HACKTIV8

BUILDING AN ETL PIPELINE: DEKORUMA E-COMMERCE DATA

COMPREHENSIVE DATA ANALYTICS PROGRAM

amandarizkipati@gmail.com

› Start Slide

HACKTIV8

REFERENCE

WEBSITE

Website Dekoruma

COLAB NOTEBOOK

Project Link (Available on Github)

amandarizkipati@gmail.com

HACKTIV8

Objective of The Project

- Build an end-to-end ETL pipeline with a manual staging step.
- Practice web scraping, data cleaning, and data preparation in Python.
- Store the processed data in a CSV file before database ingestion.
- Understand the workflow of combining automated processing with manual data import into PostgreSQL.
- Strengthen familiarity with Python–PostgreSQL integration via CSV workflows.

Project Overview

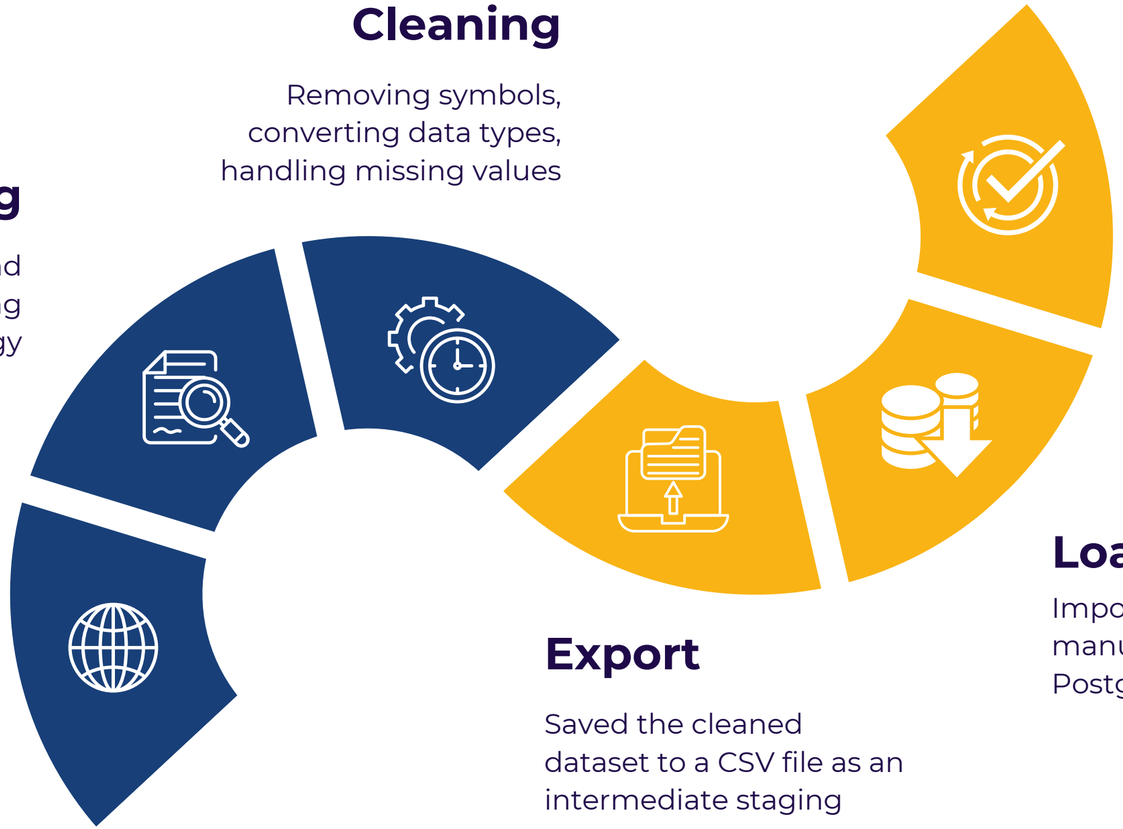
Step by Step

Cleaning

Removing symbols, converting data types, handling missing values

Pre-Processing

Data inspection and planning cleaning strategy



Final Outcome

A clean dataset ready for analysis

Extract

Extracted e-commerce product data from the Dekoruma website.te

Export

Saved the cleaned dataset to a CSV file as an intermediate staging output.

Load

Imported the CSV file manually into a PostgreSQL database.

HACKTIV8

Libraries Needed

Pandas	Tabular data manipulation.
Numpy	Numerical operations.
BeautifulSoup	HTML parsing.
Requests	HTTP requests (as backup).
Selenium	Browser automation (handling dynamic content).
Time	Delay control during scraping.

Website Scraping

Aspect	Details
Scraping Method	Selenium + BeautifulSoup
Pages Collected	5 pages
Scrolling	Automated scrolling to load all products
Captured Fields	• Nama Produk • Harga Sebelum Diskon • Diskon • Harga Setelah Diskon • Jumlah Terjual atau Ulasan • Gratis Ongkir
Total Records	120 rows

	Nama Produk	Harga Sebelum Diskon	Diskon	Harga Setelah Diskon	Jumlah Terjual atau Ulasan	Gratis Ongkir
0	Heim Studio KOZU Coffee Table 2in1	Rp1,485,000	42%	Rp859,000	(4055 Terjual)	se-Jabodetabek
1	Heim Studio KOZU Meja Kerja	Rp1,980,000	47%	Rp1,049,000	(3612 Terjual)	se-Jabodetabek
2	Heim Studio NANA Meja TV 2in1	Rp2,178,000	52%	Rp1,029,000	(2661 Terjual)	se-Jabodetabek
3	Heim Studio ISY Meja TV	Rp1,434,000	39%	Rp869,000	(3153 Terjual)	se-Jabodetabek
4	Heim Studio YVER Meja Makan Multifungsi	Rp3,636,000	48%	Rp1,859,000	(3135 Terjual)	se-Jabodetabek
...	...	...	...	...	...	...
115	Tenzo Living ALANO Meja Tamu	Rp8,670,000	25%	Rp6,499,000	None	se-Jabodetabek
116	Tenzo Living VERLIN Meja Tamu	Rp10,000,000	25%	Rp7,499,000	None	se-Jabodetabek
117	Heim Studio SOKA Meja Makan	Rp4,053,000	42%	Rp2,349,000	(5 Ulasan)	se-Jabodetabek
118	Cubic Cubic Meja Komputer Rumah - Meja Kerja L...	Rp430,000	50%	Rp215,000	(34 Terjual)	se-Jawa
119	Cubic Cubic Rak TV minimalis - Meja TV Minimal...	Rp1,397,000	57%	Rp592,000	(1 Ulasan)	se-Jawa

HACKTIV8

Data Exploration

No.	Column Name	Description	Non-Null Count/ Total	Data Type	Notes
1	Nama Produk	Product names scraped from Dekoruma	120/120	object	No missing values; data type correct (text)
2	Harga Sebelum Diskon	Original prices before discounts	120/120	object	Should be integer; must remove “Rp” and commas
3	Diskon	Discount percentages	114/120	object	Should be float; contains missing values (products without discounts); must remove “%” and scale to decimals
4	Harga Setelah Diskon	Prices after discounts	120/120	object	Should be integer; must remove “Rp” and commas
5	Jumlah Terjual atau Ulasan	Number sold or reviews	101/120	object	Missing values, contains brackets “()”
6	Gratis Ongkir	Free shipping information	120/120	object	No missing values; data type correct (text)

HACKTIV8

Data Cleaning

Column	Cleaning Actions	Result
Harga Sebelum Diskon	• Remove “Rp”and commas • Convert to integer	Int64, clean format
Diskon	• Remove “%” • Convert to float • Divide by 100 (decimal) • Replace NaN with 0	float64, no missing values
Harga Setelah Diskon	• Remove “Rp” and commas • Convert to integer	Int64, clean format
Jumlah Terjual/Ulasan	• Remove brackets “()” • Replace missing values with “Belum Terjual atau Diulas”	object, no missing values
Nama Produk & Gratis Ongkir	No changes needed	object, unchanged

HACKTIV8

Data Exploration (After Cleaning)

Column	Non-Null Count / Total	Data Type
Nama Produk	120 / 120	object
Harga Sebelum Diskon	120 / 120	Int64
Diskon	120 / 120	float64
Harga Setelah Diskon	120 / 120	Int64
Jumlah Terjual/Ulasan	120 / 120	object
Gratis Ongkir	120 / 120	object

All data cleaned and ready for further analysis.

Save Data to CSV
cleaned_retail_data.csv
(without index)

SQL Preparation

Create PostgreSQL database and table for staging

Use COPY command to load CSV

If required, normalize fields into separate tables

```sql
-- 1. Membuat Database
-- Query ini digunakan untuk membuat database
CREATE DATABASE milestone1; --(Note: Ini seharusnya berada pada halaman query yang berbeda, tetapi saya copy di sini  agar menjadi 1 file sql)

-- 2. Membuat Table Staging
-- Query ini digunakan untuk membuat table "staging"
CREATE TABLE staging (
                id SERIAL,
                nama_produk VARCHAR (250),
                harga_sebelum_diskon INT,
                diskon FLOAT,
                harga_setelah_diskon INT,
                jumlah_terjual_ulasan VARCHAR (250),
                gratis_ongkir VARCHAR (250)
                );

-- Query ini digunakan untuk mengambil semua data dari tabel "staging", tetapi dalam hal ini digunakan untuk memastikan apakah kolom pada table "staging" sudah sesuai
SELECT * FROM staging;

-- Query ini digunakan untuk memasukkan data pada tabel "staging", di mana data yang digunakan berasal dari file csv
COPY staging (nama_produk, harga_sebelum_diskon, diskon, harga_setelah_diskon, jumlah_terjual_ulasan, gratis_ongkir)
FROM 'C:\Users\cleaned_retail_data.csv' DELIMITER ',' CSV HEADER;
```

HACKTIV8

THANK YOU

COMPREHENSIVE DATA ANALYTICS PROGRAM

amandarizkipati@gmail.com

› End Slide