



Cosmos



Apache
Airflow



dbt



Metabase



docker



Final Project Online Retail ELT

Kelompok 11 Dibimbing – Data Engineering



Kelompok 11,



Chintya Dewi
Prawitasuri



Navi Latul
Ulya



Dhevita Intan
Ervandra



Dimas Wahyu Saputro



Fionjufo
Fahrezi

Table of contents



01

Introduction

02

Project Description

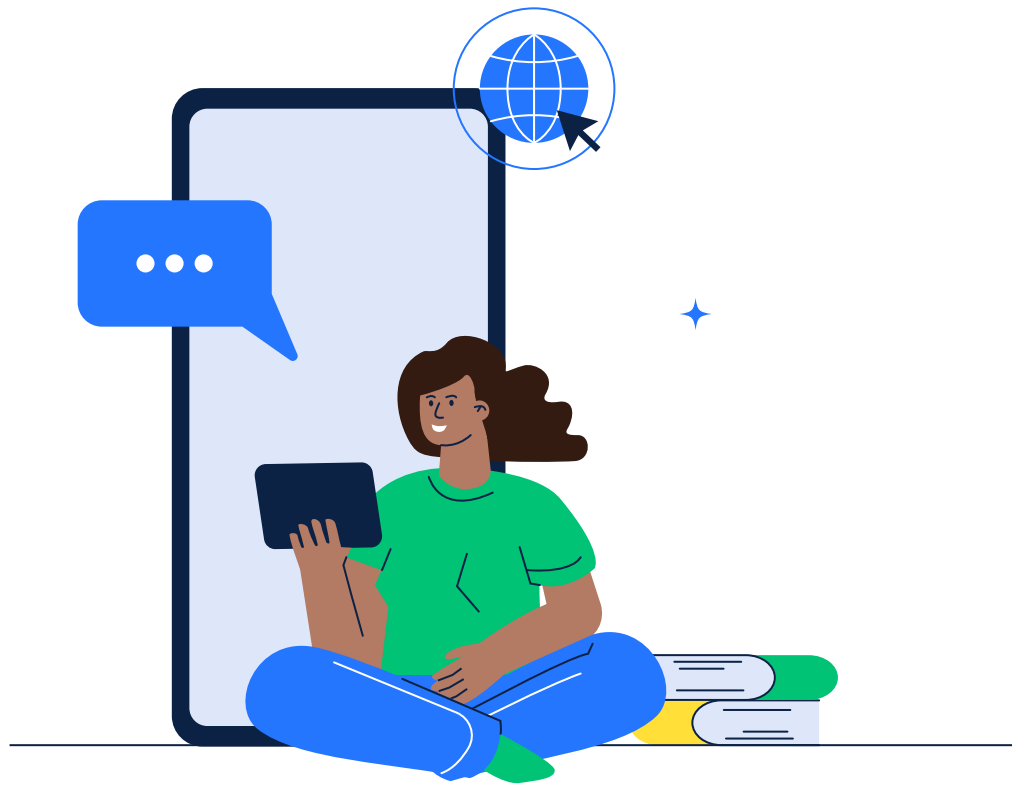
03

Analysis and Visualization

04

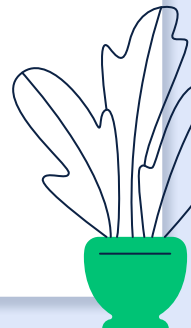
Future Improvement





01.

Introduction



Introduction



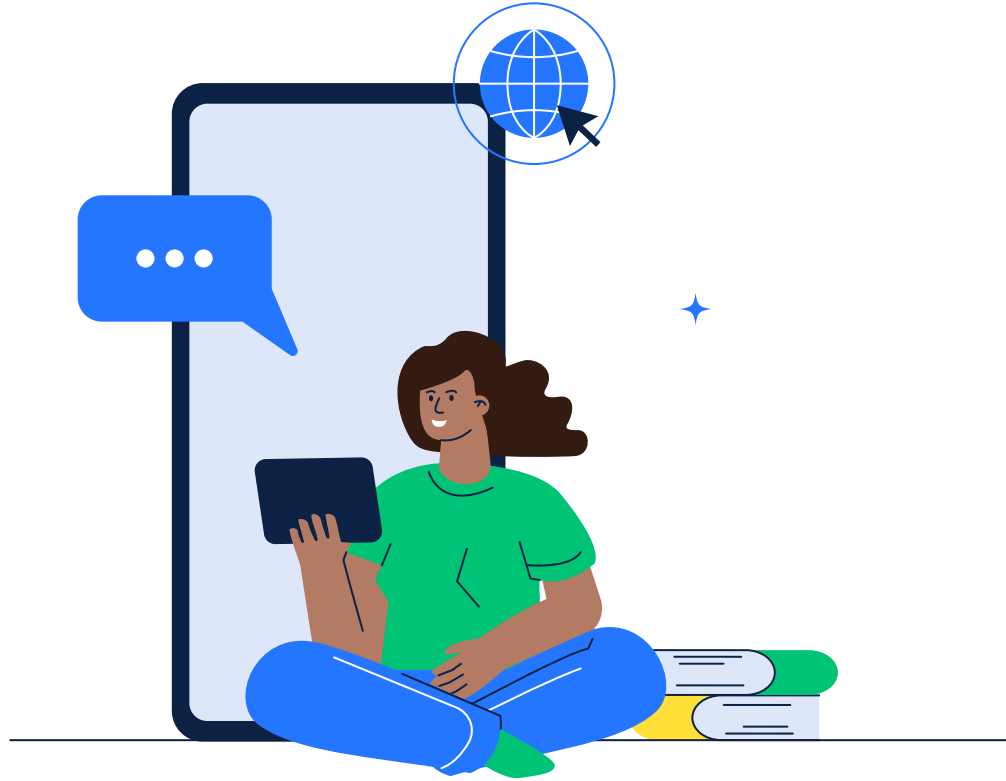
Backgrounds

Congratulations on your first role as Data Engineer!. You are just hired at a US online retail company that sells general customer products directly to customers from multiple suppliers around the world. **Your challenge is to build-up the data infrastructure using generated data crafted to mirror real-world data from leading tech companies.**

Tasks

- ETL/ELT **Job Creation** using Airflow
- **Data Modeling** in Postgres
- **Dashboard Creation** with Data Visualization
- **Craft a Presentation** Based on Your Work





02.

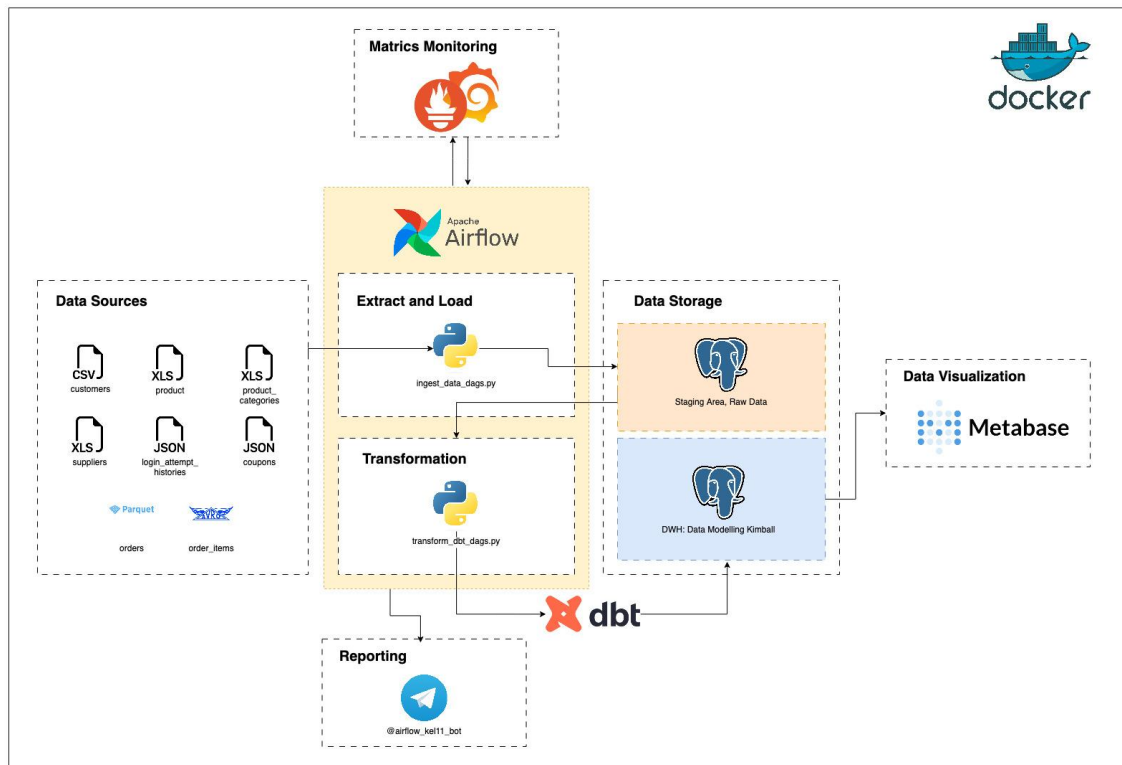
Project Description



Extract, Load, Transform



Data Platform Architecture Diagram



Tech Stack:

- Docker
- Python
- Airflow
- Postgresql
- Metabase
- **dbt**
- **Grafana**
- **Prometheus**
- **Telegram API**



Perubahan yang Dilakukan

01. Edit file requirements.txt

```
requirements.txt
1 pandas=2.1.0
2 psycogp2-binary
3 avro
4 pyarrow
5 sqlalchemy
6 xlrd
7 python-snappy
8 astronomer-cosmos[dbt-postgres]
9 apache-airflow-providers-telegram=4.1.1
```

02. Tambahkan .env baru

```
24
25 TELEGRAM_TOKEN=6830522859:AAE1UcudolrLTZhetqBx1psYK_1KQR4wFv8
26 TELEGRAM_CHAT_ID=844199573
```

03. Edit file docker-compose.yml

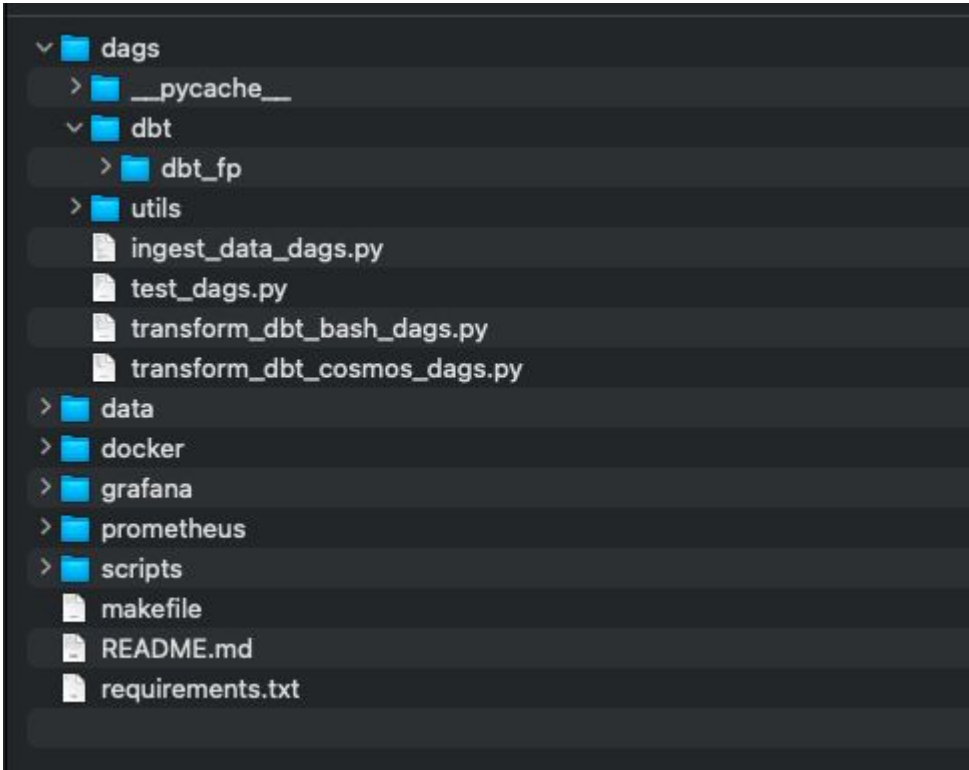
```
- AIRFLOW_CORE_DAG_FILE_PROCESSOR_TIMEOUT=999999
- AIRFLOW_CORE_DAGBAG_IMPORT_TIMEOUT=999999
- AIRFLOW_WEBSERVER_SECRET_KEY=123456789
- TELEGRAM_TOKEN=${TELEGRAM_TOKEN}
- TELEGRAM_CHAT_ID=${TELEGRAM_CHAT_ID}
- AIRFLOW_SCHEDULER_STATSD_ON=True
- AIRFLOW_SCHEDULER_STATSD_HOST=statsd-exporter
- AIRFLOW_SCHEDULER_STATSD_PORT=8125
- AIRFLOW_SCHEDULER_STATSD_PREFIX=airflow
```

di environment webserver dan scheduler

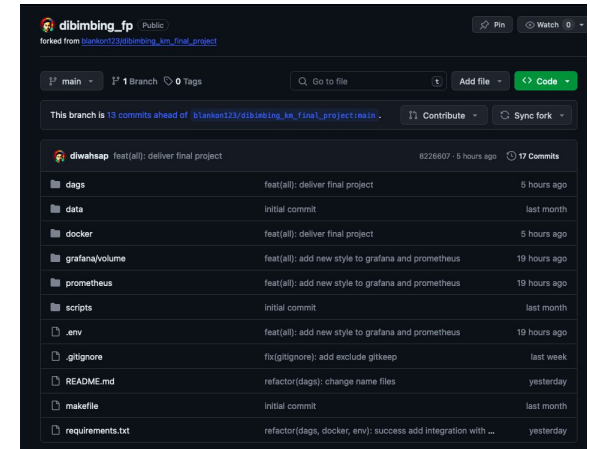
```
134 statsd-exporter:
135   image: prom/statsd-exporter
136   container_name: airflow-statsd-exporter
137   command: "--statsd.listen-udp=8125 --web.listen-address=:9102"
138   ports:
139     - 9123:9102
140     - 8125:8125/udp
141
142 prometheus:
143   image: prom/prometheus
144   container_name: airflow-prometheus
145   user: "0"
146   ports:
147     - 9090:9090
148   volumes:
149     - ../prometheus/prometheus.yml:/etc/prometheus/prometheus.yml
150     - ../prometheus/volume/prometheus
151
152 grafana:
153   image: grafana/grafana:7.1.5
154   container_name: airflow-grafana
155   environment:
156     GF_SECURITY_ADMIN_USER: admin
157     GF_SECURITY_ADMIN_PASSWORD: password
158     GF_PATHS_PROVISIONING: /grafana/provisioning
159   ports:
160     - 3000:3000
161   volumes:
162     - ../grafana/volume/data:/grafana
163     - ../grafana/volume/datasources:/grafana/datasources
164     - ../grafana/volume/dashboards:/grafana/dashboards
165     - ../grafana/volume/provisioning:/grafana/provisioning
```

Tambahkan services baru, ++ file tambahan

Struktur Folder

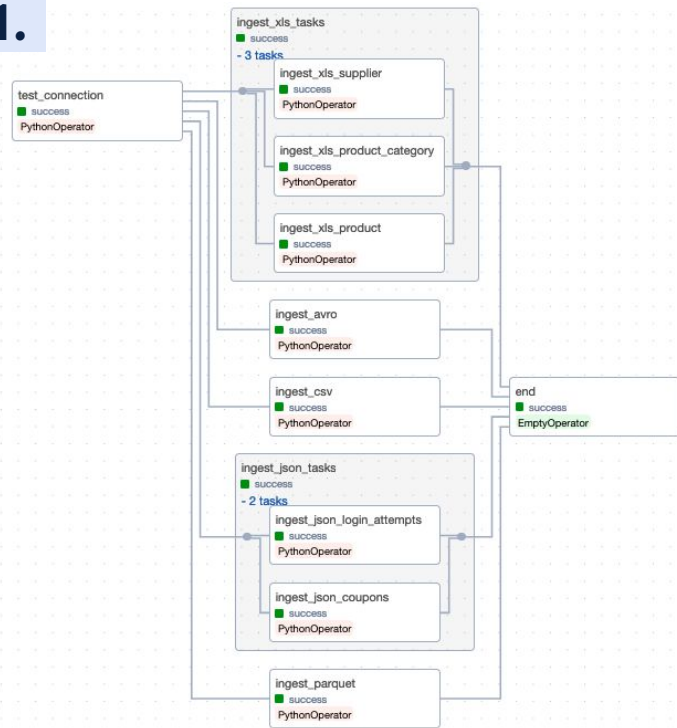


Dapat diakses di
Github



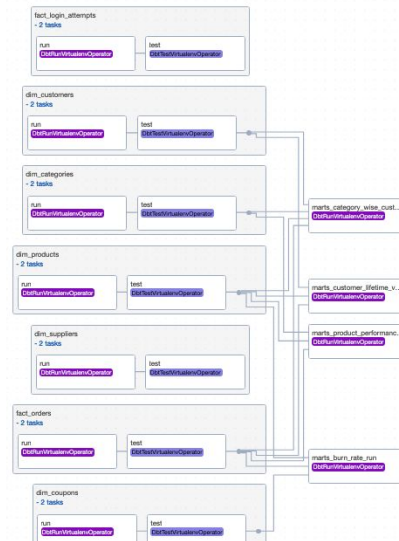
DAGs

01.



ingest_data_dags

02. (i) transform_dbt_bash_dags



02. (ii) transform_dbt_cosmos_dags

Notifikasi Telegram

The screenshot displays the Apache Airflow web interface in a browser. The main view shows the DAG 'ingest_data_dags' in the 'Graph' tab. The DAG is a linear workflow with the following tasks:

- test_connection (PythonOperator, success)
- ingest_xls_tasks (PythonOperator, success, + 3 tasks)
- ingest_avro (PythonOperator, success)
- ingest_csv (PythonOperator, success)
- ingest_json_tasks (PythonOperator, success, + 2 tasks)
- ingest_parquet (PythonOperator, success)
- end (EmptyOperator, success)



The DAG is scheduled to run at 2023-11-27, 07:00:00 WIB. The interface also shows a sidebar with a task list and a top navigation bar with various tabs like DAGs, Cluster Activity, Datasets, Security, Browse, Admin, and Docs.

Overlaid on the right side of the Airflow interface is a Telegram chat window. The chat is with a bot named 'kell1dibimbing'. It contains five messages, all indicating successful DAG runs:

- Airflow Run Successful! DAG: ingest_data_dags Task Name: ingest_json_tasks.ingest_json_coupons Duration 2.076779 10:56
- Airflow Run Successful! DAG: ingest_data_dags Task Name: ingest_xls_tasks.ingest_xls_product Duration 5.092607 10:56
- Airflow Run Successful! DAG: ingest_data_dags Task Name: ingest_parquet Duration 8.032941 10:56
- Airflow Run Successful! DAG: ingest_data_dags Task Name: ingest_json_tasks.ingest_json_login_attempt Duration 49.075862 10:57
- Airflow Run Successful! DAG: ingest_data_dags Task Name: ingest_avro Duration 138.904613 10:58
- Airflow Run Successful! DAG: ingest_data_dags Task Name: end Duration 0.210451 10:58

The Telegram chat window also shows a 'Write a message...' input field at the bottom.

Preview Raw Data at PSQL



	id	first_name	last_name	gender	address	zip_code
1	4,000	Nancy	Gonzalez	F	58724 Holloway Wall	68,028
2	4,001	Alex	Vargas	M	887 Cervantes Station Suite 099	98,241
3	4,002	Stacey	Lopez	F	5521 Brenda Villages	41,550
4	4,003	Christopher	Yoder	M	660 Francis Trail Suite 489	82,012
5	4,004	Megan	Chan	F	134 Patterson Locks	97,085
6	4,005	Karl	Bell	M	1374 Alexandra Village Suite 294	39,138
7	4,006	Sarah	Jefferson	F	1475 Nicholas Roads Apt. 302	68,019
8	4,007	Jason	Henry	M	562 Brown Rapid Apt. 128	60,619
9	4,008	Robin	Wade	F	507 Amanda Point	85,020
10	4,009	Blake	Oconnell	M	005 Khan Burg Suite 942	6,088
11	4,010	Felicia	Dominguez	F	7268 Dixon Cliffs Apt. 459	89,372
12	4,011	Larry	Mcclure	M	01412 Ronald Well	58,080
13	4,012	Laurie	Norris	F	607 Moore Crescent	99,855
14	4,013	Robert	Clark	M	09240 Tammy Lock	26,250
15	4,014	Kristy	Gonzales	F	6853 Kristina Unions Apt. 462	47,446
16	4,015	Joseph	Jimenez	M	62702 Koch Prairie	90,073
17	4,016	Chelsea	Kim	F	4966 James Ferry	81,620
18	4,017	Adam	Williams	M	690 David Shoals	8,865
19	4,018	Natalie	Green	F	28915 Herrera Road Apt. 888	39,467
20	4,019	Mark	Roberts	M	792 Jimenez Dam	73,108
21	4,020	Heather	Sanchez	F	51484 Kristin Ridge Apt. 871	58,210

customers

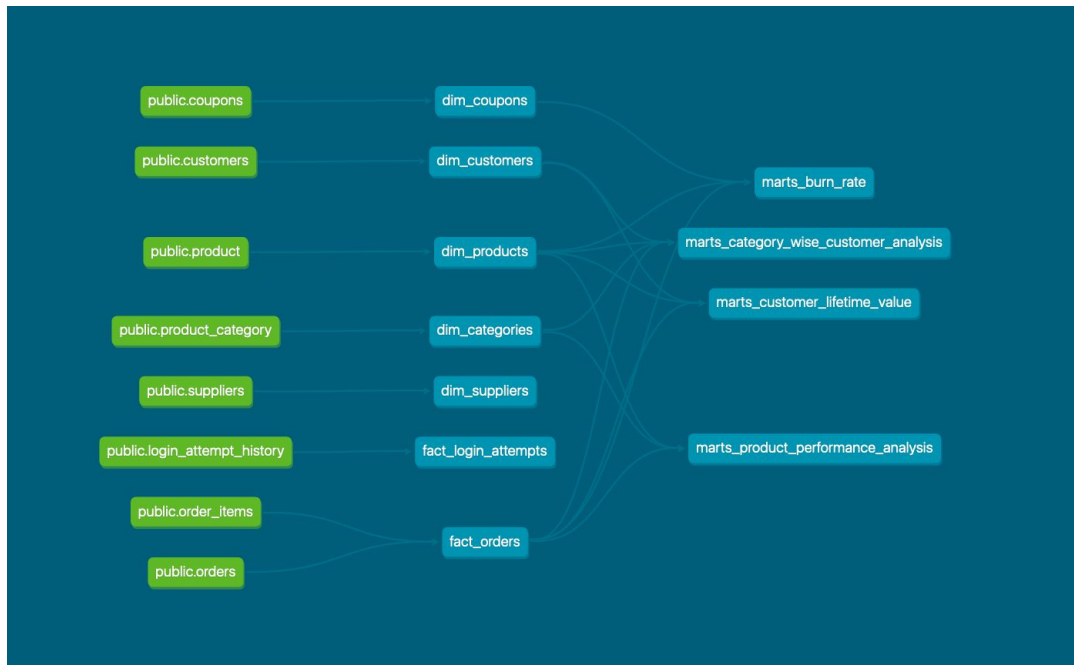
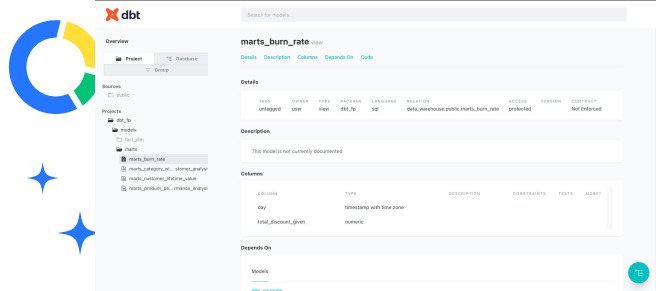
	id	customer_id	login_successful	attempted_at
1	159,424	8,207	[v]	2023-02-01 19:11:25.000
2	159,425	4,374	[v]	2023-02-16 21:45:35.000
3	159,426	2,329	[v]	2023-02-22 09:22:35.000
4	159,427	2,338	[v]	2023-01-10 02:43:28.000
5	159,428	5,658	[v]	2023-03-17 12:28:10.000
6	159,429	8,351	[v]	2023-10-09 10:28:14.000
7	159,430	7,494	[]	2023-09-26 13:28:40.000
8	159,431	4,478	[v]	2023-10-30 04:19:29.000
9	159,432	1,655	[v]	2023-02-22 08:53:59.000
10	159,433	4,516	[]	2023-01-03 18:03:53.000
11	159,434	623	[v]	2023-10-29 03:52:41.000
12	159,435	6,218	[v]	2023-09-05 07:11:44.000
13	159,436	9,745	[v]	2023-08-10 09:19:33.000
14	159,437	8,225	[v]	2023-04-23 10:52:50.000
15	159,438	906	[v]	2023-05-11 06:38:37.000
16	159,439	762	[v]	2023-07-06 15:01:41.000
17	159,440	1,632	[v]	2023-08-22 07:39:31.000
18	159,441	823	[v]	2023-09-24 05:15:44.000
19	159,442	3,392	[v]	2023-01-16 23:31:08.000

login_attempt_history

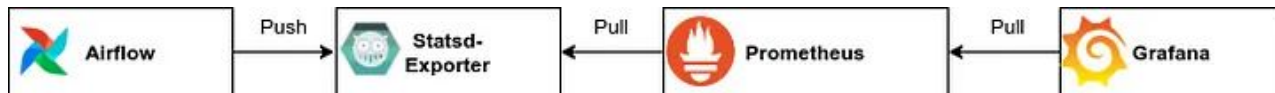
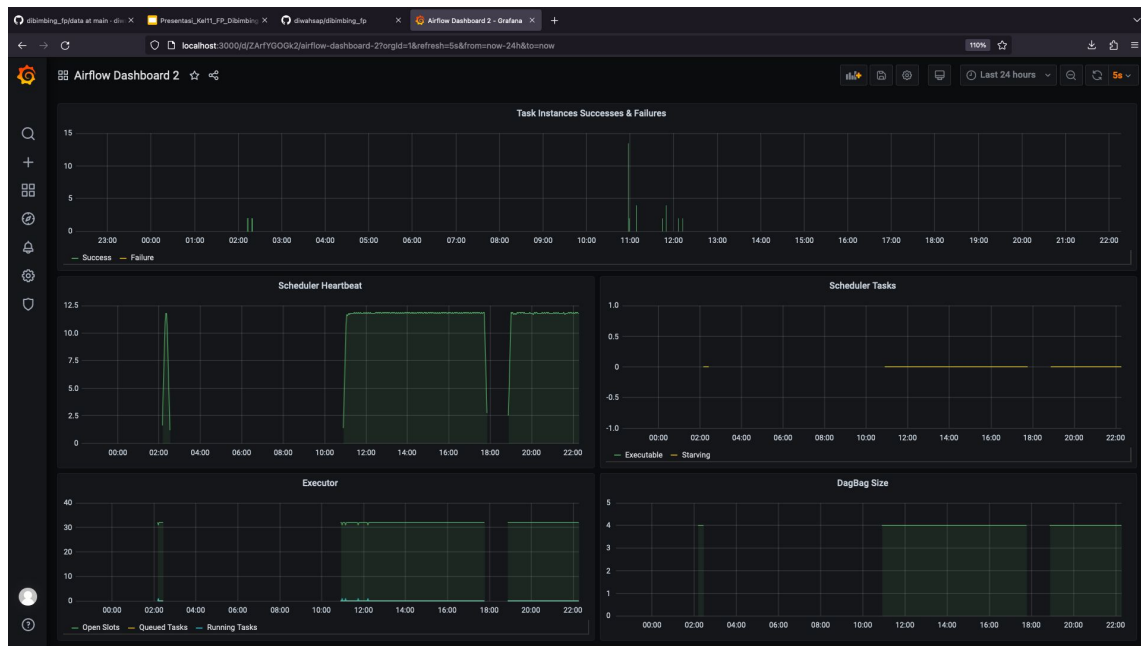
Data Lineage

Hal yang menarik dbt!

- Bisa **otomatis test** setiap data,
 - unique
 - not_null
 - Referenced
- **Otomatis generate dokumentasi** Data

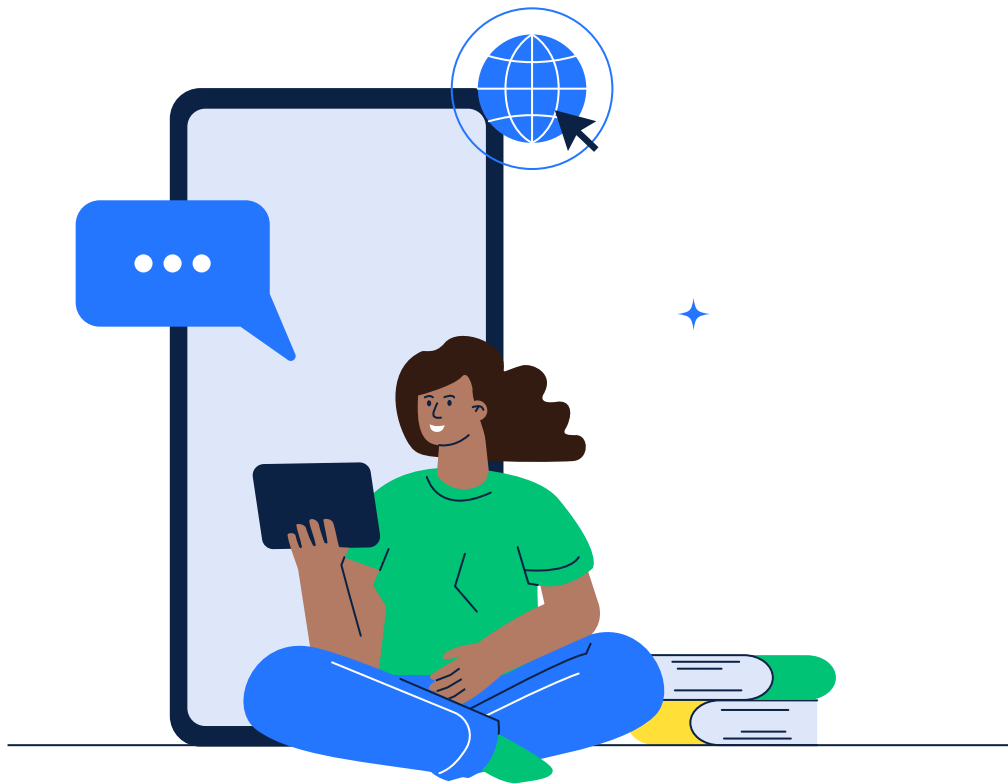


Metrics Monitoring, Grafana

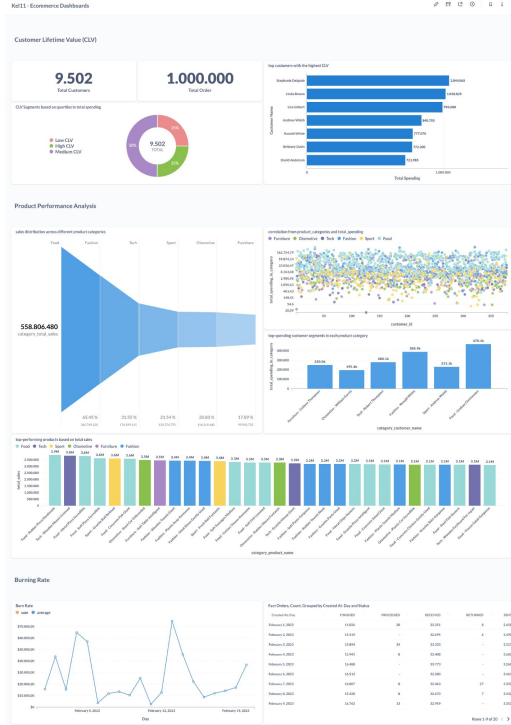


03.

Analysis and Visualization



Visualisasi Metabase



Cukup panjang ya, hehe..

Terdapat tiga bagian utama,

- Customer Lifetime Value (CLV)
- Product Performance Analysis
- Burning Rate

Customer Lifetime Value (CLV)



Kel11 - Ecommerce Dashboards



Customer Lifetime Value (CLV)

9.502

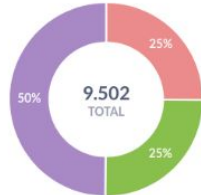
Total Customers

1.000.000

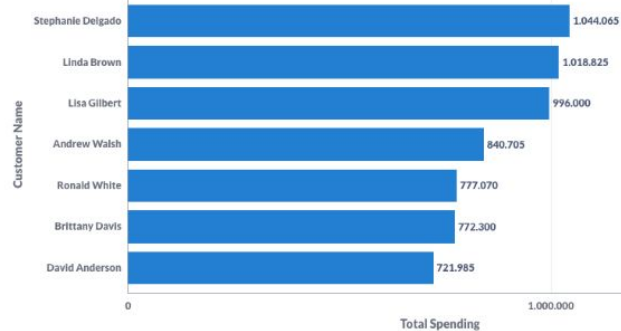
Total Order

CLV Segments based on quartiles in total spending

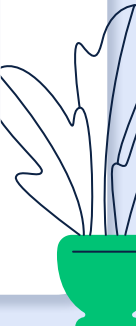
- Low CLV
- High CLV
- Medium CLV



top customers with the highest CLV



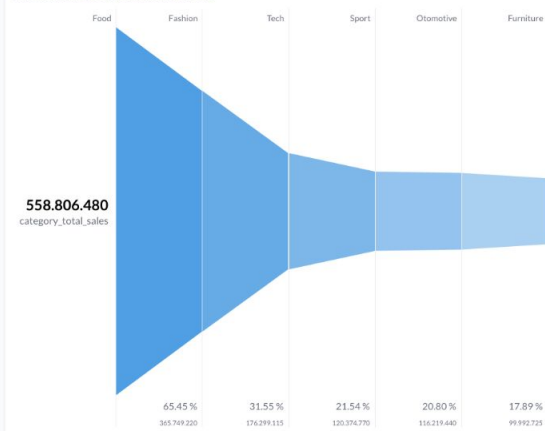
Mengidentifikasi **pelanggan bernilai tinggi** dan **memahami pola belanja**. Hal ini membantu dalam menyesuaikan strategi pemasaran dan meningkatkan retensi pelanggan.



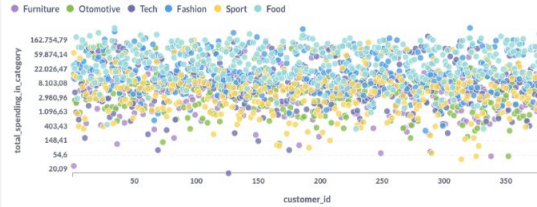
Product Performance Analysis

Product Performance Analysis

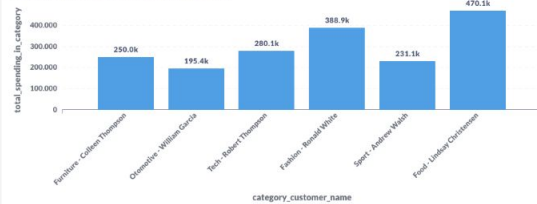
sales distribution across different product categories



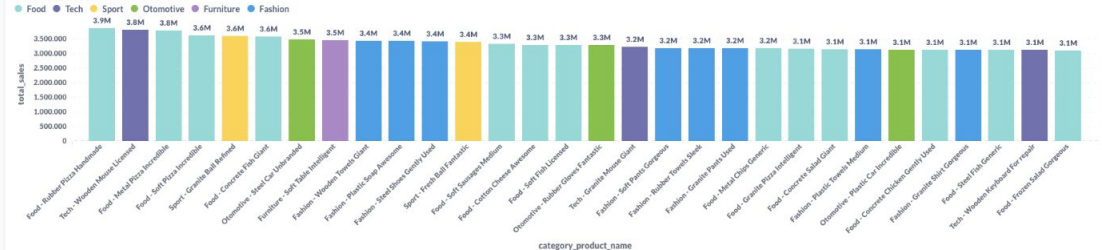
correlation from product categories and total spending



top-spending customer segments in each product category



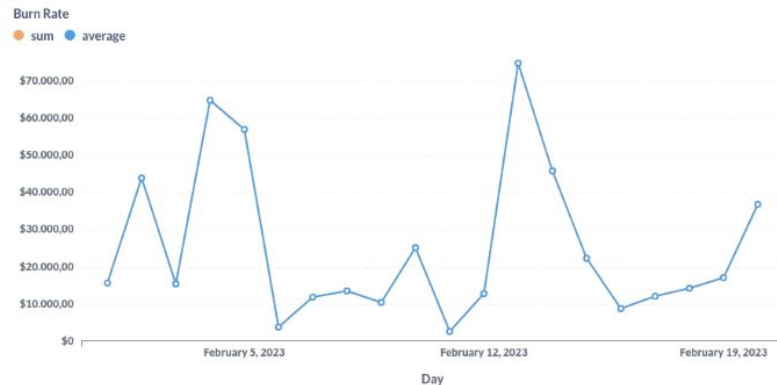
top-performing products based on total sales



Menyoroti produk dan kategori dengan performa terbaik. Gunakan data ini untuk mengelola inventaris secara efektif dan merencanakan strategi pengembangan produk.

Burning Rate

Burning Rate



Fact Orders, Count, Grouped by Created At: Day and Status

Created At: Day	FINISHED	PROCESSED	RECEIVED	RETURNED	SENT
February 1, 2023	11.026	20	22.251	8	2.458
February 2, 2023	15.419	-	32.695	6	3.490
February 3, 2023	15.894	34	33.333	-	3.317
February 4, 2023	15.941	8	32.408	-	3.682
February 5, 2023	16.488	-	33.773	-	3.260
February 6, 2023	16.512	-	32.580	-	3.461
February 7, 2023	16.807	8	32.463	27	3.392
February 8, 2023	15.428	8	32.670	7	3.432
February 9, 2023	16.762	13	32.959	-	3.352

Rows 1-9 of 20 < >

Ini mengukur tingkat di mana perusahaan membelanjakan modalnya.

04.

Future Improvement



Future Improvements



Use Datahub

Enables **data discovery**, **data observability** and **federated governance** to help the complexity of data ecosystem.



Use Cloud, and Automate with Terraform

In order to enhance the efficiency and scalability of IT infrastructure, future improvements could involve leveraging **cloud technologies and implementing automation** through tools such as Terraform. Embracing cloud computing allows for the **flexibility to scale resources on-demand, optimizing cost and performance**.



Thanks!

Do you have any questions?

youremail@freepik.com

+34 654 321 432

yourwebsite.com



CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution

