Final Project Presentation

Rizky Fajar Aditya – JCDEOL001

Released February 2025

Github: https://github.com/rizkyjarr/final-project-purwadhika

Table of Contents:

Study Case: GoHailing! company is facing risks that threatens its business sustainability

Tight market competition

GoHailing!, a provider of on-demand transportation services, is facing intense competition due to the rise of new players with the same business model, increasing market pressure.

Go Hailing!

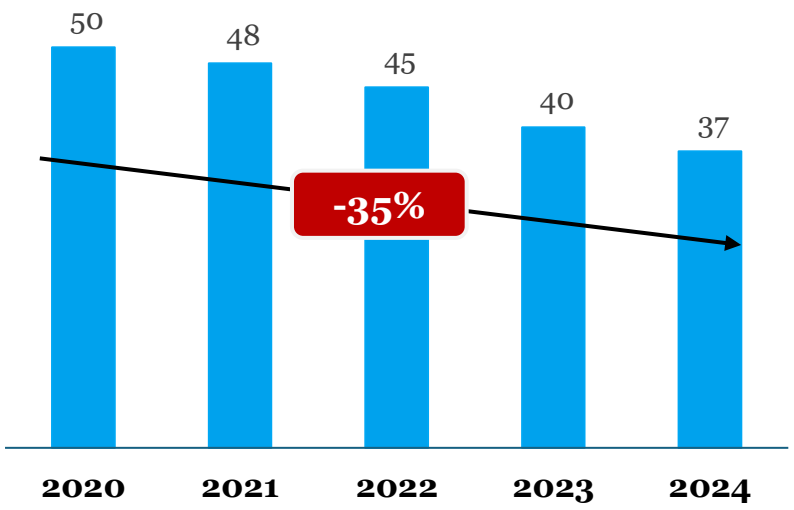
competitor

RideShare · Urbanxit · tridar · EXPRESS TAXI · RIDE24

Revenue declining sharp

Go-Hailing saw its revenue declining by **35%** over the past few years, significantly threatening its sustainability amid rising costs.

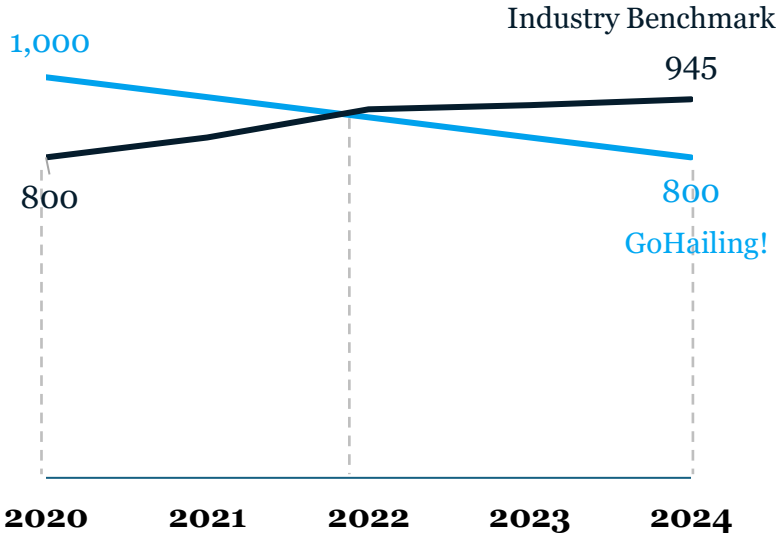
Go-Hailing revenue, in IDR Bn
2020 - 2024



Productivity decreases

GoHailing's business productivity has declined significantly compared to industry benchmarks, signaling increased customer and driver partner churn.

Productivity competition, in rides per hour
2020 - 2024



Note: The case study and data presented in this slide are hypothetical and created for illustrative purposes only

What kind of data that can help GoHailing in navigating those risks?

As competition intensifies, collecting insights from customers and drivers is crucial for developing an effective contingency plan

1 Customer insights

- How many customers that are rarely using the app? How can we stimulate them to use more our app?
- What is **the ratio of cancelled rides** compared to completed rides?
- How much is the **potential revenue loss** due to canceled rides?

2 Driver insights

- How many **drivers that are currently registered** within the database?
- What is the **ratio of idle driver** compared to the active ones in generating revenue?
- How can we stimulate driver to be more active and gain more revenue?

Contingency Plan

With all these data available, GoHailing! can formulate contingency plan to mitigate the risks, it could be through:

Driver incentives program

Customer promo program

Currently these data are **unavailable.** GoHailing! must develop **a robust and scalable data pipeline** to provide **actionable insights** for better **decision making** through **Data Engineering**

Note: The case study and data presented in this slide are hypothetical and created for illustrative purposes only

Three project deliverables for Data Engineer were set to optimize building data pipeline

Design pipeline architecture

Designing pipeline that's robust and scalable to meet business needs:

- **Technology stack selection**: script language, data warehouse, transactional database, orchestration tools, etc
- **Data Processing**: ELT or ETL, etc.
- **Dimensional data model**: Star Schema, Snowflake, etc
- **Data ingestion method:** Batch, stream, etc.

Build automation logic flow

Designing automation logical flow within the data pipeline architecture, for example:

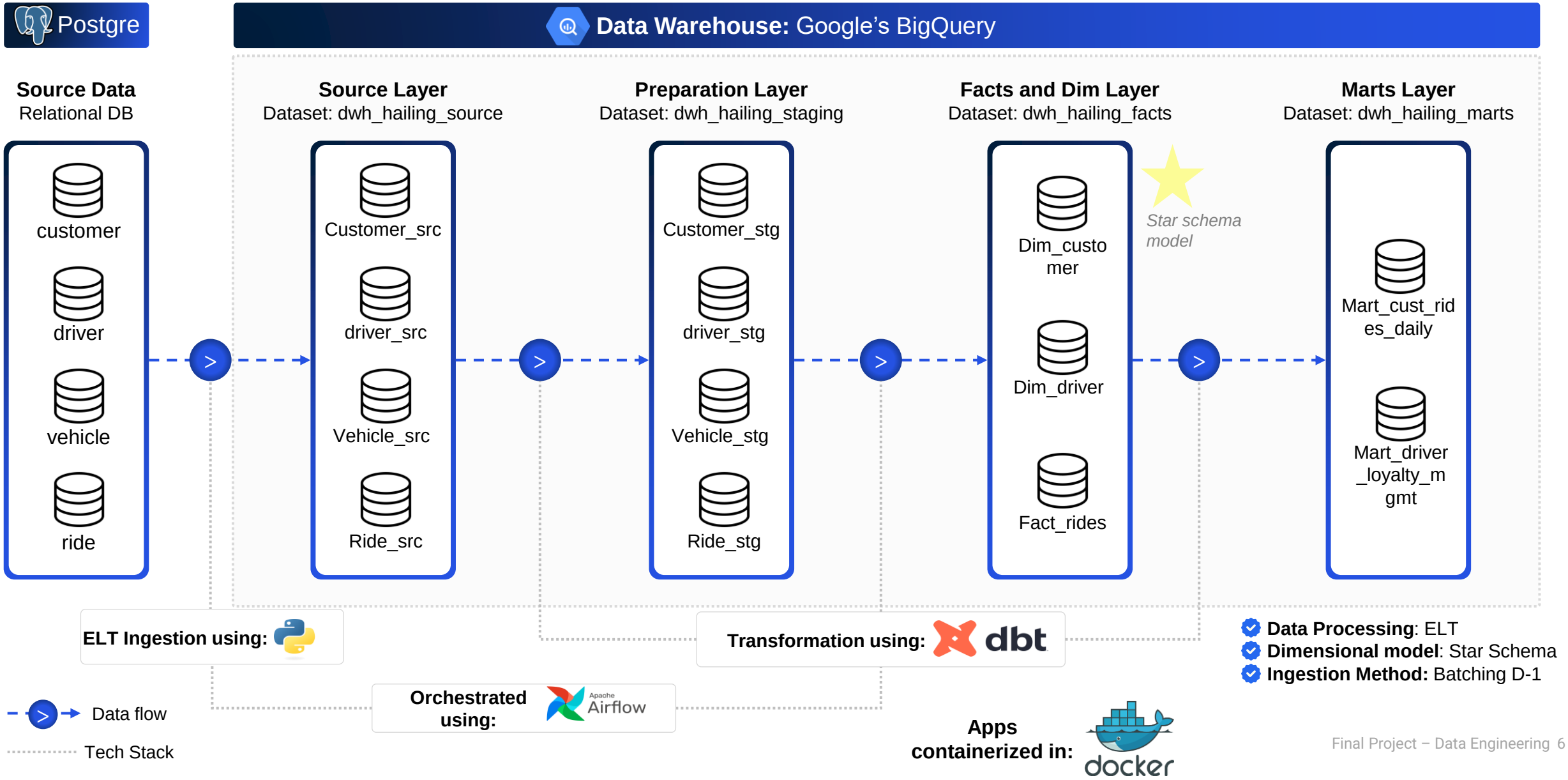
- What happens if error happens when ingesting the data?
- What needs to be done if there's any datatype mismatch between database and data warehouse?
- What would happen if the required table has not existed in data warehouse?

Result effectiveness

As a data engineer, it's essential to:

- Ensure the data engineering design is well implemented
- Ensure the data generated by the pipeline is **accessible and usable** for analyst or relevant stakeholders
- **Enabling data-driven decision making** by providing actionable insights through well-structured and reliable data
- Ensure **data privacy** is secure

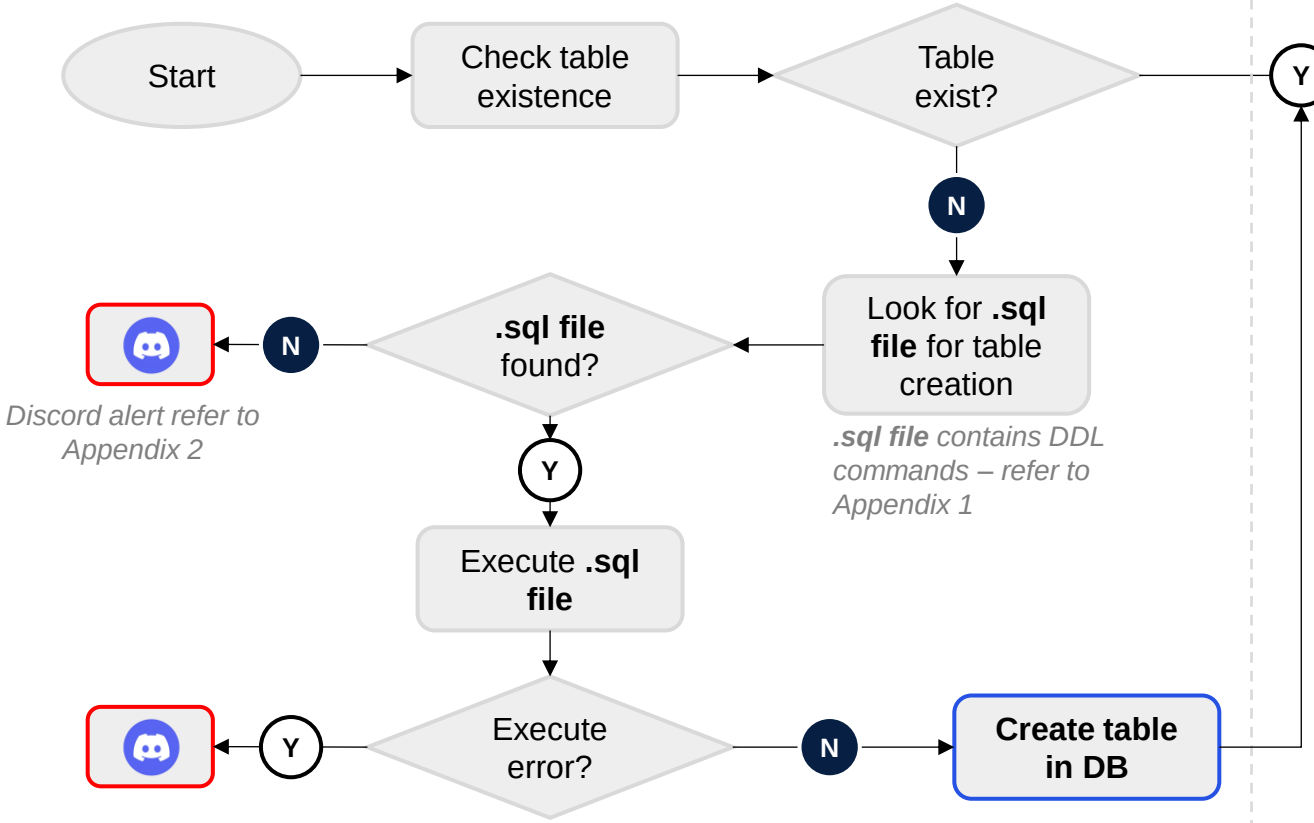
1st project deliverable: Design Pipeline Architecture



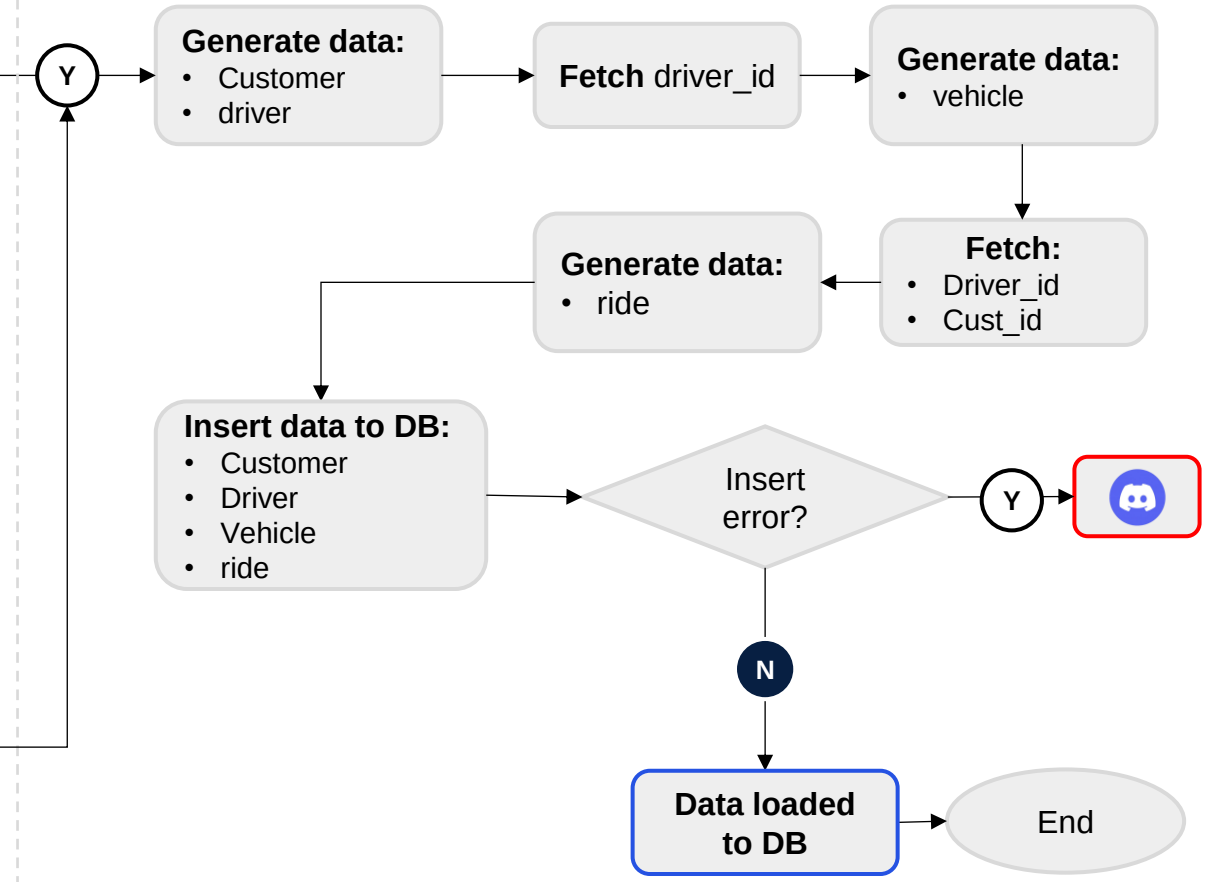
2nd project deliverable: Automation logic flow

DAG_NAME: DAG1_load_to_postgre.py | **OBJECTIVE:** Generate data and load to PostgreSQL relational database

Task 1: Ensure table existence in PostgreSQL



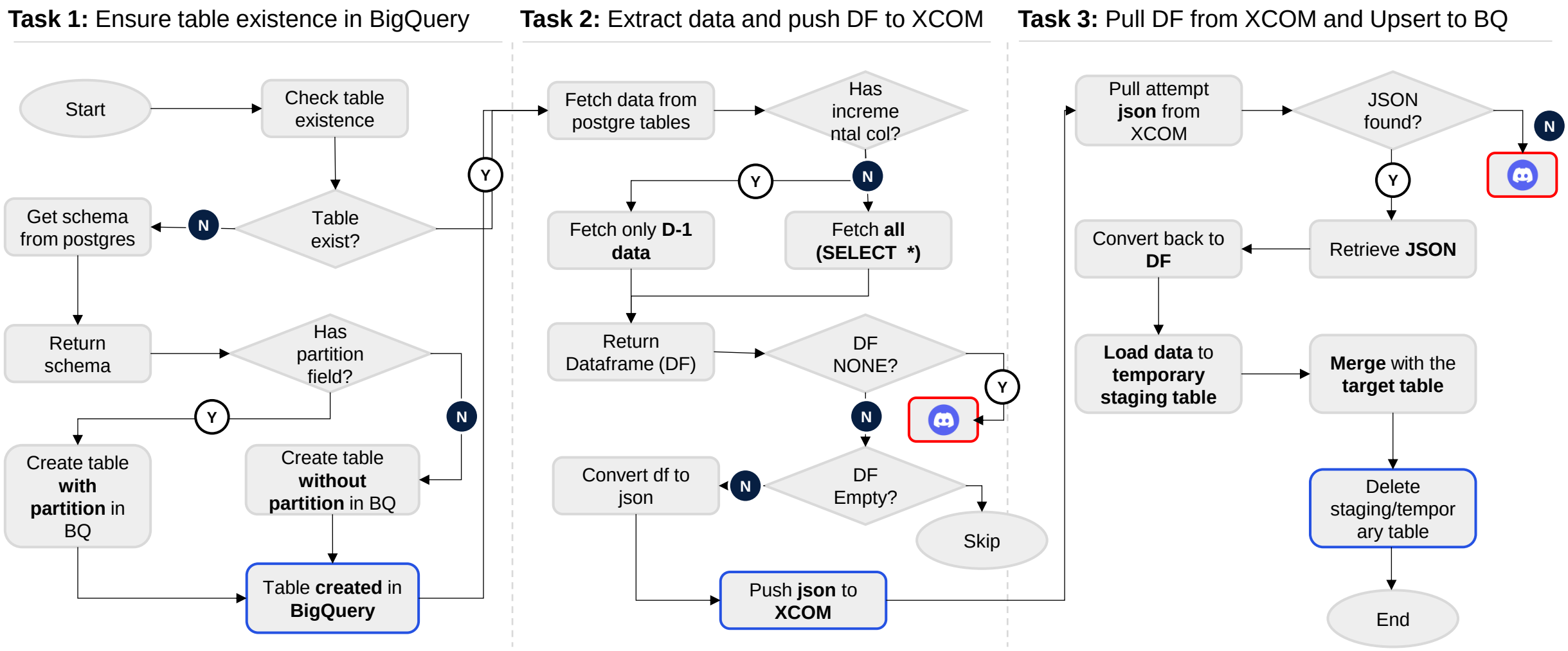
Task 2: Generate data and load to PostgreSQL/DB



Y Yes N No Send error/retry alert to discord Last process in a task

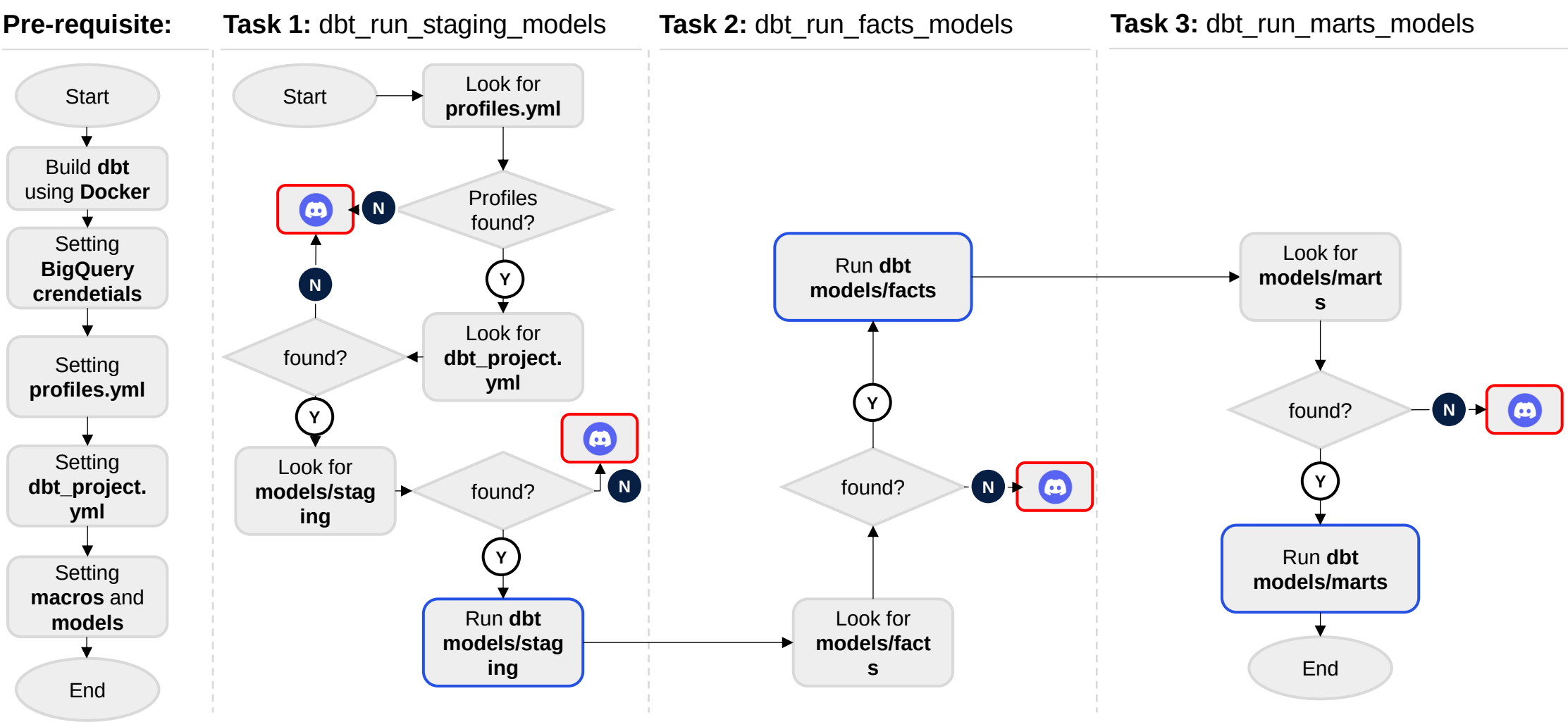
2nd project deliverable: Automation logic flow

DAG_NAME: DAG2_load_to_bigquery.py **| OBJECTIVE:** FETCH data from PostgreSQL and UPSERT to BigQuery



2nd project deliverable: Automation logic flow

DAG_NAME: DAG3_dbt_run_dwh.py | **OBJECTIVE:** Data transformation for raw data to business-level data in DWH



How well did the data pipeline design perform after implementation?

It performs well, ensuring seamless data ingestion into the warehouse and transforming raw data into business-ready insights

Data Generation

- Successfully **generate and load data** into PostgreSQL, as well as **building relationships between tables –** all orchestrated using **Airflow**
- Send discord alert if either error or retry in logic automation *(refer to appendix-2)*

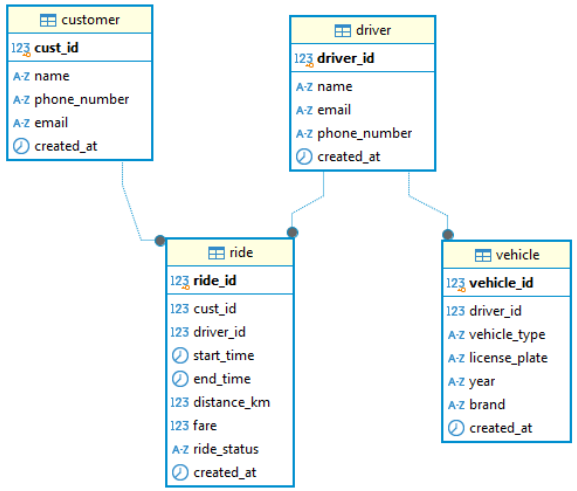


Fig 1. ERD in PostgreSQL

Data Warehouse Modelling

- Successfully executed **ELT process**, enabling **incremental data** ingestion from PostgreSQL to BigQuery using **upsert method**.
- **Transformed data** across layers from source to mart, ensuring business-ready insights. Implemented a **star schema** in facts layer to optimize analytical performance

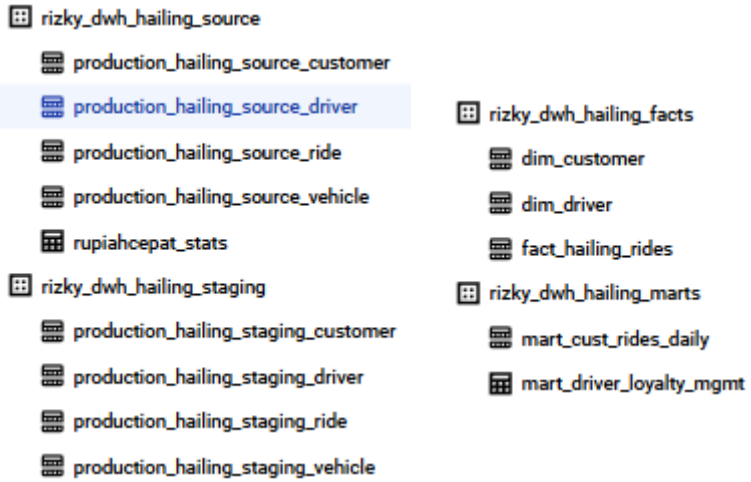


Fig 2. Dataset in BigQuery

Fulfillment of business needs

- The data can answer the questions that are asked in project's background, such as:

50% *of rides were cancelled.. need further investigations..*

438.2$ *in potential revenue was lost due to cancellations*

150 *drivers were registered within the database..*

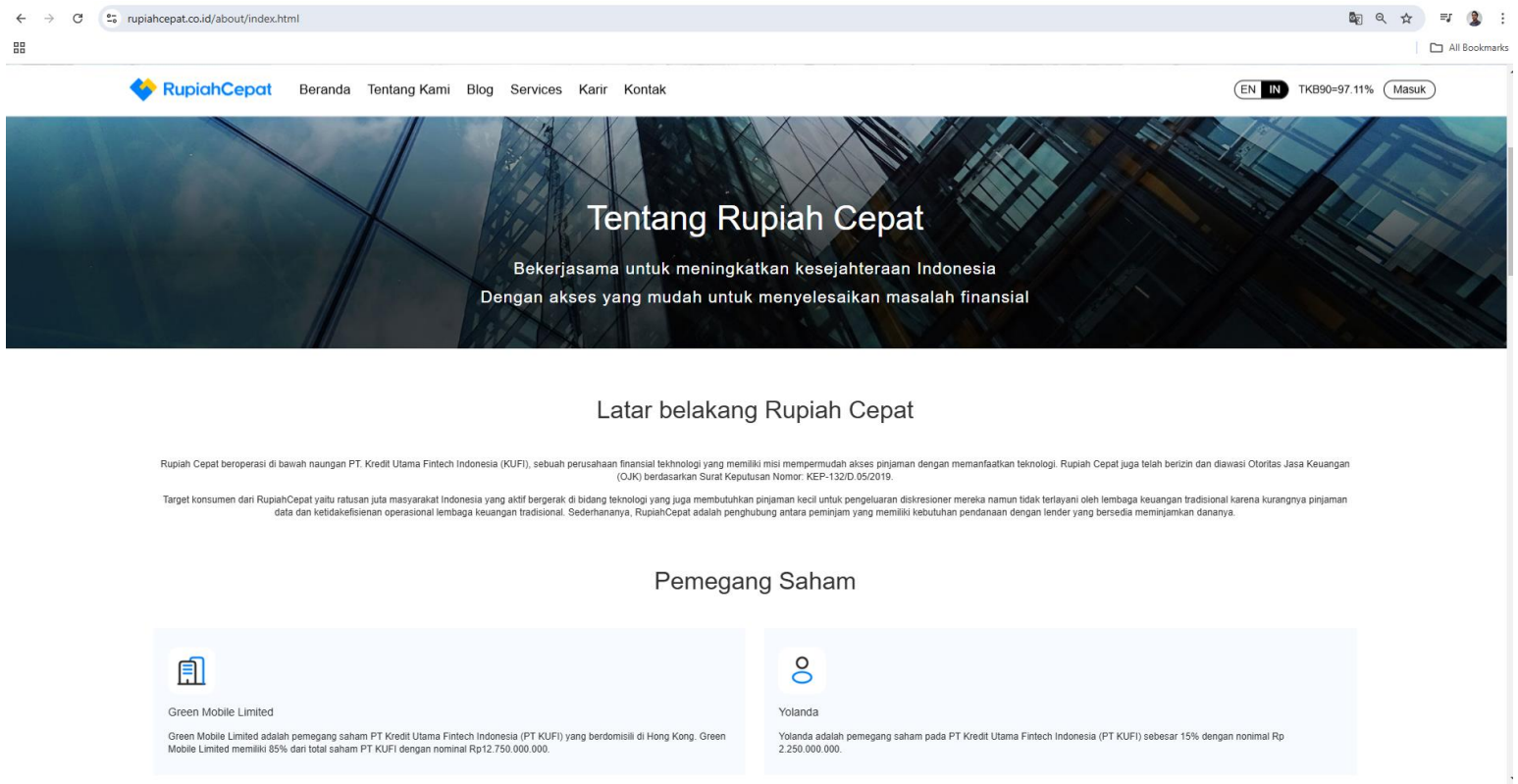
15% *of drivers are not generating any revenue.. Need to stimulate for retention...*

**Data as of 12th February 2025*

Final Project - Web Scraping Using Selenium

Github: https://github.com/rizkyjarr/final-project-purwadhika

Study Case: Web Scraping dynamic website of Rupiah Cepat



Project Objective: Perform web scraping from one of P2P lending website (RupiahCepat) and load to Data Warehouse (BigQuery)

Website scraping deliverable: Design Pipeline Architecture

Data flow ..... Tech Stack

Data Source: Rupiah Cepat's website statistic

Jumlah penerima dana sejak awal berdiri	Jumlah pemberi dana sejak awal berdiri	Jumlah pendanaan yang tersalurkan sejak awal berdiri
6.4 jt	1922	29.6 T
Jumlah penerima dana tahun berjalan	**Jumlah pemberi dana tahun berjalan**	**Jumlah pendanaan yang tersalurkan tahun berjalan**
390.9 ribu	57	886.5 M
Jumlah penerima dana di posisi akhir	**Jumlah pemberi dana di posisi akhir**	**Jumlah pendanaan yang tersalurkan di posisi akhir**
1191.3 ribu	56	1495.6 M

Data Warehouse: Google's BigQuery

Facts and Dim Layer
Dataset: rizky_dwh_hailing_source

Rupiahcep at_stats

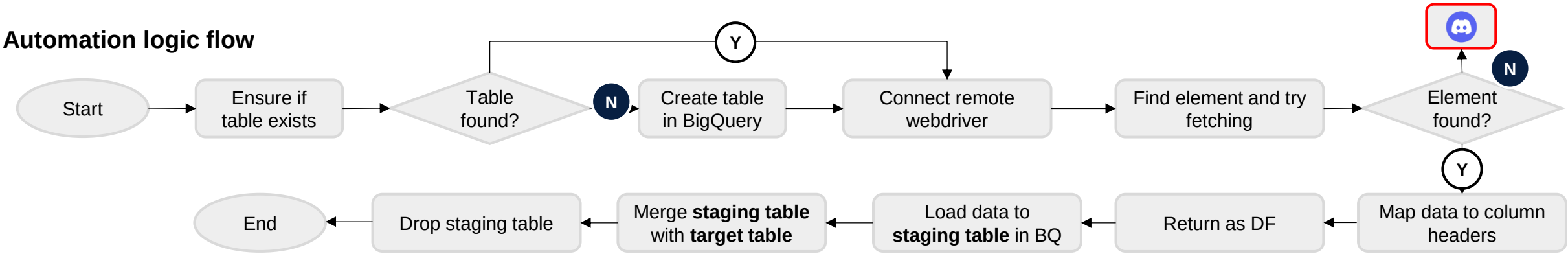
Scraped using: Selenium

Orchestrated using: Apache Airflow

Remarks:

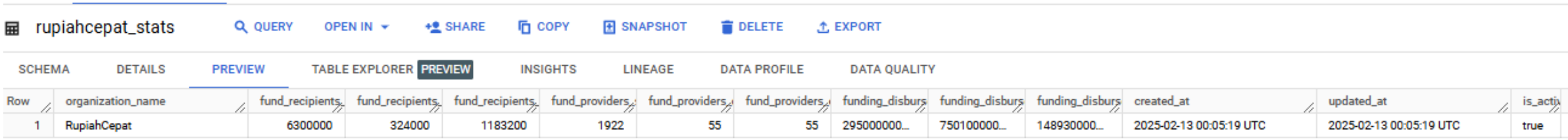
- **Data Processing**: ETL
- Insert new row if there's any changes from the website

Automation logic flow



How well did the data pipeline design perform after implementation?

- Data successfully **scraped** using **selenium**, and **upsert** to data warehouse – all orchestrated using **Airflow**
- If there's any changes in website, it will insert a new row of data and set the old data **inactive** in **is_active column**



Row	organization_name	fund_recipients	fund_recipients	fund_recipients	fund_providers	fund_providers	fund_providers	funding_disburs	funding_disburs	funding_disburs	created_at	updated_at	is_activ
1	RupiahCepat	6300000	324000	1183200	1922	55	55	295000000...	750100000...	148930000...	2025-02-13 00:05:19 UTC	2025-02-13 00:05:19 UTC	true

Fig 1. Data Preview in BigQuery

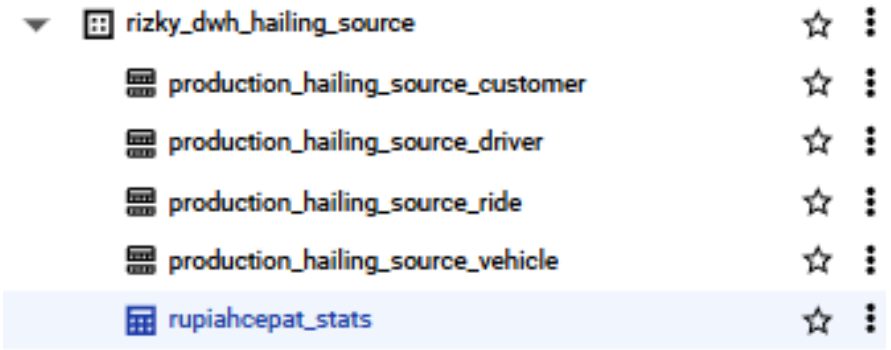


Fig 2. Table preview in BigQuery dataset

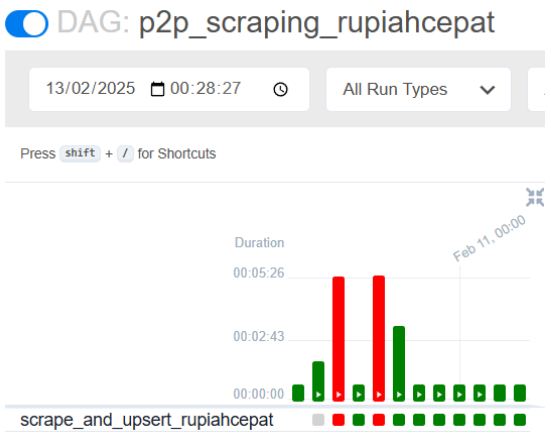


Fig 3. DAG4 Historical Task Logs

**Data as of 12th February 2025*

Thank you!

Any feedback would be very much appreciated ☺

Github: https://github.com/rizkyjarr/final-project-purwadhika

Appendix 1 – SQL FILES

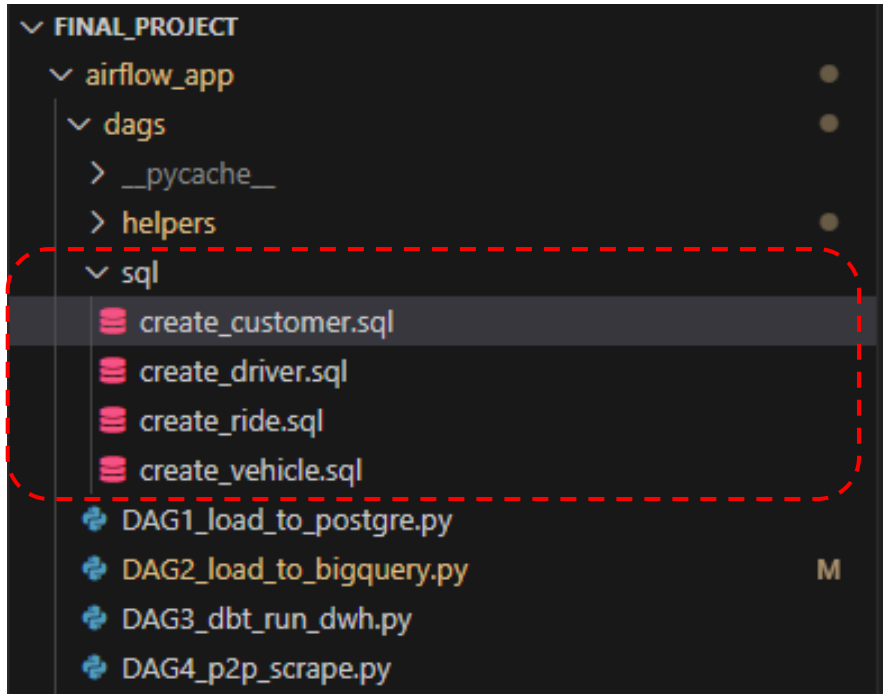


Fig 1. SQL folder location

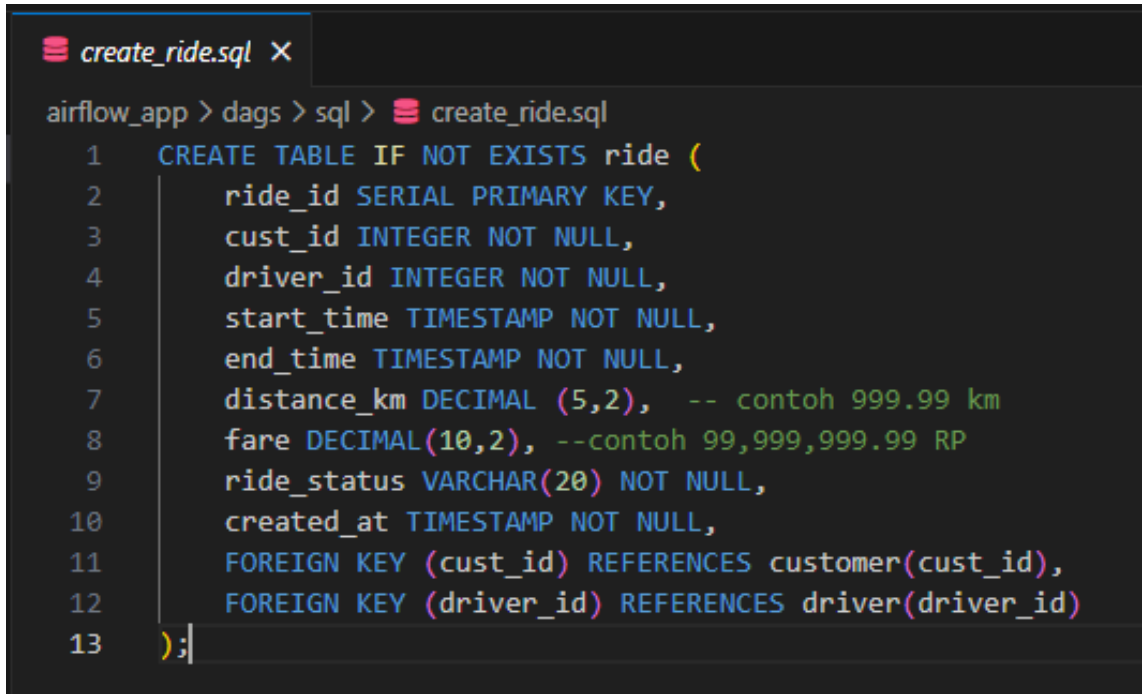


Fig 2. Code example

Appendix 2 – Discord Webhook Alert

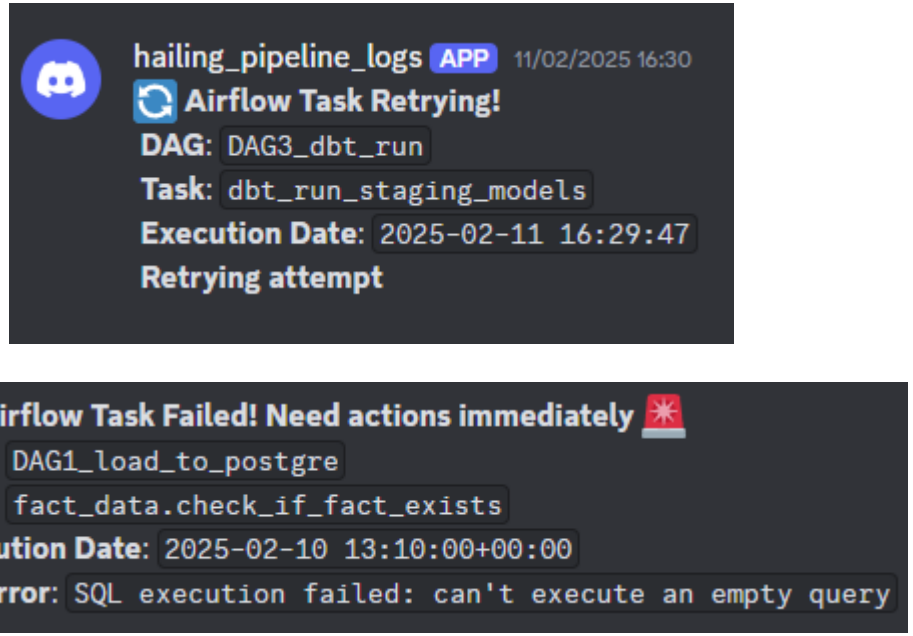


Fig 1. Error and retry alert on Discord

```python
DISCORD_WEBHOOK_URL = "https://discord.com/api/webhooks/1338491764238979132/gxkcO89HefdilVEvU0JeNmo_Mqsyg3Rg-V_S4Rs5k1yzysSA7_el9XvcNtsS_6fUXwAU"

def send_discord_alert(context, alert_type="failure"):

    dag_id = context.get('dag_run').dag_id
    task_id = context.get('task_instance').task_id
    execution_date = context.get('execution_date')
    exception = context.get('exception')

    jakarta_tz = pytz.timezone("Asia/Jakarta")
    execution_date_jakarta = execution_date.astimezone(jakarta_tz)
    execution_date_str = execution_date_jakarta.strftime("%Y-%m-%d %H:%M:%S")

    if alert_type == "failure":
        message = f"🚨 **Airflow Task Failed! Need actions immediately** \n" \
                  f" **DAG**: `{dag_id}`\n" \
                  f" **Task**: `{task_id}`\n" \
                  f" **Execution Date**: `{execution_date_str}`\n" \
                  f" **Error**: `{exception}`"
    elif alert_type == "retry":
        message = f"🔄 **Airflow Task Retrying!** \n" \
                  f" **DAG**: `{dag_id}`\n" \
                  f" **Task**: `{task_id}`\n" \
                  f" **Execution Date**: `{execution_date_str}`\n" \
                  f" **Retrying attempt**"
    elif alert_type == "success":
        message = f"✅ **Airflow Task Completed!** \n" \
                  f" **DAG**: `{dag_id}`\n" \
                  f" **Task**: `{task_id}`\n" \
                  f" **Execution Date**: `{execution_date_str}`\n" \
                  f" **Success**"

    payload = {"content": message}
    headers = {"Content-Type": "application/json"}

    response = requests.post(DISCORD_WEBHOOK_URL, data=json.dumps(payload), headers=headers)

    if response.status_code != 204:
        print(f"Failed to send Discord alert: {response.text}")
```

Fig 2. Discord alert's script

Appendix 3 – Data Preview in BigQuery

mart_cust_rides_daily

QUERY OPEN IN SHARE COPY SNAPSHOT DELETE EXPORT

This is a partitioned table. Learn more

SCHEMA DETAILS PREVIEW TABLE EXPLORER PREVIEW INSIGHTS LINEAGE DATA PROFILE DATA QUALITY

Row	cust_id	ride_date	cust_name	cust_email	cust_phone_number	no_of_rides	total_ride_durati
1	13	2025-02-03	Carla Schultz	carla.schultz@gmail.com	6825803430	1	16
2	16	2025-02-03	Maria Charles	maria.charles@hotmail.com	2954253781	1	9
3	2	2025-02-03	Justin Davis	justin.davis@yahoo.com	2507697049	1	28
4	5	2025-02-03	Michael Taylor	michael.taylor@yahoo.com	3893217677	2	46
5	11	2025-02-03	Steven Norton	steven.norton@gmail.com	4627285681	3	19
6	4	2025-02-03	Joshua Hull	joshua.hull@yahoo.com	2383802133	4	32
7	3	2025-02-03	Tina Cardenas	tina.cardenas@hotmail.com	5592886928	4	0
8	1	2025-02-03	Brenda Berry	brenda.berry@gmail.com	9694194121	5	14
9	54	2025-02-08	William Moore	william.moore@yahoo.com	7236132326	1	0
10	50	2025-02-08	Jessica Chambers	jessica.chambers@hotmail.com	9912912811	1	0
11	4	2025-02-08	Joshua Hull	joshua.hull@yahoo.com	2383802133	1	0
12	26	2025-02-08	Taylor Cooper	taylor.cooper@gmail.com	8766756323	1	0
13	22	2025-02-08	Sandra Hart	sandra.hart@hotmail.com	9622726664	1	0
14	52	2025-02-08	Kellie Shepherd	kellie.shepherd@hotmail.com	5386118723	1	0
15	74	2025-02-08	Antonio Reed	antonio.reed@hotmail.com	2637464529	1	8
16	65	2025-02-08	Sarah Andrade	sarah.andrade@hotmail.com	3753433576	1	8
17	29	2025-02-08	Patricia Hopkins	patricia.hopkins@yahoo.com	5389841129	1	12
18	72	2025-02-08	Kathleen Barnes	kathleen.barnes@hotmail.com	6995993317	1	15
19	42	2025-02-08	Carlos Watkins	carlos.watkins@hotmail.com	2439495985	1	20
20	1	2025-02-08	Brenda Berry	brenda.berry@gmail.com	9694194121	1	21

Fig 1. Mart_cust_rides_daily table

fact_hailing_rides

QUERY OPEN IN SHARE COPY SNAPSHOT DELETE EXPORT

This is a partitioned table. Learn more

SCHEMA DETAILS PREVIEW TABLE EXPLORER PREVIEW INSIGHTS LINEAGE DATA PROFILE DATA QUALITY

Row	ride_id	ride_start_time	ride_end_time	distance_km	fare	ride_status	cust_id
1	59	2025-02-06 20:42:35 UTC	2025-02-06 20:42:35 UTC	9.64	0	Cancelled	25
2	69	2025-02-06 20:44:49 UTC	2025-02-06 20:44:49 UTC	6.37	0	Cancelled	56
3	67	2025-02-06 20:15:28 UTC	2025-02-06 20:44:28 UTC	9.11	9.11	Completed	22
4	58	2025-02-06 20:17:01 UTC	2025-02-06 20:42:01 UTC	3.82	3.82	Completed	53
5	54	2025-02-06 20:28:05 UTC	2025-02-06 20:41:05 UTC	9.22	9.22	Completed	18
6	53	2025-02-06 20:40:53 UTC	2025-02-06 20:40:53 UTC	8.86	0	Cancelled	39
7	63	2025-02-06 20:43:25 UTC	2025-02-06 20:43:25 UTC	6.45	0	Cancelled	56
8	65	2025-02-06 20:34:59 UTC	2025-02-06 20:43:59 UTC	2.57	2.57	Completed	6
9	62	2025-02-06 20:43:11 UTC	2025-02-06 20:43:11 UTC	6.5	0	Cancelled	34
10	68	2025-02-06 20:44:44 UTC	2025-02-06 20:44:44 UTC	2.76	0	Cancelled	30
11	55	2025-02-06 20:11:18 UTC	2025-02-06 20:41:18 UTC	3.61	3.61	Completed	33
12	51	2025-02-06 20:40:37 UTC	2025-02-06 20:40:37 UTC	5.64	0	Cancelled	4
13	52	2025-02-06 20:40:37 UTC	2025-02-06 20:40:37 UTC	8.02	0	Cancelled	40
14	56	2025-02-06 20:12:29 UTC	2025-02-06 20:41:29 UTC	1.38	1.38	Completed	38
15	61	2025-02-06 20:42:58 UTC	2025-02-06 20:42:58 UTC	2.04	0	Cancelled	22
16	66	2025-02-06 20:44:14 UTC	2025-02-06 20:44:14 UTC	5.41	0	Cancelled	22
17	64	2025-02-06 20:43:38 UTC	2025-02-06 20:43:38 UTC	7.46	0	Cancelled	44
18	60	2025-02-06 20:42:47 UTC	2025-02-06 20:42:47 UTC	9.18	0	Cancelled	27
19	70	2025-02-06 20:44:55 UTC	2025-02-06 20:44:55 UTC	4.71	0	Cancelled	1
20	57	2025-02-06 20:33:40 UTC	2025-02-06 20:41:40 UTC	8.89	8.89	Completed	45

Fig 2. Fact_hailing_rides

mart_driver_loyalty_mgmt

QUERY OPEN IN SHARE COPY SNAPSHOT DELETE EXPORT

SCHEMA DETAILS PREVIEW TABLE EXPLORER PREVIEW INSIGHTS LINEAGE DATA PROFILE DATA QUALITY

Row	driver_id	driver_name	phone_number	email	vehicle_type	no_of_rides	total_ride_durati
1	135	Renee Keller	3093785575	renee.keller@hotmail.com	Car	0	null
2	143	Regina Blake	8899002851	regina.blake@yahoo.com	Car	0	null
3	44	Corey Ray	7206503831	corey.ray@gmail.com	Car	0	null
4	128	Adam Foster	3147849778	adam.foster@hotmail.com	Car	0	null
5	75	William Sanchez	8179030128	william.sanchez@hotmail.com	Car	0	null
6	89	Michael Shah	3227248410	michael.shah@yahoo.com	Car	0	null
7	30	Kathryn Jones	5043728416	kathryn.jones@yahoo.com	Car	0	null
8	81	Derrick Goodman	2088221624	derrick.goodman@gmail.com	Car	0	null
9	103	Diane Santos	9503134848	diane.santos@gmail.com	Car	0	null
10	118	Joanne Smith	7024801748	joanne.smith@gmail.com	Car	0	null
11	130	Ashlee Brown	2366658453	ashlee.brown@hotmail.com	Car	0	null
12	138	Rebecca Garrison	2799940471	rebecca.garrison@gmail.com	Car	0	null
13	41	Robin Arias	9454398541	robin.arias@gmail.com	Car	0	null
14	120	Todd Little	4803110726	todd.little@yahoo.com	Car	0	null
15	133	Lance Smith	7226209865	lance.smith@gmail.com	Car	0	null
16	78	Dylan Turner	8687589839	dylan.turner@yahoo.com	Car	0	null
17	59	Michele Jones	8225626336	michele.jones@yahoo.com	Car	0	null
18	98	Patrick Lawson	4416126026	patrick.lawson@yahoo.com	Car	0	null
19	20	Oscar Jordan	4077955521	oscar.jordan@gmail.com	Car	0	null
20	77	Jacob Cole	4085527454	jacob.cole@yahoo.com	Car	0	null

Fig 3. Mart_driver_loyalty_mgmt table

dim_customer

QUERY OPEN IN SHARE COPY SNAPSHOT DELETE EXPORT

This is a partitioned table. Learn more

SCHEMA DETAILS PREVIEW TABLE EXPLORER PREVIEW INSIGHTS LINEAGE DATA PROFILE DATA QUALIT

Row	cust_id	name	phone_number	email	created_at
1	52	Kellie Shepherd	5386118723	kellie.shepherd@hotmail.com	2025-02-06 20:40:32 UTC
2	51	Jennifer Nelson	8317242999	jennifer.nelson@hotmail.com	2025-02-06 20:40:32 UTC
3	53	Erica Wyatt	6835362631	erica.wyatt@yahoo.com	2025-02-06 20:40:48 UTC
4	54	William Moore	7236132326	william.moore@yahoo.com	2025-02-06 20:41:01 UTC
5	55	Chris Diaz	6337555613	chris.diaz@hotmail.com	2025-02-06 20:41:14 UTC
6	56	Richard Bean	3856953454	richard.bean@gmail.com	2025-02-06 20:41:25 UTC
7	57	Monica Perez	8503995003	monica.perez@hotmail.com	2025-02-06 20:41:36 UTC
8	58	Christine Williams	8045304140	christine.williams@gmail.com	2025-02-06 20:41:57 UTC
9	59	Laura Robinson	9884812646	laura.robinson@hotmail.com	2025-02-06 20:42:30 UTC
10	60	Margaret Davis	7307032406	margaret.davis@hotmail.com	2025-02-06 20:42:43 UTC
11	61	Gregory Edwards	9463292118	gregory.edwards@yahoo.com	2025-02-06 20:42:53 UTC
12	62	Amanda Murphy	6537326375	amanda.murphy@gmail.com	2025-02-06 20:43:06 UTC
13	63	Zachary Watts	4729363046	zachary.watts@hotmail.com	2025-02-06 20:43:20 UTC
14	64	Felicia White	4612171120	felicia.white@hotmail.com	2025-02-06 20:43:33 UTC
15	65	Sarah Andrade	3753433576	sarah.andrade@hotmail.com	2025-02-06 20:43:55 UTC
16	66	Paul Kelly	6699402493	paul.kelly@hotmail.com	2025-02-06 20:44:09 UTC
17	67	Maria Owens	9135886399	maria.owens@hotmail.com	2025-02-06 20:44:24 UTC
18	68	Sara Williams	3038947857	sara.williams@gmail.com	2025-02-06 20:44:39 UTC
19	69	Zachary Patterson	9256311067	zachary.patterson@hotmail.com	2025-02-06 20:44:44 UTC
20	70	Chad Morgan	2818580852	chad.morgan@hotmail.com	2025-02-06 20:44:50 UTC

Fig 4. dim_customer table