ibimbing

Data Engineering Challenge

Keizia Purba

Contents

1 Extract

2 Transform

3 Load & Data Visualization

Overview

Sebagai seorang **Data Engineer**, kamu memiliki tugas untuk mengolah data dari berbagai sumber, melakukan proses **Extract, Transform, Load (ETL)**, sampai dengan mempresentasikannya dalam bentuk yang dapat digunakan oleh stakeholder. Kamu akan diberikan challenge dalam beberapa area, yaitu proses **Extract** dari berbagai sumber, seperti melakukan **Web Scraping** dan mengintegrasikan **API**, kemudian melakukan **Data Cleansing** dari data yang sudah kita kumpulkan, serta melakukan **Data Transformation** menjadi format yang sesuai dengan kebutuhan.

Extract

Pada tahap ini, terdapat 2 metode yang dilakukan untuk memperoleh raw data:

1. Web Scrapping
   - Mengambil data Gross Domestic Product (GDP) dari provinsi - provinsi di Indonesia di tahun 2022 menggunakan url: https://en.wikipedia.org/wiki/List_of_Indonesian_provinces_by_GDP
   - Mengubah data kedalam bentuk DataFrame menggunakan library Pandas
   - Lakukan data cleaning
2. Application Programming Interface (API)
   - Menggunakan Geocoding API dari website: https://opencagedata.com/ untuk mendapatkan key API, latitude dan longitude
   - Mengubah data kedalam bentuk DataFrame

Hasil data dari Web Scrapping

	rank	province	region	gdp_in_billion_rp	gdp_in_billion_usd	gdp_ppp_in_billion_usd
5	1	Jakarta	Java	3,186,470	214.59	669.63
6	2	East Java	Java	2,730,907	183.91	573.89
7	3	West Java	Java	2,422,782	163.16	509.14
9	4	Central Java	Java	1,560,899	105.12	328.02
11	5	Riau	Sumatra	991,589	66.78	208.38

Hasil data dari API

	province	latitude	longitude
0	Jakarta	-6.175247	106.827049
1	East Java	-7.697740	112.491420
2	West Java	-6.889190	107.640472
3	Central Java	-7.303241	110.004414
4	Riau	0.500411	101.547581
5	North Sumatra	2.192352	99.381220

Transform

Setelah melakukan ektraksi data, selanjutnya melakukan transformasi data. Terdapat 3 tahapan yang dilakukan, yaitu:

1. Data Ingestion
   - Pada tahap ini, dilakukan penggabungan 2 data yang telah didapatkan pada proses ekstraksi data
2. Data Cleaning
   - Mengubah tipe data yang telah digabungkan, agar tipenya sesuai
3. Data Enrichment
   - Menambahkan satu kolom baru yang merupakan gabungan dari kolom latitude dan longitude

Data Ingestion

```
# Merge DataFrame for Extraction
df = pd.merge(gdp_prov, components_df, on ='province')
df.head()
```

	rank	province	region	gdp_in_billion_rp	gdp_in_billion_usd	gdp_ppp_in_billion_usd	latitude	longitude
0	1	Jakarta	Java	3,186,470	214.59	669.63	-6.175247	106.827049
1	2	East Java	Java	2,730,907	183.91	573.89	-7.697740	112.491420
2	3	West Java	Java	2,422,782	163.16	509.14	-6.889190	107.640472
3	4	Central Java	Java	1,560,899	105.12	328.02	-7.303241	110.004414
4	5	Riau	Sumatra	991,589	66.78	208.38	0.500411	101.547581

Data Cleaning

```
# Membersihkan nilai dari tanda koma dan mengubah tipe data ke integer
df['gdp_in_billion_rp']= df['gdp_in_billion_rp'].str.replace(',', '').astype(int)

#Change data type to integer
df[['rank', 'gdp_in_billion_rp']]= df[['rank', 'gdp_in_billion_rp']].astype(int)

#Change data type to float
df[['gdp_in_billion_usd', 'gdp_ppp_in_billion_usd']]= df[['gdp_in_billion_usd', 'gdp_ppp_in_billion_usd']].astype(float)

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 34 entries, 0 to 33
Data columns (total 8 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   rank                   34 non-null     int64
 1   province               34 non-null     object
 2   region                 34 non-null     object
 3   gdp_in_billion_rp      34 non-null     int64
 4   gdp_in_billion_usd     34 non-null     float64
 5   gdp_ppp_in_billion_usd  34 non-null     float64
 6   latitude               34 non-null     float64
 7   longitude              34 non-null     float64
dtypes: float64(4), int64(2), object(2)
memory usage: 2.4+ KB
```

Data Enrichment

```
#Add column lat_long
df['lat_long'] = df['latitude'].astype(str) + ',' + df['longitude'].astype(str)

df.head()
```

province	region	gdp_in_billion_rp	gdp_in_billion_usd	gdp_ppp_in_billion_usd	latitude	longitude	lat_long
Jakarta	Java	3186470	214.59	669.63	-6.175247	106.827049	-6.175247,106.8270488
East Java	Java	2730907	183.91	573.89	-7.697740	112.491420	-7.6977397,112.4914199
West Java	Java	2422782	163.16	509.14	-6.889190	107.640472	-6.8891904,107.6404716
Central Java	Java	1560899	105.12	328.02	-7.303241	110.004414	-7.3032412,110.0044145
Riau	Sumatra	991589	66.78	208.38	0.500411	101.547581	0.5004112,101.5475811

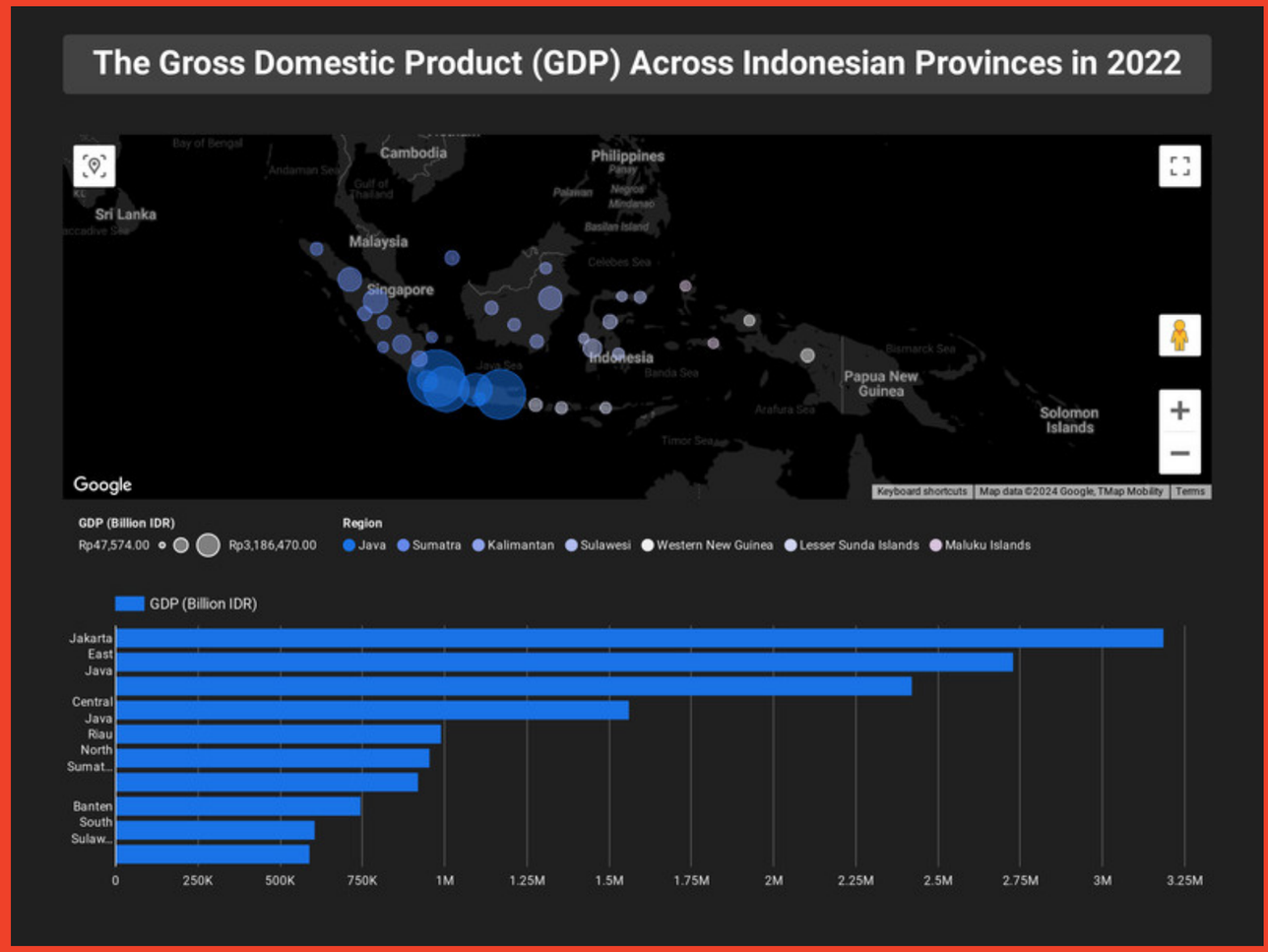
Load & Data Visualization

Setelah proses **Extract** dan **Transform** selesai, langkah selanjutnya adalah melakukan **Load** dengan mengekspor data ke dalam format file CSV. File CSV tersebut dapat digunakan untuk **memvisualisasikan data**, memudahkan pemahaman informasi yang terkandung dalam data.

Load

```
df.to_csv('etl_project.csv', index=False)
```

Data Visualization



Data Sources & Guide

- https://en.wikipedia.org/wiki/List_of_Indonesian_provinces_by_GDP
- https://opencagedata.com
- https://drive.google.com/file/d/1bfAlt4rdr5laB8Gil2oOc51rlOzmmsl3/view

ibimbing

Thank you

For your time

@keiziapurba