filkom.ub.id

DIGITAL TALENT SCHOLARSHIP 2019

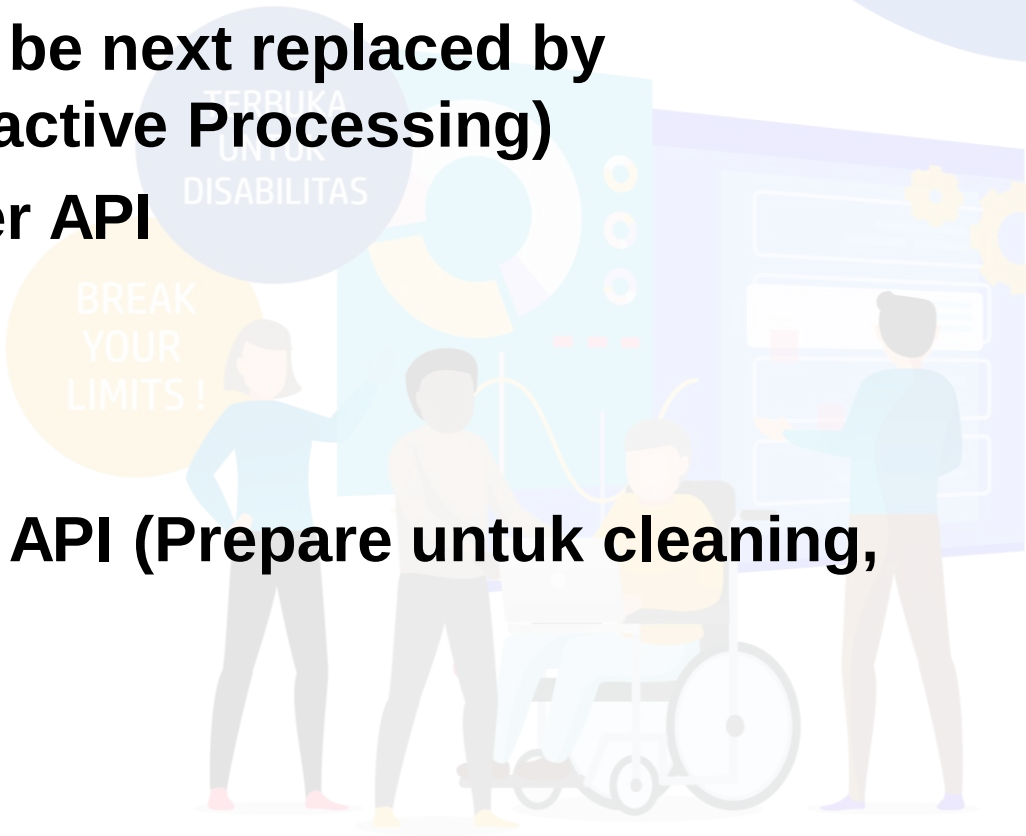
Big Data Analytics

Importing dan Exporting Data Crawling

Oleh: Imam Cholissodin | imamcs@ub.ac.id, Putra Pandu Adikara, Sufia Adha Putri
Asisten: Guedho, Sukma, Anshori, Aang dan Gusti
Fakultas Ilmu Komputer (Filkom) Universitas Brawijaya (UB)

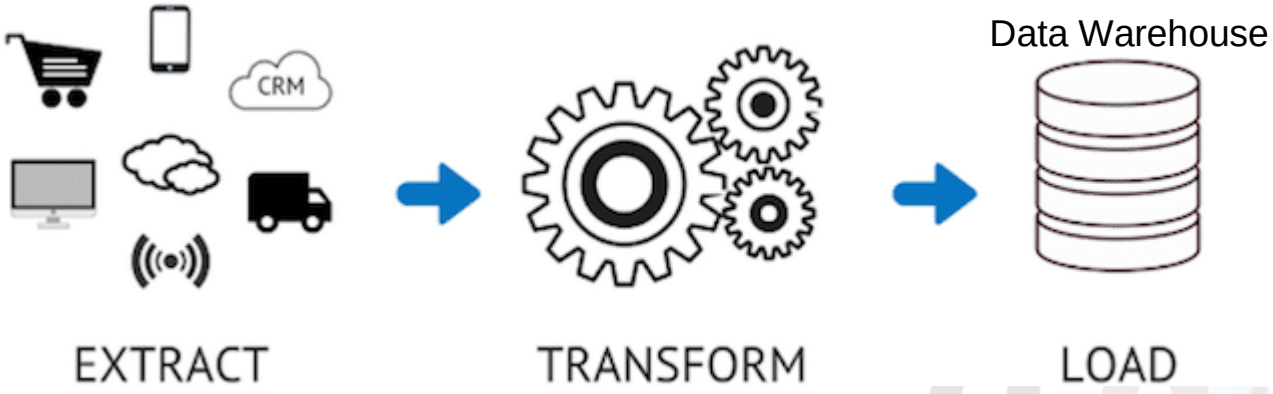
Pokok Bahasan

1. Pengenalan ETL (May be next replaced by Realtime/Stream/Interactive Processing)
2. Akses data dari Twitter API
    - Get Friends
    - Get Status
    - etc
3. Olah Data dari Twitter API (Prepare untuk cleaning, etc)
4. Tugas



Pengenalan ETL

- ETL adalah bagian pertama dari **strategi integrasi data** (integration data) yang memungkinkan bisnis dan organisasi untuk **mengumpulkan data dari berbagai sumber** dan menggabungkannya menjadi satu, lokasi terpusat (pada Data Warehouse).
- Struktur ETL (extract, transform, load) :



- **Langkah pertama (Extract)**, yaitu organisasi bisnis melakukan ekstraksi dengan mengelola **data mentah** (terstruktur, semi terstruktur, maupun yang tidak terstruktur) dari berbagai sumber untuk **diimport dalam satu repository** dengan menggunakan sistem analisis data untuk menghasilkan intelijen bisnis.

Pengenalan ETL

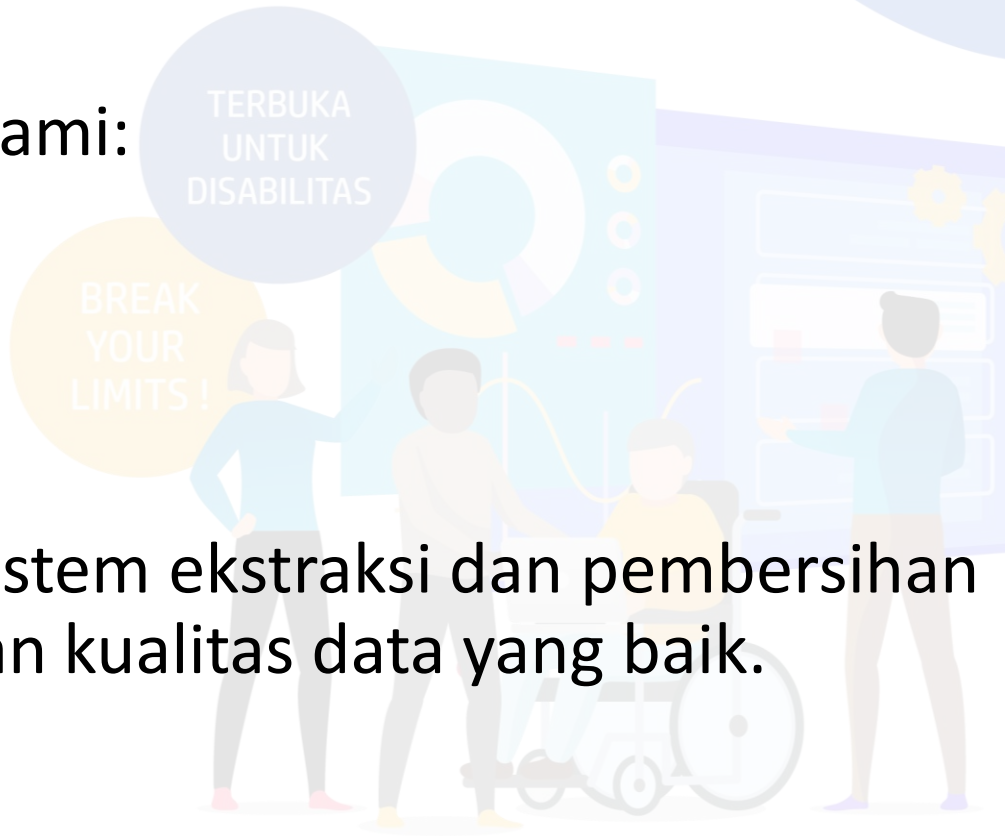
- **Langkah kedua (Transform)**, yaitu **memastikan kualitas dan aksesibilitas serta integritas data**. **Perusahaan** Anda dapat **menerapkan aturan** untuk memenuhi hasil proses transform sesuai dengan spesifikasi data yang dibutuhkan. Proses transformasi data terdiri dari beberapa sub-proses:
  - **Pembersihan**: ketidakkonsistenan dan missing value data, diselesaikan.
  - **Standarisasi**: pemformatan pada kumpulan data.
  - **Deduplikasi**: data yang duplicate dikecualikan atau dibuang.
  - **Verifikasi**: data yang tidak dapat digunakan dihapus dan anomali ditandai.
  - **Penyortiran**: data disusun menurut jenisnya.
  - Other Task: aturan tambahan / opsional dapat diterapkan untuk meningkatkan kualitas data.
- **Langkah terakhir (Loading ke data warehouse)**, yaitu **data dimuat** secara sekaligus (**full load**) atau dengan interval **yang terjadwal** (incremental load).

Kualitas Data (1)

- Penerapan system ETL adalah untuk **mendapatkan** data dengan **kualitas yang baik**.
- Kualitas data dipengaruhi oleh:
  - **Heteroginitas** sumber data.
    - Perbedaan Teknologi
    - Perbedaan Platform
  - **Ukuran Data** yang besar yang dihasilkan setiap hari oleh suatu sumber data.
  - **Permasalahan pada Data** yang dialami oleh suatu sumber data.

Kualitas Data (2)

- Permasalahan yang dialami:
  - Duplikasi data
  - Inkonsistensi data
  - Data ambigu
  - Data yang tidak lengkap
- Sehingga, dibutuhkan sistem ekstraksi dan pembersihan data untuk menghasilkan kualitas data yang baik.



Kualitas Data (3)

- Kualitas data yang baik merupakan **asset yang sangat berharga**. Sementara kualitas data yang buruk dapat membahayakan **kredibilitas dan akurasi hasil** pengolahan data.
- Kualitas data adalah sebuah **persepsi atau penilaian kelayakan data** untuk melayani tujuannya dalam konteks tertentu.

Kualitas Data (4)

- Parameter kualitas data:
  - *Correctness/Accuracy*: sejauh mana data dapat **menggambarkan secara benar** sebuah entitas nyata.
  - *Consistency*: Data **memberikan satu versi kebenaran** walau diperlakukan dalam kondisi yang berbeda.
  - *Completeness*: Sejauh mana **atribut yang** diinginkan **bisa disediakan** oleh data tersebut.
  - *Timeliness*: **Ketepatan waktu** dari kedatangan suatu data.
  - *Metadata*: **Informasi** mengenai **data** itu sendiri.

Extraction (1)

- Seperti yang ditunjukkan pada bagan alir, tahap pertama dari system ETL adalah *Extraction* (ekstraksi)
- Prinsip-prinsip dasar pada ekstraksi data adalah:
  - **Volume data** yang diambil berukuran besar.
  - **Proses** ekstraksi dilakukan se-**cepat** mungkin.
  - Proses ekstraksi dilakukan sebisa mungkin **outputnya menjadi kecil (simple)**.
  - Diharapkan, **perubahan di sumber data seminimal mungkin**.

Extraction (2)

- Sebelum melakukan ekstraksi data, diperlukan sebuah **peta logika data yang menggambarkan relasi antara *feature*** dari sumber data dan *feature* data yang akan diolah atau ditampilkan ke pengguna akhir.
- **Peta logika tersebut menjadi panduan** yang mengikat proses ETL dari awal hingga akhir.

Extraction (3)

- Langkah-langkah pembuatan peta logika data:
  - Memiliki **perencanaan yang matang** – berlandaskan pada metadata.
  - **Identifikasi kandidat-kandidat sumber data** – identifikasi sumber-sumber data yang dibutuhkan dalam pengambilan keputusan.
  - Analisa sumber data dengan aplikasi *data-profiling* – Anomali data harus dapat dideteksi dan didokumentasi dengan baik.
  - Memahami kebutuhan data dan aturan bisnis pada bagian pengguna akhir.
  - Memahami model data dari tempat penyimpanan data.
  - Melakukan validasi formula dan proses perhitungan pada data.

Extraction (4)

- Komponen-komponen pada peta logika data:
  - Sumber data
  - Parameter-parameter sumber data
  - Parameter-parameter pada keluaran data
  - Transformasi

Transformation (1)

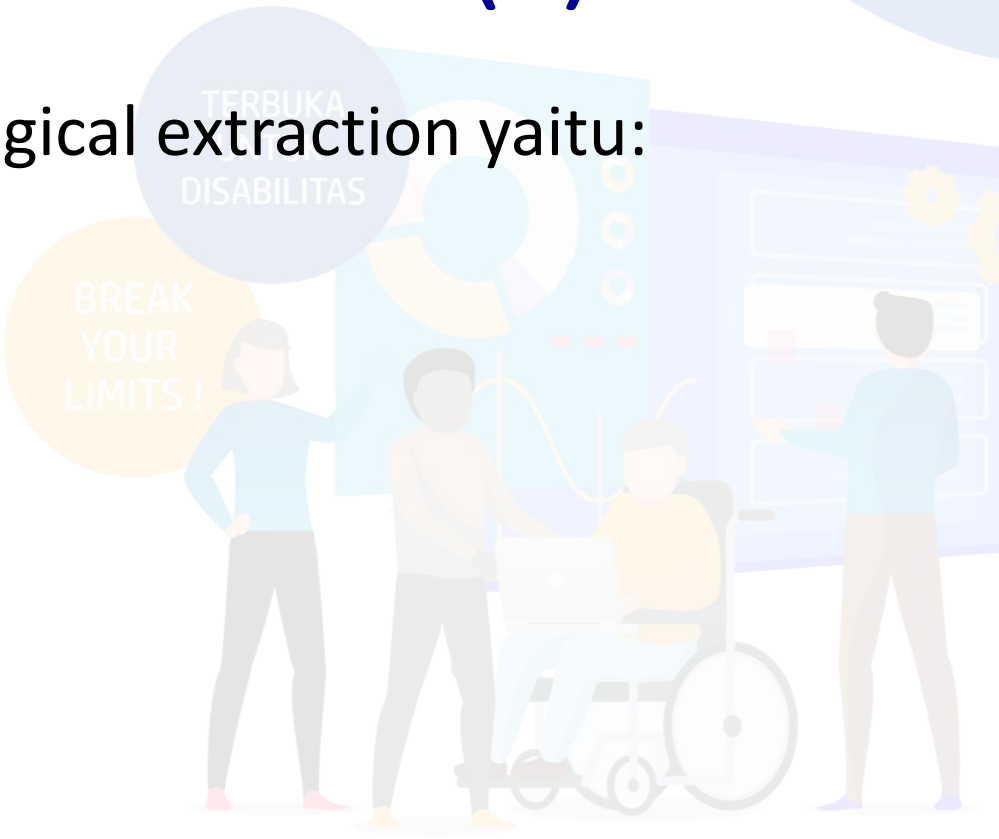
- *Transformation data* pada system ETL melingkupi aktifitas-aktifitas berikut:
  - *Formatting* dan standardisasi.
  - Mengubah ke angka atau teks tertentu atau format tanggal.
  - Terjemahkan data ke bentuk yang lain.
  - Agregasi atau merangkum data pada level yang lebih tinggi.

Transformation (2)

- Transformasi data juga melibatkan prinsip-prinsip:
  - *Leakage* (kebocoran) terjadi ketika proses ETL mengunduh data secara lengkap dari sumber data, namun pada kenyataannya terdapat beberapa *record* yang hilang.
  - *Recoverability* (pemulihan) berarti bahwa proses ETL harus *robust* sehingga jika terjadi kegagalan, ini bisa segera dipulihkan tanpa kehilangan atau kerusakan data.

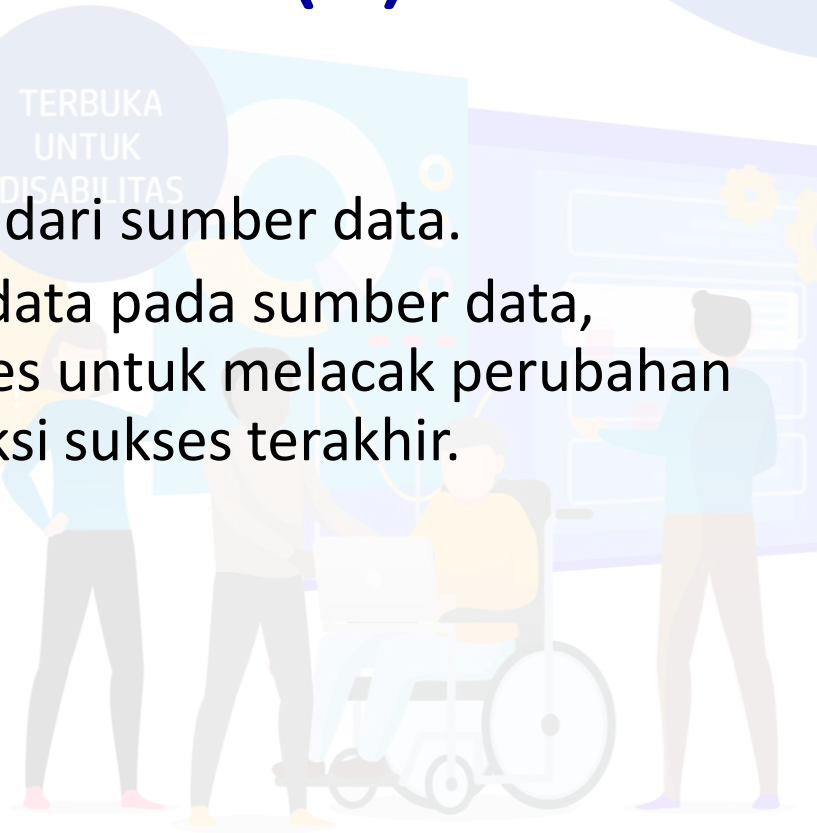
Metode *Logical Extraction* (1)

- Terdapat dua metode logical extraction yaitu:
  - Full extraction
  - Incremental extraction



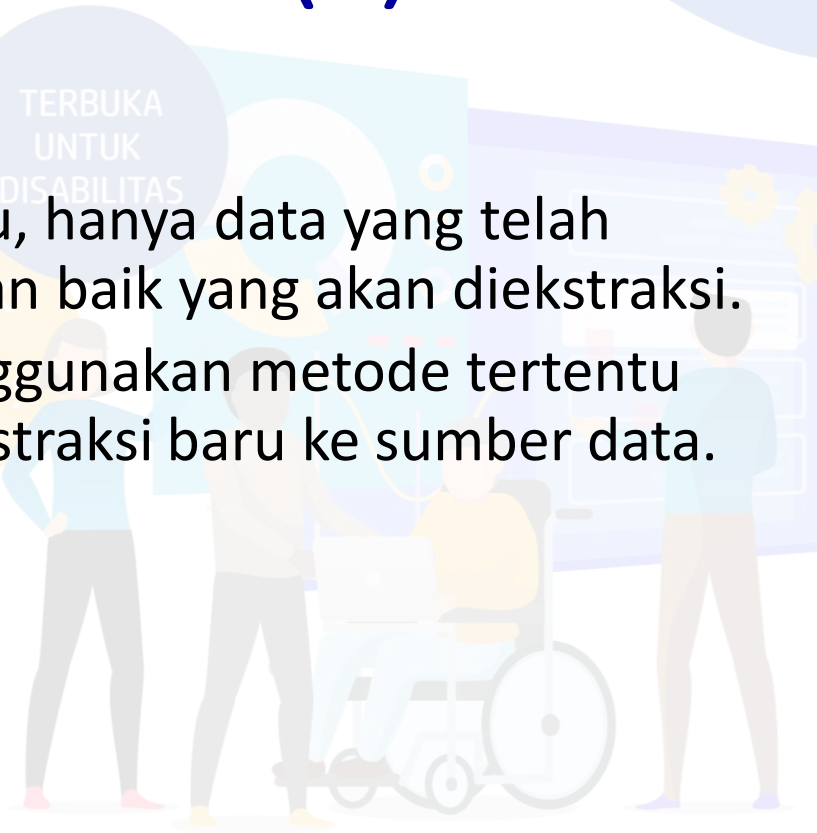
Metode *Logical Extraction* (2)

- *Full Extraction*:
  - Pengambilan keseluruhan data dari sumber data.
  - Ekstrasi ini mereplikasi semua data pada sumber data, sehingga tidak diperlukan proses untuk melacak perubahan pada sumber data sejak ekstraksi sukses terakhir.



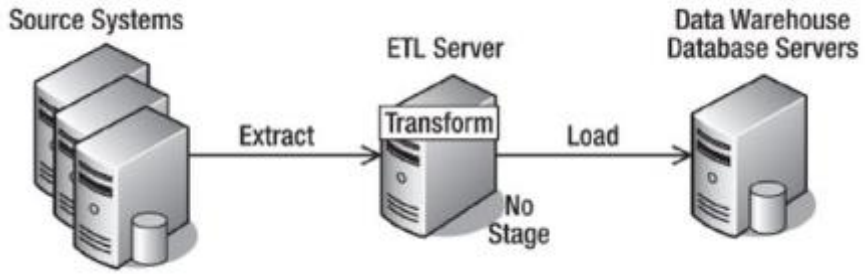
Metode *Logical Extraction* (3)

- *Incremental Extraction*:
  - Pada titik tertentu dalam waktu, hanya data yang telah berubah sejak terdefinisi dengan baik yang akan diekstraksi.
  - Dalam kebanyakan kasus, menggunakan metode tertentu untuk menambahkan logika ekstraksi baru ke sumber data.



Load (1)

- Langkah terakhir dari system ETL adalah *Load* yang berupa:
  - Penyimpanan data ke *data warehouse*
  - Menampilkan data ke aplikasi atau pengguna akhir.
- Arsitektur ETL secara keseluruhan hingga tahap load adalah sebagai berikut:



Latihan langsung di Kelas Ke-1 & Pembahasan
Link kode “http://bit.ly/2Sr4NeE”

Silahkan dicoba dijalankan dengan Jupyter notebook yang Anda buat sebelumnya di Ubuntu 16.04 atau dengan SageMaker notebook (JupyterLab) yang baru Anda buat hari ini.

Lab-Sesi14-1

Menggunakan tweepy untuk Data Crawler

In [1]:
```
#!pip install tweepy
```

In [17]:
```
import tweepy
import pandas as pd
```

In [18]:
```
class Stream2Screen(tweepy.StreamListener):
    def on_status(self, status):
        if hasattr(status, 'retweeted_status'):
            try:
                tweet = status.retweeted_status.extended_tweet["full_text"]
            except:
                tweet = status.retweeted_status.text
        else:
            try:
                tweet = status.extended_tweet["full_text"]
            except AttributeError:
                tweet = status.text
```

In [19]:
```
consumer_key = "JEj5tRSA9JWjWV6imMOrUIVWV"
consumer_secret = "7MEa0OKHpUbjxble8pdlV74qPbvW2OHqLtjt45QQraJaAzRmAh"
access_token = "935208713551364097-W9Oy0IS2M1dRUQS5MZ6Dnz18BkHUP80"
access_secret = "jCANa7K7werTP2X1mnLlcRBFDHAJt9TZSCbC77FSNCj50"

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
```

Latihan langsung di Kelas Ke-2 & Pembahasan

- Tugas latihan ke-2 ini tidak ada

filkom.ub.id

DIGITAL TALENT SCHOLARSHIP 2019

Big Data Analytics

Terimakasih

Oleh: Imam Cholissodin | imamcs@ub.ac.id, Putra Pandu Adikara, Sufia Adha Putri
Asisten: Guedho, Sukma, Anshori, Aang dan Gusti
Fakultas Ilmu Komputer (Filkom) Universitas Brawijaya (UB)

digitalent.kominfo.go.id