan datanya didapatkan dengan cara mengukur dengan alat ukur yang memakai skala tertentu. Seperti misalnya, Suhu, berat, bakat, kecerdasan, dan lainnya.

**Bedasarkan waktu**

* **Data Cross Section**, adalah data yang menunjukkan titik waktu tertentu. Contohnya laporan keuangan per 31 Desember 2020, data pelanggan PT.Data Indah bulan mei 2004, dan lain sebagainya.
* **Data Time Series / Berkala**, adalah data yang datanya menggambarkan sesuatu dari waktu ke waktu atau periode secara historis. Contoh data time series adalah data perkembangan nilai tukar dollar amerika terhadap rupiah tahun 2016 - 2020

**Pengambilan Data**

* Pengambilan data secara manual.
* Pengambilan data melalui API
  + Contoh melalui API Kaggle (pip install kaggle)(kaggle datasets download -d mauryansshivam/paytm-revenue-users-transactions)
  + Contoh melalui API Portal Data Bandung (http://data.bandung.go.id/index.php/portal/api/1db4a0cc-dc0f-4362-91f1-d0ad1b4bce98)
* Pengambilan data melalui akses langsung ke basis data relasional yang ada.
* Pengambilan data melalui web scraping.
  + Contoh pengambilan data melalui web scraping kometar video youtube