KOMDIGI
Kementerian Komunikasi dan Digital
Republik Indonesia

digitalent

VSGA Vocational School Graduate Academy

VOCATIONAL SCHOOL GRADUATE ACADEMY

Associate Data Scientist

Pertemuan: Mengumpulkan Data

Outline Pembelajaran

A. Menentukan kebutuhan data
B. Mengambil data
C. Mengintegrasikan data

A. Menentukan Kebutuhan Data

Menentukan kebutuhan data adalah proses mengidentifikasi dan mendokumentasikan data yang dibutuhkan oleh user untuk menjawab masalah bisnis

Dua jenis informasi mengenai data adalah:

1. Informasi yang menjelaskan struktur data, seperti entitas, atribut, dan relasi.
    - Informasi ini biasanya dinyatakan dalam bentuk grafik seperti *Entity-Relationship Diagrams* (E-RD).
2. Informasi yang menggambarkan aturan atau batasan yang dapat menjaga integritas data.
    - Biasanya disebut aturan bisnis (*business rules*), batasan-batasan ini harus di tuangkan dalam data dictionary/directory (atau *repository*) suatu organisasi.

Proses Menentukan Kebutuhan Data

Apa yang diperlukan pengguna? Apa yang diinginkan pengguna?
Mengapa kita perlu mendefinisikan kebutuhan pengguna?

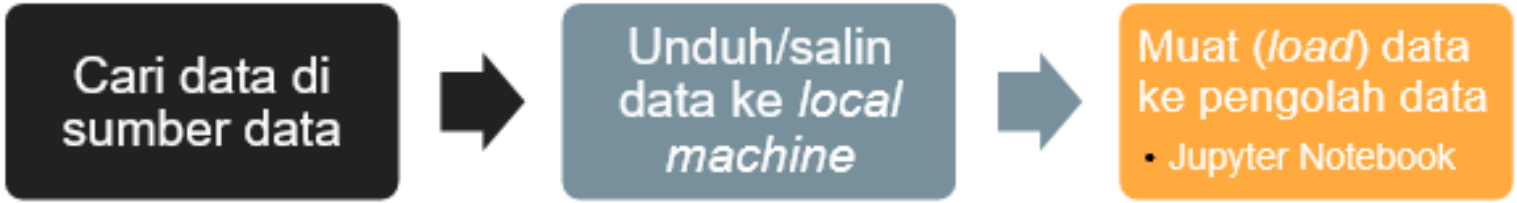
Langkah-langkah Menentukan Kebutuhan Data

1. Mendefinisikan lingkup database
2. Memilih metodologi
3. Mengidentifikasi pandangan user (User Views)
4. Model struktur data
5. Model database constraints
6. Mengidentifikasi kebutuhan operasional

B. Mengambil Data

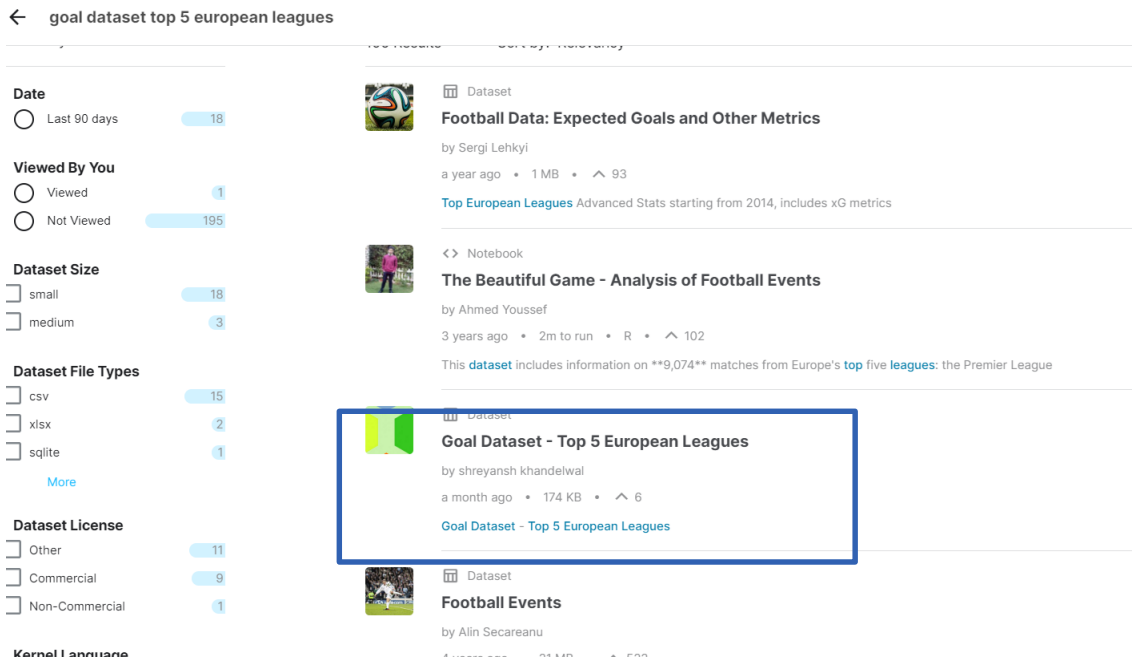
1. Pengambilan data secara manual
2. Pengambilan data melalui API, contoh melalui API Kaggle atau Twitter
3. Pengambilan data melalui web scraping
4. Pengambilan data melalui akses langsung ke basis data relasional yang ada.

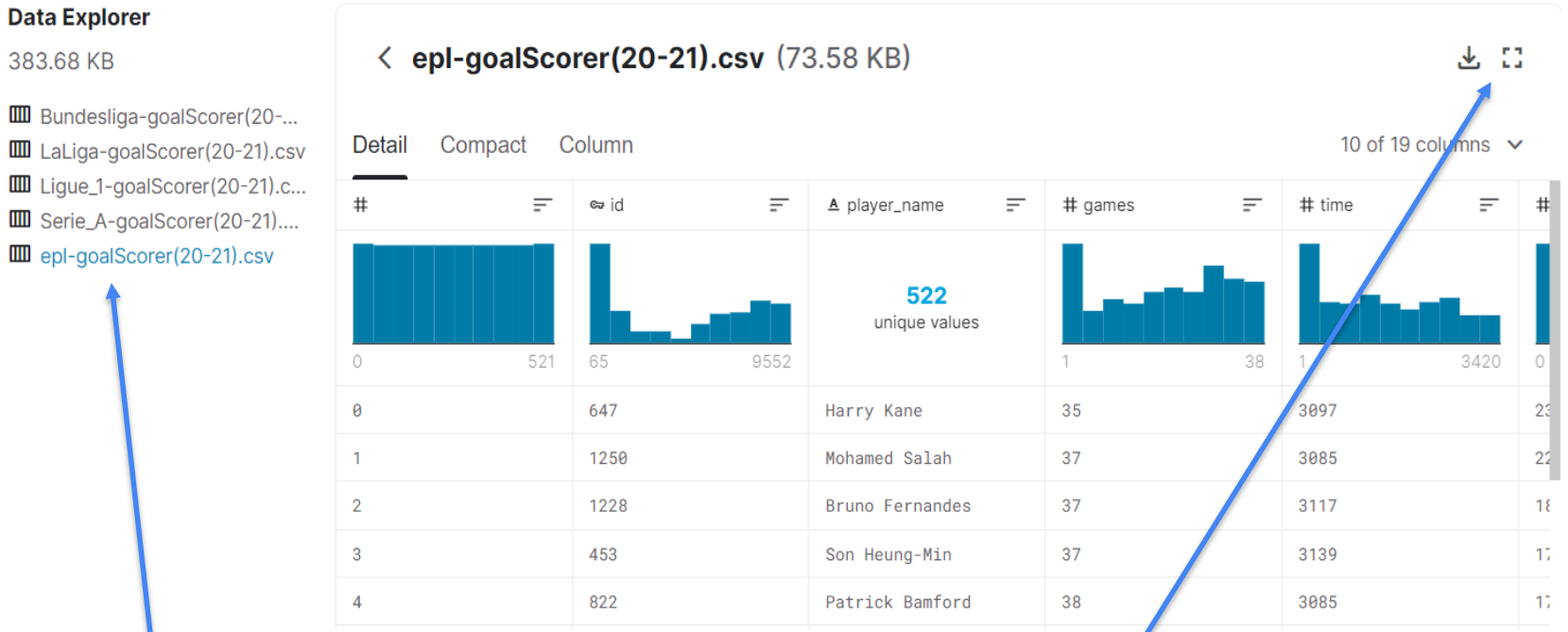
Pengambilan Data Secara Manual



Pengambilan Data (Secara Manual) dari Kaggle

1. Kita akan mengakses data dari "Goal Dataset – Top 5 European Leagues" dari Kaggle.
2. Kunjungi Kaggle.com dan login (buat akun jika perlu)
3. Lakukan pencarian "goal dataset top 5 European leagues"
4. Klik "Goal Dataset – Top 5 European Leagues"



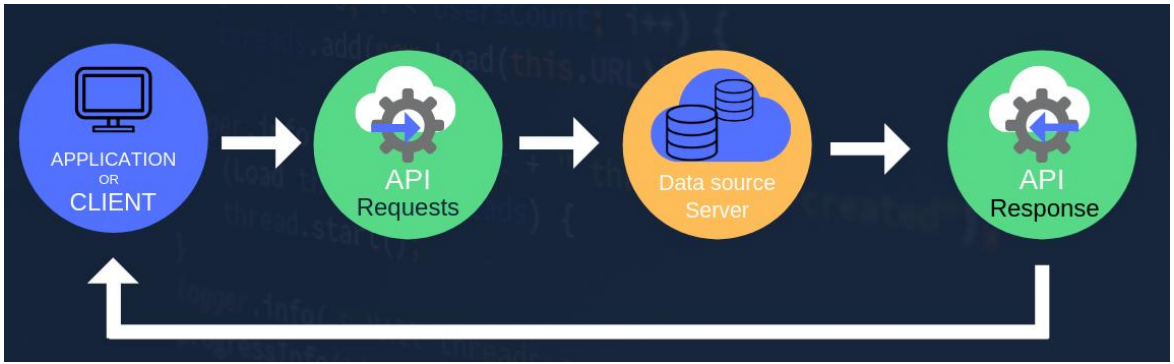


- Di halaman data explorer, pilih "epl-goalScorer (20-21).csv"
- Unduh data dengan mengklik tombol unduh di bagian kanan dan simpan di folder kerja Anda.

Pengambilan Data Melalui API

Pengertian *application programming interface* (API).

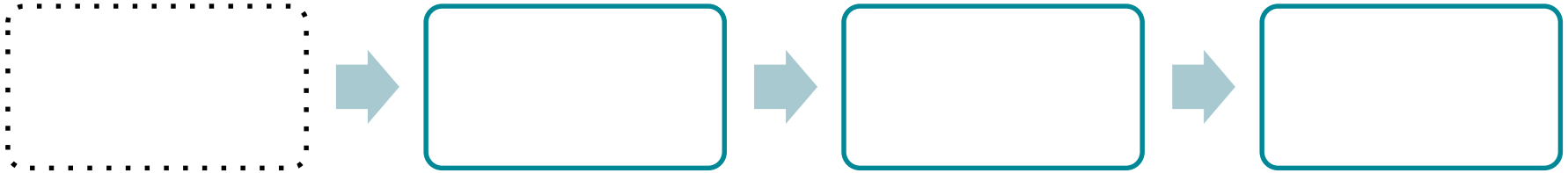
- Sekumpulan aturan yang didefinisikan untuk memfasilitasi proses komunikasi antar komputer atau program aplikasi.
- API berfungsi sebagai perantara antara program aplikasi dengan sebuah web server yang memungkinkan transfer data diantara kedua pihak yang berkomunikasi.



Pengambilan Data Melalui API

Data dapat diambil melalui *application programming interface* (API).

- API disediakan oleh beberapa layanan data seperti Kaggle.
- API token/key (mungkin) diperlukan untuk mengakses data via API.
- Proses pembuatan API token/key (jika perlu) dirinci di dokumentasi masing-masing layanan.



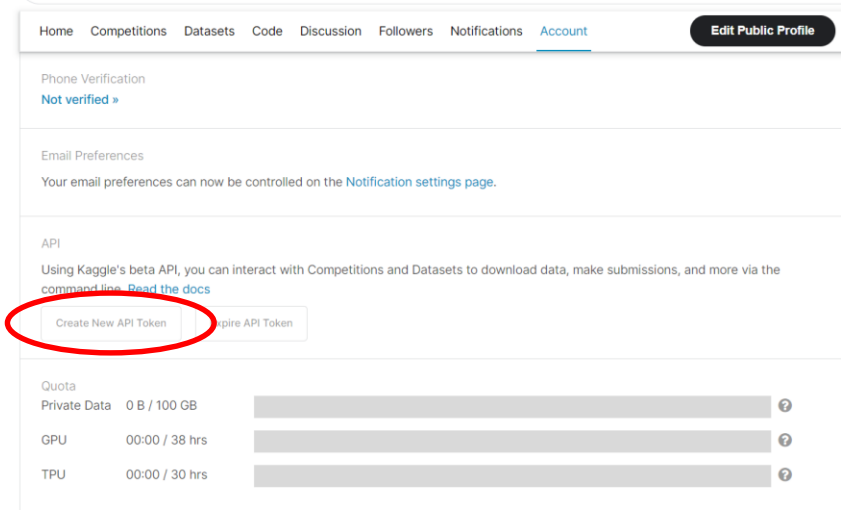
Pengambilan Data dari Kaggle Melalui API (1/6)

- Jalankan Jupyter Notebook di folder kerja Anda, lalu buka atau buat satu skrip baru (Python 3).
- Instal kaggle library (mis: dengan pip)

```
In [1]: !pip install kaggle
```

Pengambilan Data dari Kaggle Melalui API (2/6)

- Login ke Kaggle, klik foto profil Anda (di kanan atas), kemudian klik 'Your Profile' untuk membuka halaman profil Anda.
- Pada halaman profil Anda, klik tab 'Account'. Geser ke bawah sedikit, dan Anda akan menemukan tombol 'Create New API Token'



Pengambilan Data dari Kaggle Melalui API (3/6)

- Klik 'Create New API Token'. Jika tombol tidak berfungsi, klik 'Expire API Token' lebih dahulu.
  - Browser akan mengunduh file `kaggle.json` ke folder unduhan (Downloads) Anda.
- Kaggle API secara default mengasumsikan bahwa file `kaggle.json` tersebut berada di dalam folder:

  `~/.kaggle/` (Linux/Mac) atau
  `C:\Users\<Windows-username>\.kaggle\` (Windows)
  - Jika folder tersebut belum ada, buat dulu dengan perintah `mkdir` di shell/command line.
  - Pindahkan file `kaggle.json` ke folder tersebut (menggunakan File/Windows Explorer atau melalui perintah `mv` atau `move` di shell)

Pengambilan Data dari Kaggle Melalui API (4/6)

- Kaggle API memiliki empat perintah
  - `kaggle competitions {list, files, download, submit, submissions, leaderboard}`
  - `kaggle datasets {list, files, download, create, version, init}`
  - `kaggle kernels {list, init, push, pull, output, status}`
  - `kaggle config {view, set, unset}`

- Dokumentasi Kaggle API dapat dilihat di https://github.com/Kaggle/kaggle-api
- Untuk keperluan modul ini, kita hanya menggunakan perintah kaggle datasets

Pengambilan Data dari Kaggle Melalui API (5/6)

- Untuk melakukan pencarian dataset: `kaggle datasets list -s <keyword>`
- Jika terjadi masalah gagal akses, dsb., bisa dicoba dengan membuat ulang API Token.
- Nama dataset berada di kolom ref pada tabel output pencarian. Misalnya kita ingin mengunduh "Goal Dataset – Top 5 European Leagues, maka nama dataset adalah: `shreyanshkhandelwal/goal-dataset-top-5-european-leagues`.

```
In [2]: !kaggle datasets list -s "goal leagues"
```

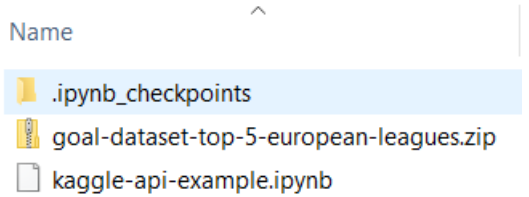
```
ref                                                         title                                              size  lastUpd
ated          downloadCount  voteCount  usabilityRating
----------------------------------------------------------  -------------------------------------------------  -----  -------
------------  -------------  ---------  ---------------
slehkyi/extended-football-stats-for-european-leagues-xg     Football Data: Expected Goals and Other Metrics      1MB  2020-08
-02 17:28:39           2733         94  1.0
secareanualin/football-events                               Football Events                                    21MB  2017-01
-25 01:19:19          19416        525  0.7647059
shreyanshkhandelwal/goal-dataset-top-5-european-leagues     Goal Dataset - Top 5 European Leagues             174KB  2021-05
-23 21:20:09             25          6  0.5294118
chaibapat/fantasy-premier-league                            Fantasy Premier League - 2016/2017                476MB  2017-05
-16 18:56:26           1466         31  0.85294116
yamaerenay/most-popular-soccer-leagues                      Most Popular Soccer Leagues                        30KB  2020-08
-01 16:59:30             78          5  1.0
```

Pengambilan Data dari Kaggle Melalui API (6/6)

- Unduh dataset yang diinginkan dengan perintah `kaggle datasets download`

```
In [3]: !kaggle datasets download shreyanshkhandelwal/goal-dataset-top-5-european-leagues
```

- Dataset akan terunduh di folder aktif dalam bentuk file terkompresi zip.



- Selanjutnya, kita ekstraksi dataset tersebut dengan perintah `unzip`, dan dataset berupa berkas-berkas csv siap digunakan.

```
In [4]: !unzip goal-dataset-top-5-european-leagues.zip

Archive:  goal-dataset-top-5-european-leagues.zip
  inflating: Bundesliga-goalScorer(20-21).csv
  inflating: LaLiga-goalScorer(20-21).csv
  inflating: Ligue_1-goalScorer(20-21).csv
  inflating: Serie_A-goalScorer(20-21).csv
  inflating: epl-goalScorer(20-21).csv
```

- Berkas csv dapat langsung dimuat ke Pandas DataFrame

Pengambilan Data Melalui API

- Kaggle dan beberapa layanan data lainnya menyediakan akses melalui API.
- Langkah-langkah mengakses API biasanya melalui proses pembuatan API token/API key yang dirinci di dokumentasi masing-masing layanan.
- Selain API, teknik pengambilan data yang bersifat lanjut mencakup web scraping serta akses data langsung dari basis data relasional.

Pengambilan Data Menggunakan *Web Scraping*

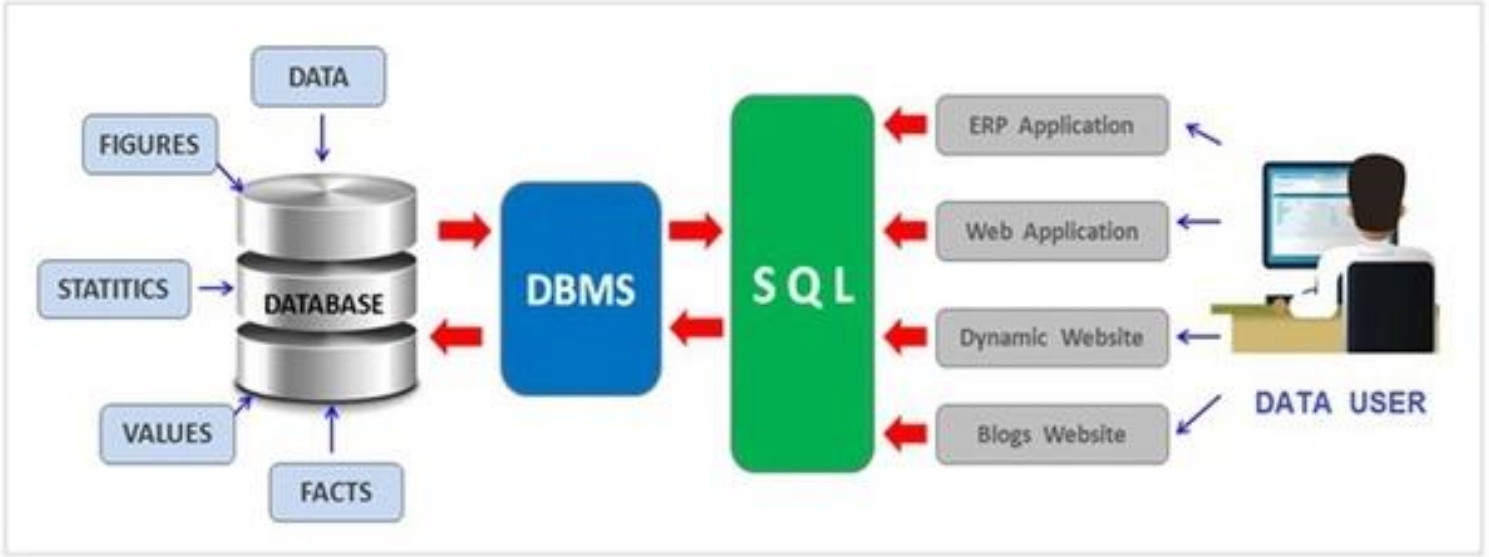
- *Web scraping* = mengekstraksi data secara langsung dari suatu halaman web.
- Langkah-langkah umum (contoh detil dapat dilihat di https://realpython.com/beautiful-soup-web-scraper-python/)
  - Tentukan URL halaman web (HTML) yang akan di-*scrape*.
  - Gunakan fungsi requests.get untuk mengakses URL tersebut. Teks HTML akan tersimpan pada atribut text dari object yang dikembalikan requests.get.
  - Lakukan *parsing* pada HTML dengan library beautifulsoup untuk memperoleh tabel data yang diinginkan (dengan mengekstraksi elemen-elemen HTML yang relevan).

Pengambilan Data Menggunakan *Web Scraping*

Pengambilan Data dari Relasional DBMS

- Data juga dapat bersumber dari basis data relasional DBMS (RDBMS) organisasi.
- Langkah-langkah umum:
  - Import pandas
  - Import library penghubung RDB, misal: mysql.connector untuk MySQL
  - Gunakan method connect dari penghubung RDB untuk membuka koneksi ke RDB.
  - Siapkan SQL query dalam string.
  - Gunakan pandas.read_sql dengan argument string SQL query dan koneksi RDB untuk mengeksekusi SQL dan memuat hasilnya ke dalam DataFrame.
  - Tutup koneksi.
- Proses antara membuka hingga menutup koneksi biasanya ditaruh dalam blok try-except
- Pembukaan koneksi membutuhkan kredensial (username, password) ke RDBMS yang di-hardcode secara langsung. Ini dapat disembunyikan dengan teknik pengamanan yang tidak dibahas di sini.
- Contoh singkat dapat dilihat di: https://medium.com/analytics-vidhya/importing-data-from-a-mysql-database-into-pandas-data-frame-a06e392d27d7

Pengambilan Data dari Relasional DBMS



Pengambilan Data dari Relasional DBMS

Jenis-jenis Integritas data yang menjadi batasan (*constraint*) sebuah database dapat dikelompokkan menjadi:

1. Integritas *Entity:* Kunci utama dari entitas tidak boleh bernilai null.
2. Integritas *Domain:* Nilai setiap fitur (variabel, atribut) entitas harus berasal dari sebuah domain tertentu.
3. Integritas *Referential:* Nilai kunci utama (*primary key*) sebuah tabel entitas berasal dari nilai kunci utama tabel lain yang menjadi referensi (*foreign key*).
4. Integritas *User-defined:* Aturan dan batasan mengenai entitas dibuat oleh pengelola (admin) database untuk menyesuaikan dengan kebutuhan atau penggunaan database.

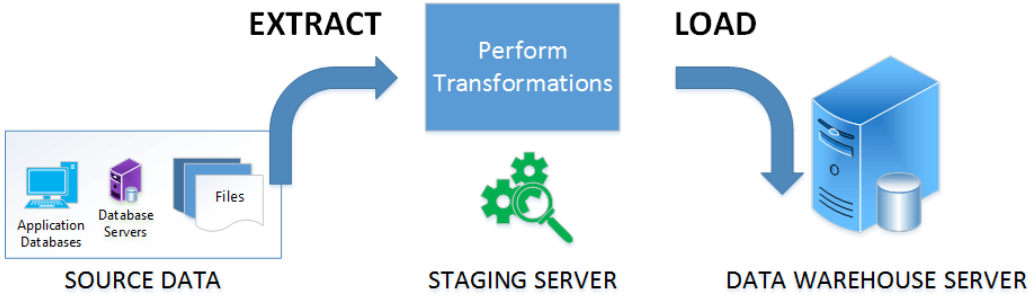
C. Mengintegrasikan Data

- Integrasi data adalah proses menggabungkan atau mengkombinasikan dua atau lebih set data yang berasal dari sumber yang berbeda ke dalam suatu penyimpanan seperti data warehouse.
- Salah satu manfaat yang didapatkan dengan melakukan integrasi data adalah terhindar dari duplikat data.
- Jika terdapat duplikat data maka akan mengganggu proses selanjutnya yang hendak dilakukan seperti analisis data karena nilai yang diperoleh bisa tidak konsisten
- Proses integrasi data diimplementasikan kedalam proses ETL (Extract, Transform, Load) atau ELT (Extract, Load, Transform).
- Pemilihan ETL atau ELT tergantung kepada tujuan data science.

C. Mengintegrasikan Data

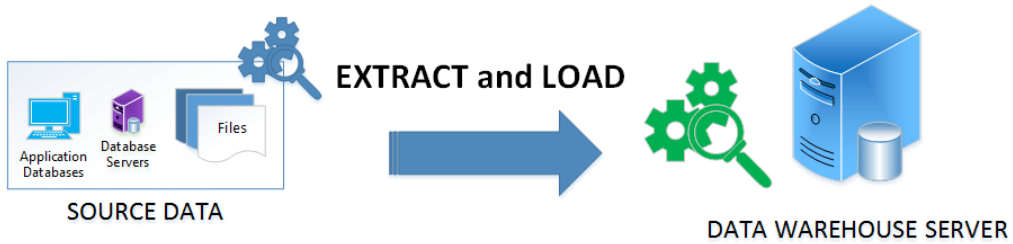
- Beberapa kendala dari proses integrasi data adalah:
  - Volume data yang besar
  - Keragaman format fitur data
  - Data dari berbagai sumber seringkali memiliki kualitas yang berbeda
  - Adanya keterlambatan dari ketersediaan data dari sumber yang berbeda
  - Hasil pengumpulan data dari beberapa sumber data mengandung duplikasi data.
  - Automasi proses integrasi membutuhkan sejumlah tools.

C. Mengintegrasikan Data



ETL (Extract – Transform – Load)

ELT (Extract – Load - Transform)



Rangkuman

Unit Kompetensi mengumpulkan data, berhubungan dengan pengetahuan, keterampilan dan sikap kerja yang dibutuhkan dalam mengumpulkan data untuk data science, mempunyai tiga elemen kompetensi yaitu: Menentukan kebutuhan data, mengambil data dan mengintegrasikan data.

1. Menentukan kebutuhan data adalah proses mengidentifikasi dan mendokumentasikan data yang dibutuhkan oleh user dalam sebuah database untuk memenuhi kebutuhan informasi saat ini dan masa yang akan dating
2. Mengambil data adalah proses pengumpulan, manipulasi dan pemrosesan data berdasarkan data yang dikumpulkan agar dapat digunakan untuk mencapai tujuan tertentu.
3. Mengintegrasikan data adalah proses menggabungkan atau mengkombinasikan dua atau lebih set data yang berasal dari sumber yang berbeda ke dalam suatu penyimpanan seperti data warehouse

Latihan Praktek: Menentukan Kebutuhan Data

Tugas

1. Andaikan Anda diminta untuk memprediksi sebuah transaksi keuangan yang menggunakan Kartu Kredit kedalam kategori transaksi legal atau fraud.
2. Jelaskan variabel/atribut/fitur apa saja yang harus Anda kumpulkan yang terkait dengan:
    a) Identitas Pemegang Kartu Kredit
    b) Identitas Kartu Kredit yang digunakan
    c) Tempat melakukan transaksi Kartu Kredit
    d) Barang/layanan yang dibeli menggunakan Kartu Kredit
    e) Tempat/alamat terjadinya transaksi
    f) Waktu dilakukannya transaksi

Latihan Praktek: Mengumpulkan Data

Tugas:

1. Unduhlah file Bank Marketing dataset di link:UCI Marketing Repository https://archive.ics.uci.edu/ml/datasets/bank+marketing
2. Klik link: `Data Folder`
3. Unduh file: `bank.xip` dan tempatkan pada sebuah direktori komputer Anda.
4. Bukalah file `bank.xip`
5. Unggahlah file: `bank-full.csv` ke dalam direktori RapidMiner.
6. Apakah data yang diunduh dapat dipergunakan untuk menyelesaikan tujuan teknis: klasifikasi atau regresi? Jelaskan jawaban Anda.
7. Apakah ada fitur yang nilainya bukan numerik?
8. Teknik apa yang harus dilakukan untuk merubah nilai fitur yang bukan numerik menjadi numerik?

Latihan Praktek ke-1: Mengintegrasikan Data

Diberikan dua buah file data:

1) File pertama: bank-1.csv
2) File kedua: bank-2.csv
3) Periksa format fitur kedua file tersebut.
4) Apabila ada fitur didalam kedua file memiliki format data yang berbeda maka samakan format fitur kedua file tersebut

Tugas:

1) Samakan format fitur `bank-1.csv` dengan `bank-2.csv`
2) Gabungkan file hasil penyamaan format fitur diatas kedalam sebuah file `bank-3.csv`.
3) Unggahlah file hasil proses integrasi diatas ke dalam RapidMiner.

Latihan Praktek ke-2: Mengintegrasikan Data

Diberikan dua buah file data:

1) File pertama: bank-4.csv (memiliki fitur target)
2) File kedua: bank-5.csv (tidak memiliki fitur target)

Tugas:

1) Gabungkan kedua file tersebut menggunakan RapidMiner

Referensi

Dokumen silabus VSGA Associate Data Analyst, Kementerian Komunikasi dan Informatika, 2024

Standard Kompetensi Kerja Nasional Indonesia No 299 Tahun 2020 Bidang Keahlian Artificial Intelligence sub bidang Data science: https://skkni.kemnaker.go.id/tentang-skkni/dokumen

Standard Kompetensi Kerja Nasional Indonesia No 282 Tahun 2016 Bidang Software Development sub bidang Pemrograman : https://skkni.kemnaker.go.id/tentang-skkni/dokumen

Joel Grus, “Data Science from Scratch: First Principles with Python”, 2nd Edition, O’Reilly 2019

J.62DMI00.004.1, unit Kompetensi Mengumpulkan Data

https://github.com/kevinadhiguna/dqlab-career-track

Tools yang Digunakan

RapidMiner Studio Educational 09.10.011

KOMDIGI
Kementerian Komunikasi dan Digital
Republik Indonesia

digitalent

Terima Kasih