Female | Beauty

Repository

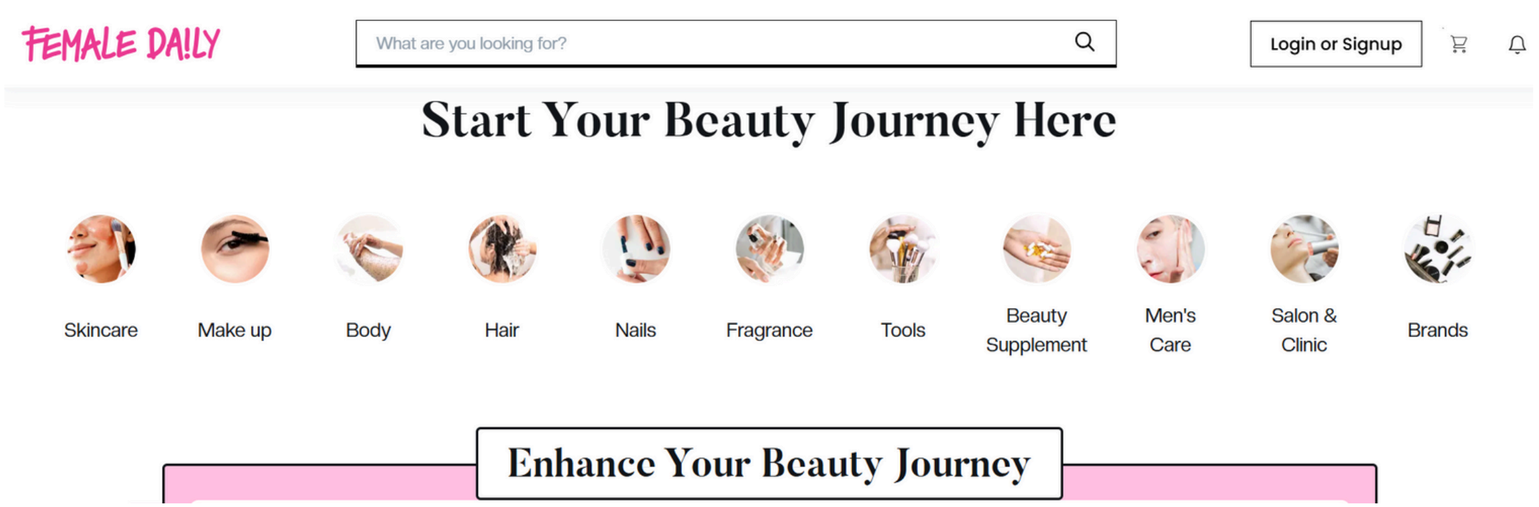
Data Pipeline for Skincare Product Reviews

created by Dyah **Isya**fira

Content

About Me

Bachelor of Science in Physics (Complex System Research Division) from **Bandung Institute of Technology (ITB)**. Passionate about transforming data into actionable insights and **continuously learning** to enhance decision-making. Currently working as a **Data Analyst** at a leading BNPL company, aspiring to **become an Analytics Engineer**.



Project Background

This project **automates data scraping** from the **Female Daily website**, collecting information on skincare categories, products, and reviews. A data pipeline is built to streamline data collection, transformation, and visualization, **fully integrated within Docker** using Selenium, Apache Airflow, and Metabase.

Why This Project Matters?

For Woman

who struggle to find the best skincare products, it **simplifies the decision-making process** by aggregating and analyzing reviews, **saving time and effort**. Instead of manually reading countless reviews, users can quickly access insights to choose the most suitable product.

For Businesses

This project provides a competitive advantage by enabling them to **monitor** their **product performance and analyze competitors** in the market. With structured data and visualized trends, brands can refine their strategies and improve their offerings based on consumer preferences.

Problem Statement

Many women **struggle to find the right skincare** products due to the overwhelming number of reviews and unstructured information online.

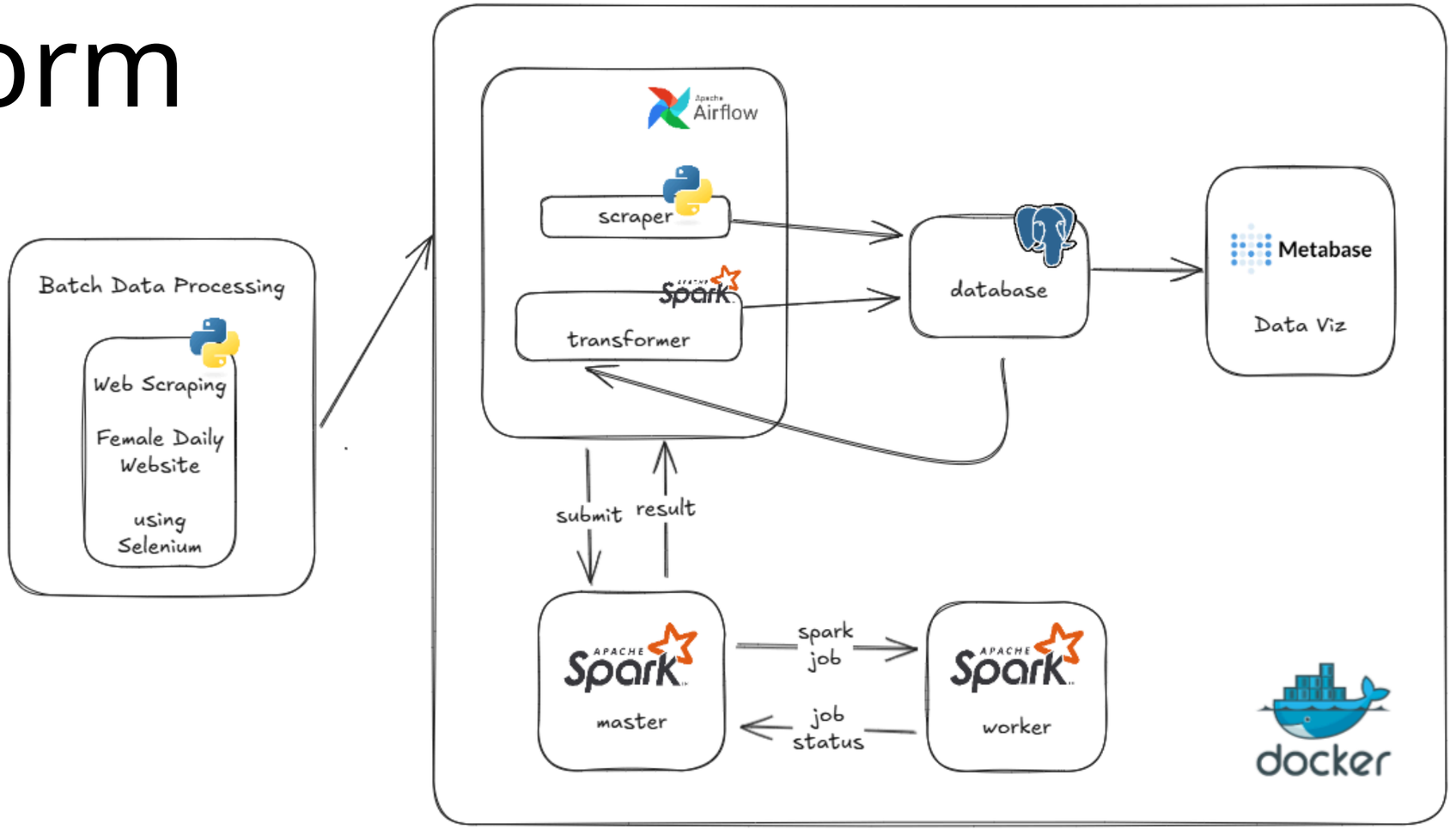
Manually reading countless reviews is **time-consuming** and inefficient.

Businesses **lack structured insights** into consumer preferences and market trends.

Project Goals

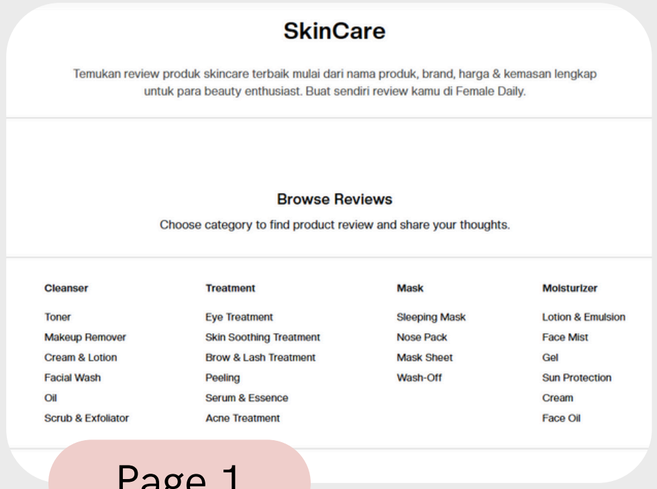
- Automate Data Collection
- Transform & Structure Data
- Visualize Insights

Data Platform



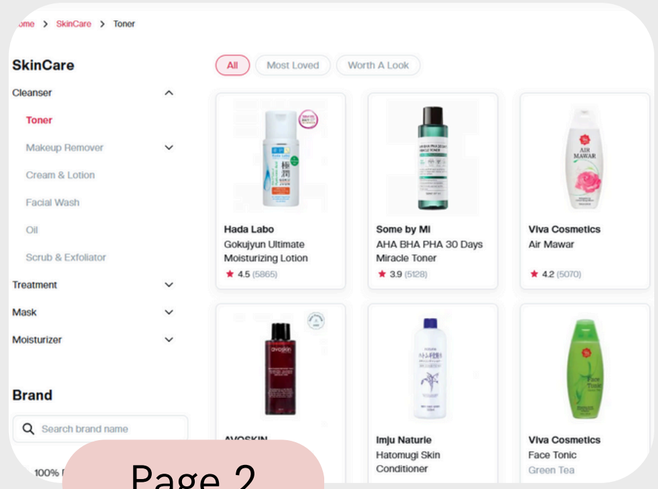
Data Understanding

Scrape from Female Daily/Skincare



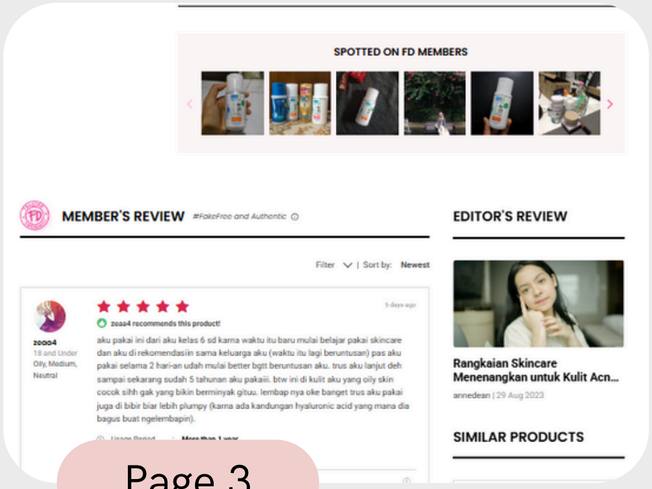
Page 1

Category Page



Page 2

Product Page

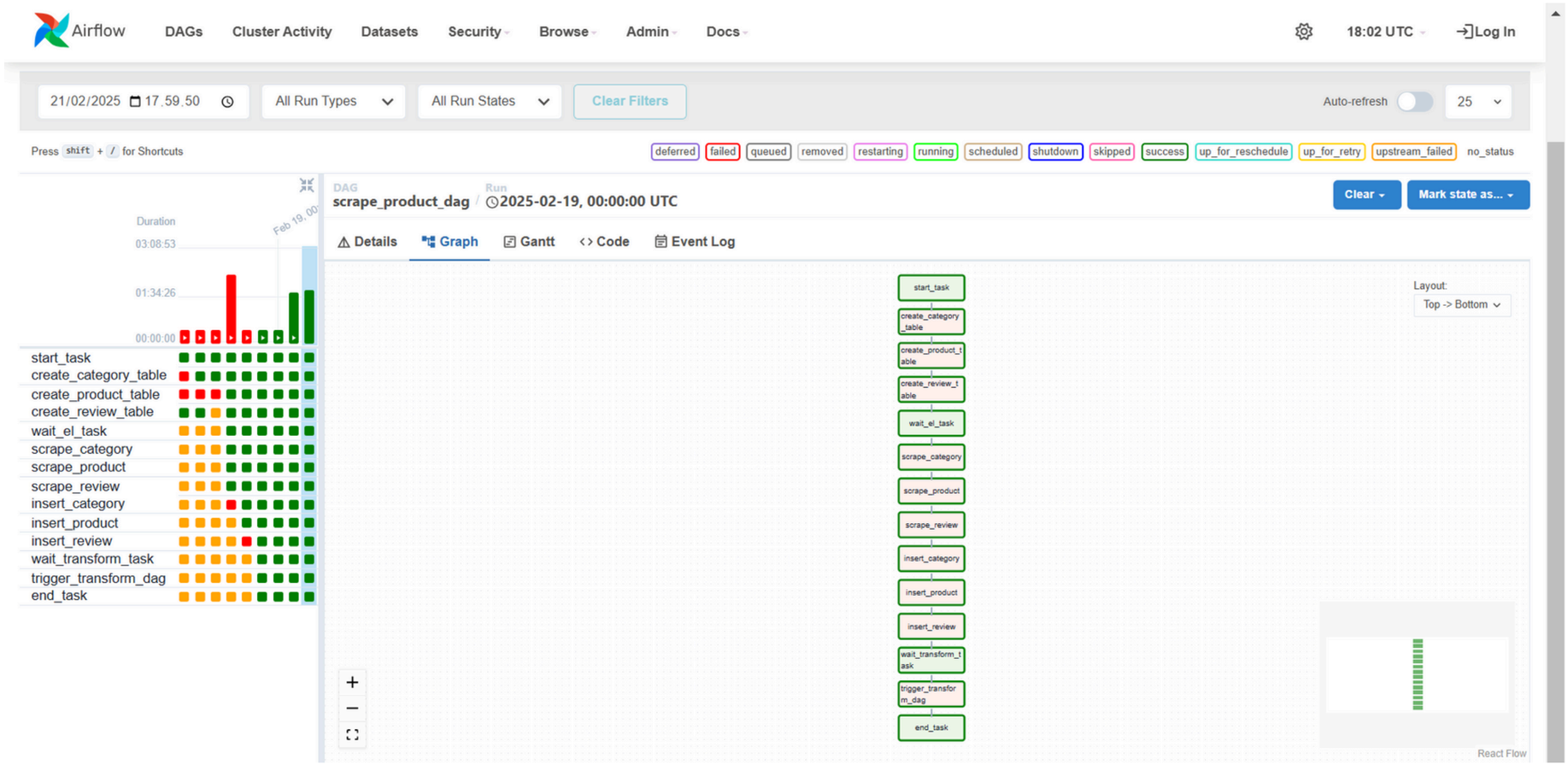


Page 3

Review Page

Data Understanding

scrape_product_dag.py



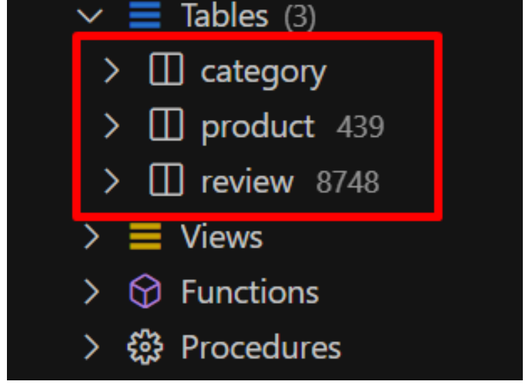
Data Understanding

scrape_product_dag.py

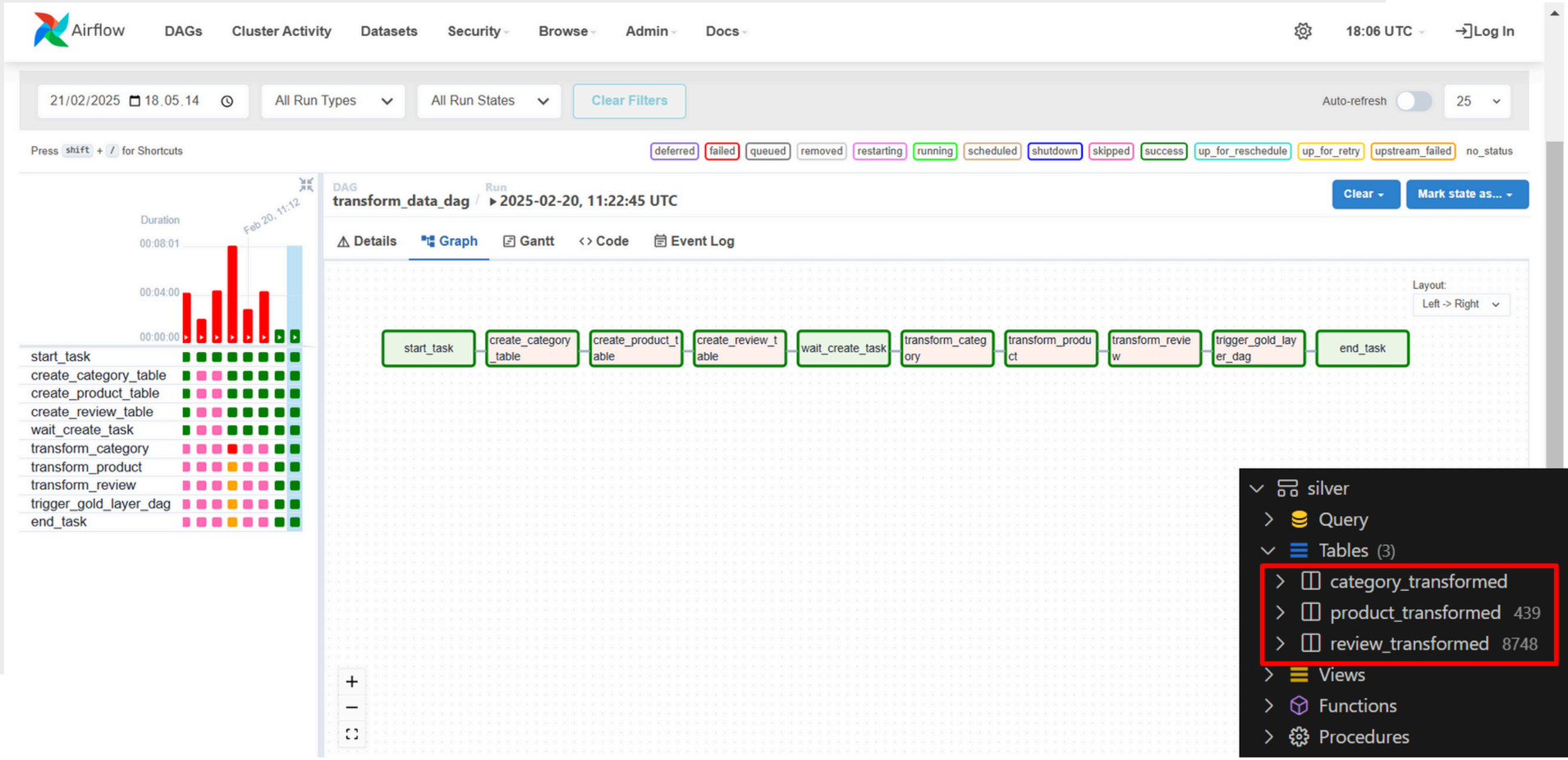
Category
Category Name
Category URL

Reviews
Reviewer Name
Reviewer Age
Reviewer Description
Rating
Thumb
Review Text
Usage Period
Purchase Point
Review Date

Product
Product Name
Brand Name
Rating
Reviews
Price
Product URL



Transformation



Category Transformation

transform_data_dag.py

index	Category	Link
0	Toner	https://reviews.femaledaily.com/products/cleanser/toner?brand=&order=popular&page=1
1	Makeup Remover	https://reviews.femaledaily.com/products/cleanser/face?brand=&order=popular&page=1
2	Cream & Lotion	https://reviews.femaledaily.com/products/cleanser/cream-lotion?brand=&order=popular&page=1
3	Facial Wash	https://reviews.femaledaily.com/products/cleanser/facial-wash?brand=&order=popular&page=1
4	Oil	https://reviews.femaledaily.com/products/cleanser/oil-2?brand=&order=popular&page=1
5	Scrub & Exfoliator	https://reviews.femaledaily.com/products/cleanser/scrub-exfoliator?brand=&order=popular&page=1
6	Eye Treatment	https://reviews.femaledaily.com/products/treatment/eye-serum?brand=&order=popular&page=1
7	Skin Soothing Treatment	https://reviews.femaledaily.com/products/treatment/skin-soothing-treatment?brand=&order=popular&page=1
8	Brow & Lash Treatment	https://reviews.femaledaily.com/products/treatment/brow-lash-treatment?brand=&order=popular&page=1
9	Peeling	https://reviews.femaledaily.com/products/treatment/peeling?brand=&order=popular&page=1
10	Serum & Essence	https://reviews.femaledaily.com/products/treatment/serum-essence?brand=&order=popular&page=1
11	Acne Treatment	https://reviews.femaledaily.com/products/treatment/acne-treatment?brand=&order=popular&page=1
12	Sleeping Mask	https://reviews.femaledaily.com/products/mask/sleeping-mask?brand=&order=popular&page=1
13	Nose Pack	https://reviews.femaledaily.com/products/mask/nose-pack?brand=&order=popular&page=1
14	Mask Sheet	https://reviews.femaledaily.com/products/mask/mask-sheet?brand=&order=popular&page=1
15	Wash-Off	https://reviews.femaledaily.com/products/mask/wash-off?brand=&order=popular&page=1
16		https://reviews.femaledaily.com/products/moisturizer/lotion-emulsion?brand=&order=popular&page=1
17		https://reviews.femaledaily.com/products/moisturizer/face-mist?brand=&order=popular&page=1
18		https://reviews.femaledaily.com/products/moisturizer/gel?brand=&order=popular&page=1
19		https://reviews.femaledaily.com/products/moisturizer/sun-protection-1?brand=&order=popular&page=1
20		https://reviews.femaledaily.com/products/moisturizer/cream-1?brand=&order=popular&page=1
21		https://reviews.femaledaily.com/products/moisturizer/face-oil?brand=&order=popular&page=1

- **Fill null** in **category_name** if it is null or empty using information extracted from the category_url
- **Removed duplicates** from category_name and category_url

Product Transformation

transform_data_dag.py

- **Brand and product names** are populated from the product_url when these values are missing.
- Remove () from **reviews column**
- Remove currency symbols and commas in **price column**

id integ	product_name varchar(255)	brand_name varchar(255)	rating varchar(50)	reviews varchar(50)	price varchar(50)	product_url text
1	Gokujyun Ultimate Moisturi	Hada Labo	4.5	(5865)	Rp. 29.000	https://reviews.femaledaily.c
2	AHA BHA PHA 30 Days Mira	Some by Mi	3.9	(5127)	Rp. 189.000	https://reviews.femaledaily.c
3					Rp. 4.800	https://reviews.femaledaily.c
4					Rp. 190.000	https://reviews.femaledaily.c
5	Miraculous Refining Toner	AVOSKIN	4.2	(4222)	Price not available	https://reviews.femaledaily.c
6	Hatomugi Skin Conditioner	Imju Naturie	4.3	(3626)	Rp. 115.000	https://reviews.femaledaily.c
7					Rp. 5.950	https://reviews.femaledaily.c
8					Rp. 200.000	https://reviews.femaledaily.c
9	Gokujyun Ultimate Moisturi	Hada Labo	4.3	(2567)	Rp. 30.000	https://reviews.femaledaily.c

Review Transformation

transform_data_dag.py

review_date is normalized to a standard date format, handling values such as "X days ago."

review_date varchar(100)
a day ago
8 hours ago
7 hours ago
7 days ago
6 days ago
5 days ago
4 hours ago
4 days ago

reviewer_description text
Oily, Medium, Neutral
Normal, Light, Neutral
Oily, Medium, Warm
Combination, Medium Light, Cool

skin_type varchar(50)	skin_tone varchar(50)	undertone varchar(50)
Oily	Medium	Neutral
Normal	Light	Neutral
Oily	Medium	Warm
Combination	Light	Cool

Skin type, skin tone, and undertone are extracted from the reviewer_description column

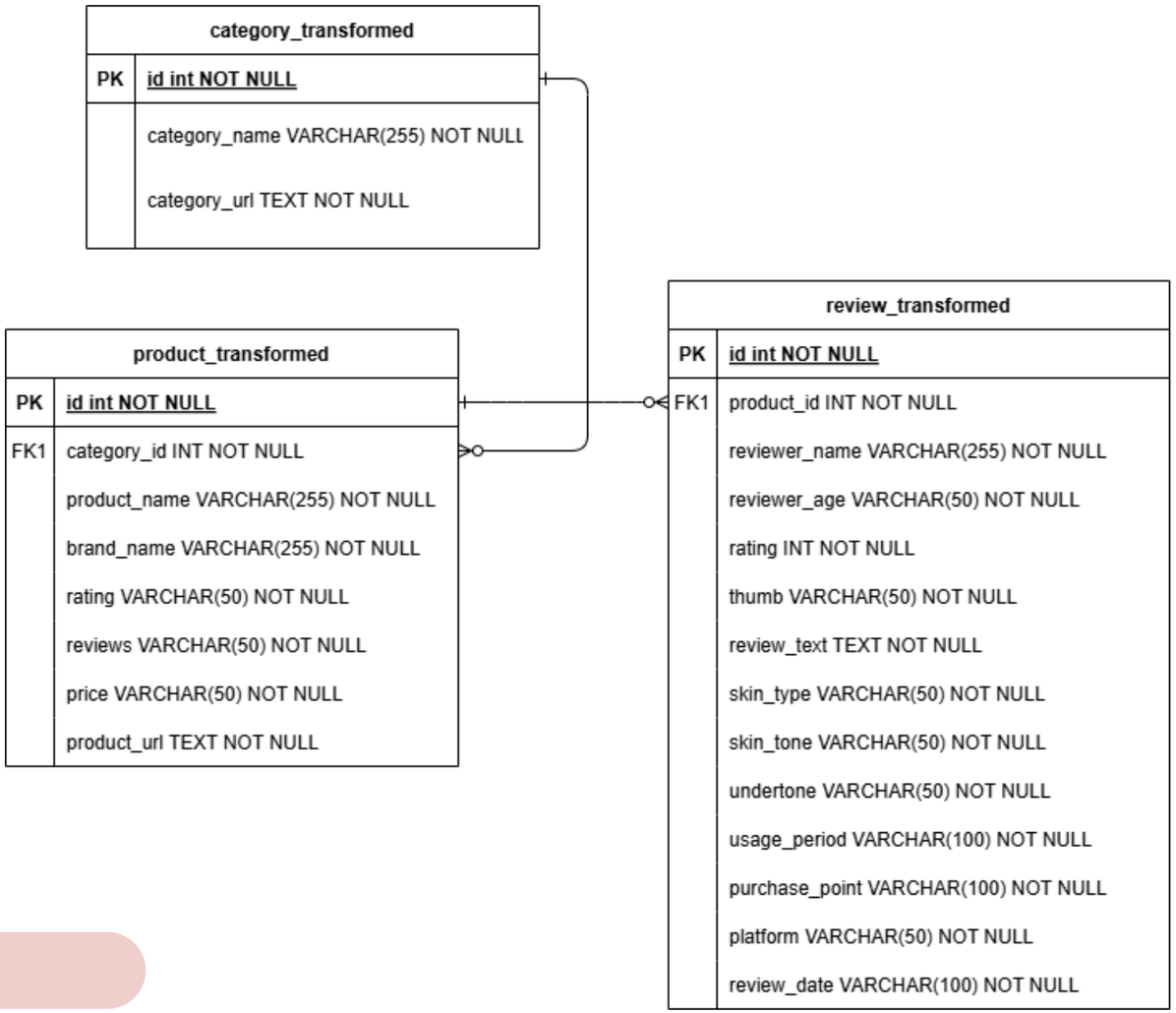
thumb varchar(50)
thumb_up
thumb_down

thumb varchar(50)
1
0

purchase_point varchar(100)	platform varchar(50)
Blibli.com	E-commerce
Female Daily Event	Event
Hypermart	Supermarket
Brand website	Official Website

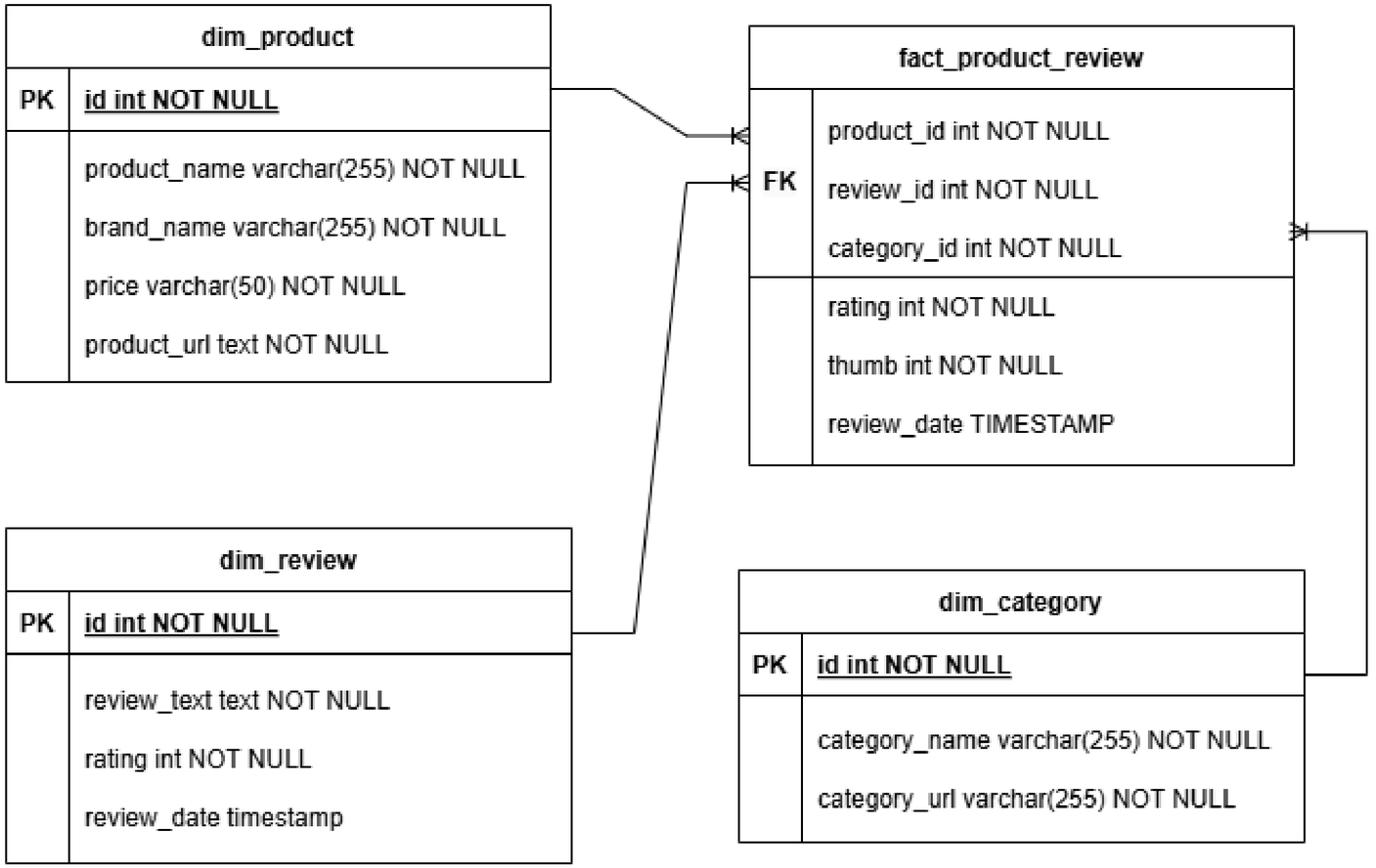
Mapped purchase_point into predefined categories (**platform**)

Data Modelling



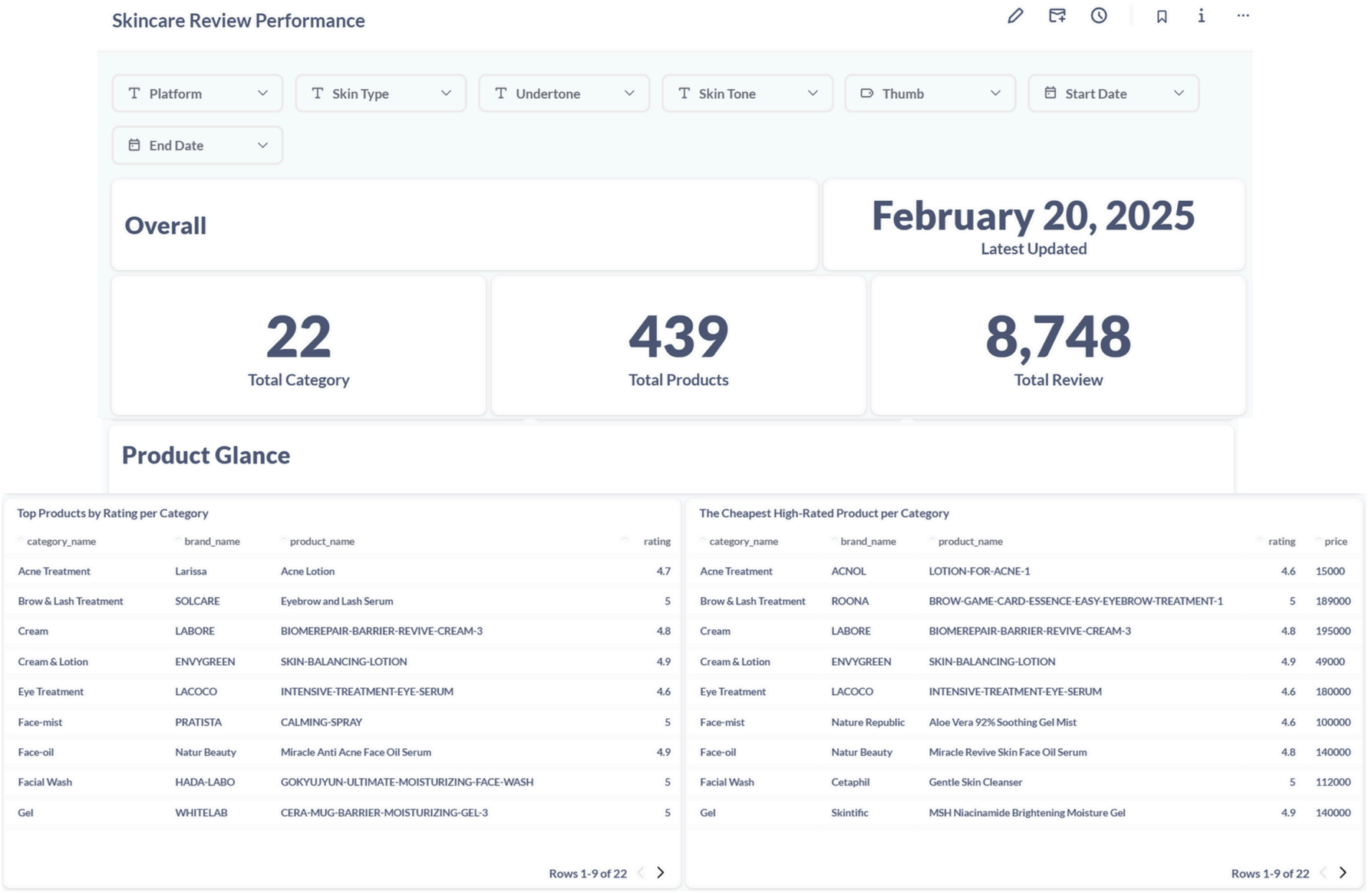
silver schema

Data Modelling



gold schema

Dashboard


Skincare Review Performance
Platform
Skin Type
Undertone
Skin Tone
Thumb
Start Date
End Date
Overall
February 20, 2025
Latest Updated
22
Total Category
439
Total Products
8,748
Total Review
Product Glance
Top Products by Rating per Category
category_name
brand_name
product_name
rating
Acne Treatment Larissa Acne Lotion 4.7
Brow & Lash Treatment SOLCARE Eyebrow and Lash Serum 5
Cream LABORE BIOMEREPAIR-BARRIER-REVIVE-CREAM-3 4.8
Cream & Lotion ENVYGREEN SKIN-BALANCING-LOTION 4.9
Eye Treatment LACOCO INTENSIVE-TREATMENT-EYE-SERUM 4.6
Face-mist PRATISTA CALMING-SPRAY 5
Face-oil Natur Beauty Miracle Anti Acne Face Oil Serum 4.9
Facial Wash HADA-LABO GOKYUJYUN-ULTIMATE-MOISTURIZING-FACE-WASH 5
Gel WHITELAB CERA-MUG-BARRIER-MOISTURIZING-GEL-3 5
Rows 1-9 of 22
The Cheapest High-Rated Product per Category
category_name
brand_name
product_name
rating
price
Acne Treatment ACNOL LOTION-FOR-ACNE-1 4.6 15000
Brow & Lash Treatment ROONA BROW-GAME-CARD-ESSENCE-EASY-EYEBROW-TREATMENT-1 5 189000
Cream LABORE BIOMEREPAIR-BARRIER-REVIVE-CREAM-3 4.8 195000
Cream & Lotion ENVYGREEN SKIN-BALANCING-LOTION 4.9 49000
Eye Treatment LACOCO INTENSIVE-TREATMENT-EYE-SERUM 4.6 180000
Face-mist Nature Republic Aloe Vera 92% Soothing Gel Mist 4.6 100000
Face-oil Natur Beauty Miracle Revive Skin Face Oil Serum 4.8 140000
Facial Wash Cetaphil Gentle Skin Cleanser 5 112000
Gel Skintific MSH Niacinamide Brightening Moisture Gel 4.9 140000
Rows 1-9 of 22

Dashboard


How these Review are distributed
Review per Undertone
Neutral
Warm
Cool
Unknown
57%
23%
16%
Review per Skin Type
Combina...
Oily
Dry
Normal
Unknown
44%
25%
17%
11%
Review per Skin Tone
Light
Medium
Dark
Unknown
50.1%
37.7%
Review per Usage Period
More than 1 year
1 month - 3 months
3 months - 6 months
1 week - 1 month
6 months - 1 year
Less than 1 week
Usage Period
Count
Review by Rating
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Count
1
2
3
4
5
5.5
Rating
Review per Reviewer Age
19 - 24
25 - 29
30 - 34
18 and Under
35 - 39
40 - 44
45 and Above
Reviewer Age
Count

Conclusion

Successful Pipeline: Built an automated pipeline using Selenium, Apache Spark, PostgreSQL, and Airflow for scraping, transforming, and storing skincare data.

Effective Data Transformation: Cleaned and structured raw data from bronze to silver schemas for easy access.

Data Reporting: The transformed data is stored in PostgreSQL and can be easily accessed for reporting and analysis via Metabase with interactive visualization, allowing end-users to explore insights on skincare products, categories, and customer reviews.

Recommendation

Scalability: The pipeline is scalable to accommodate larger datasets in the future. With tools like Apache Spark, it can efficiently handle increasing amounts of data.

Real-Time Data Processing: Consider implementing a streaming pipeline to capture real-time data updates. This would help deliver more up-to-date insights, especially for time-sensitive trends in the skincare market.

Enhanced Analytics: Future improvements can include incorporating more sophisticated analysis like sentiment analysis of reviews, customer segmentation, or integrating external datasets (e.g., competitor product data).

User Access & Collaboration

Thank You!

github.com/iisyafiraa

linkedin.com/in/dyahisyafira