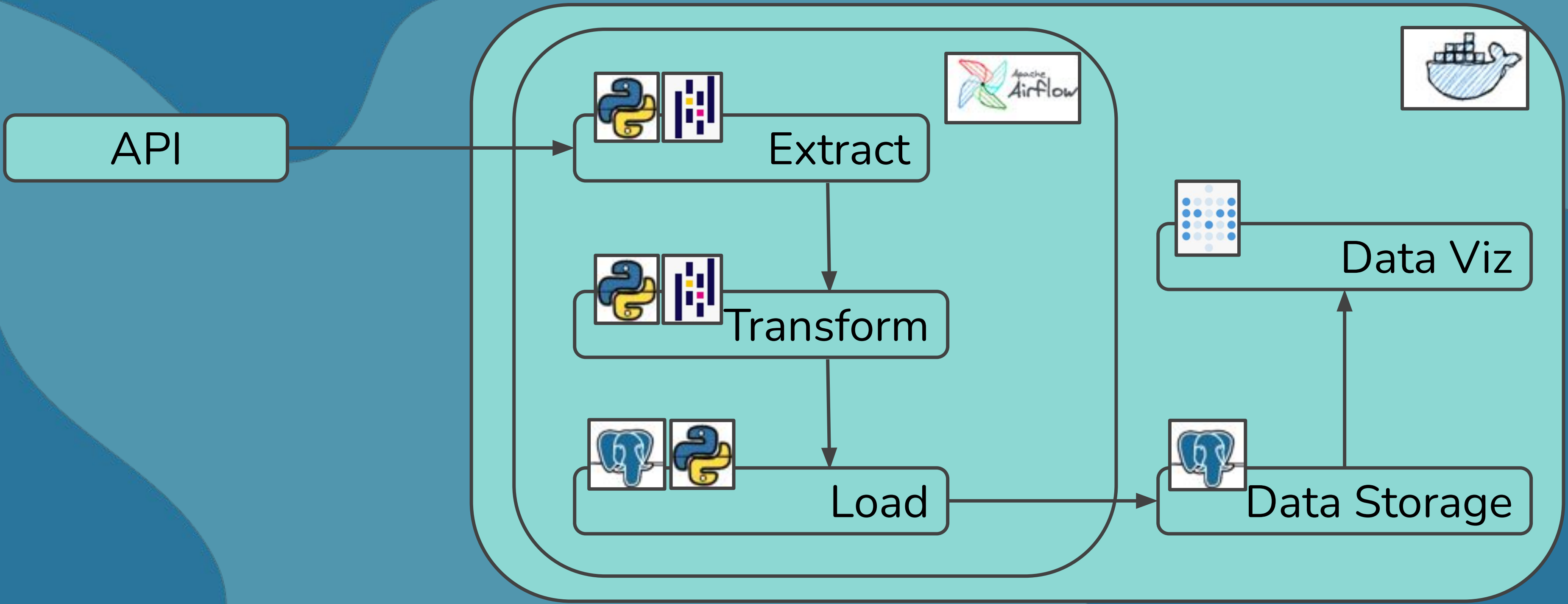
Final Project Dibimbing DE 3

Anime Data Pipeline

Ahmad Belva

Data Pipeline



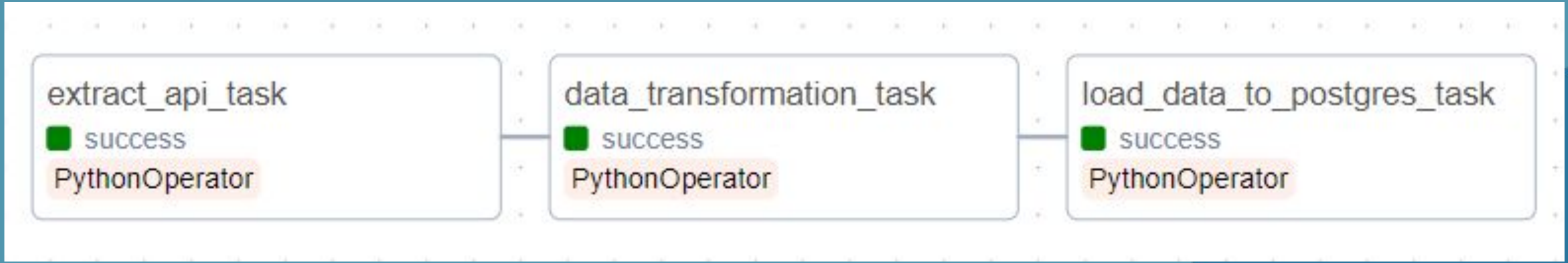
API

The API used in this project is Jikan API, which is an open-source API that provides information from myanimelist.net. Information about the API can be accessed here: https://jikan.moe/

For this project, the following endpoint is used:
https://api.jikan.moe/v4/seasons/{year}/{season}?page={i}

{year}: the year in which data is requested (e.g. 2020, 2021, ...)
{season}: the season in which data is requested (winter, spring, summer, fall)
{page}: the page number of the requested data (e.g. 1, 2, 3, ...)

DAG Architecture

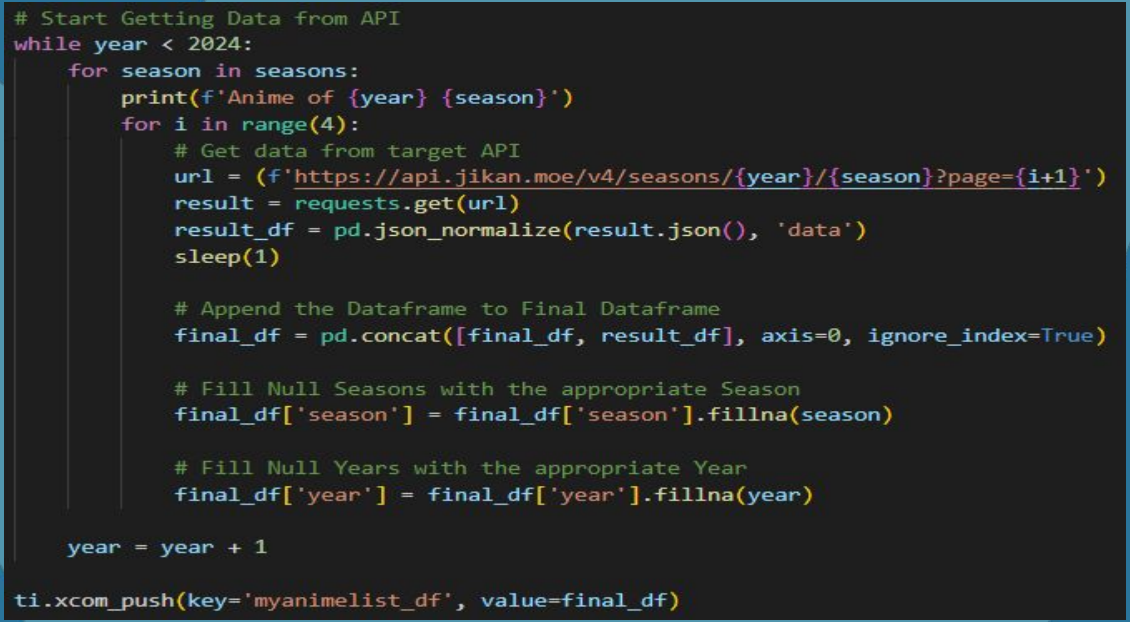


extract_api_task
🟩 success
PythonOperator

data_transformation_task
🟩 success
PythonOperator

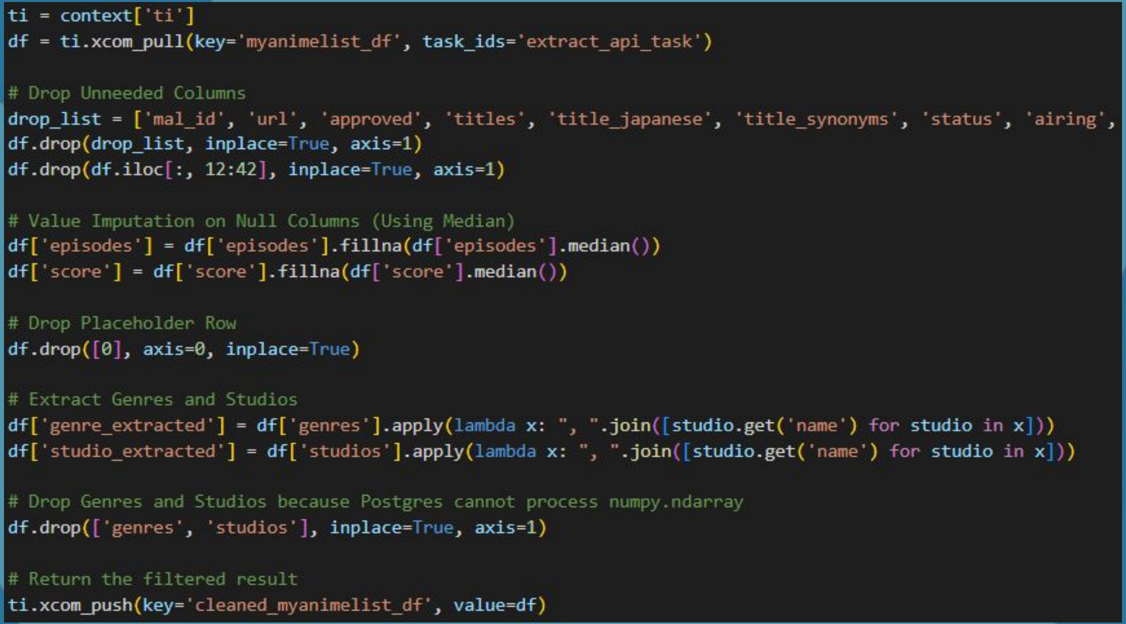
load_data_to_postgres_task
🟩 success
PythonOperator

Extract



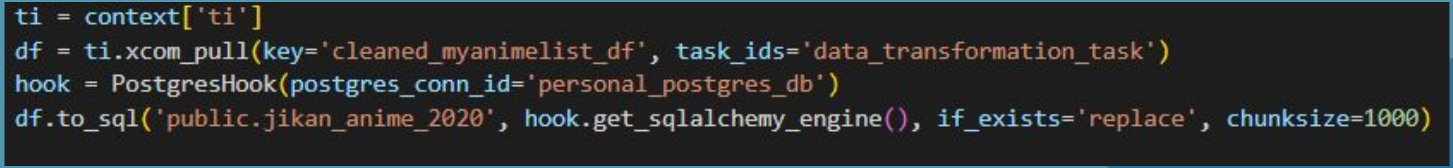
```
# Start Getting Data from API
while year < 2024:
    for season in seasons:
        print(f'Anime of {year} {season}')
        for i in range(4):
            # Get data from target API
            url = (f'https://api.jikan.moe/v4/seasons/{year}/{season}?page={i+1}')
            result = requests.get(url)
            result_df = pd.json_normalize(result.json(), 'data')
            sleep(1)

            # Append the Dataframe to Final Dataframe
            final_df = pd.concat([final_df, result_df], axis=0, ignore_index=True)

            # Fill Null Seasons with the appropriate Season
            final_df['season'] = final_df['season'].fillna(season)

            # Fill Null Years with the appropriate Year
            final_df['year'] = final_df['year'].fillna(year)

    year = year + 1

ti.xcom_push(key='myanimelist_df', value=final_df)
```

Transform



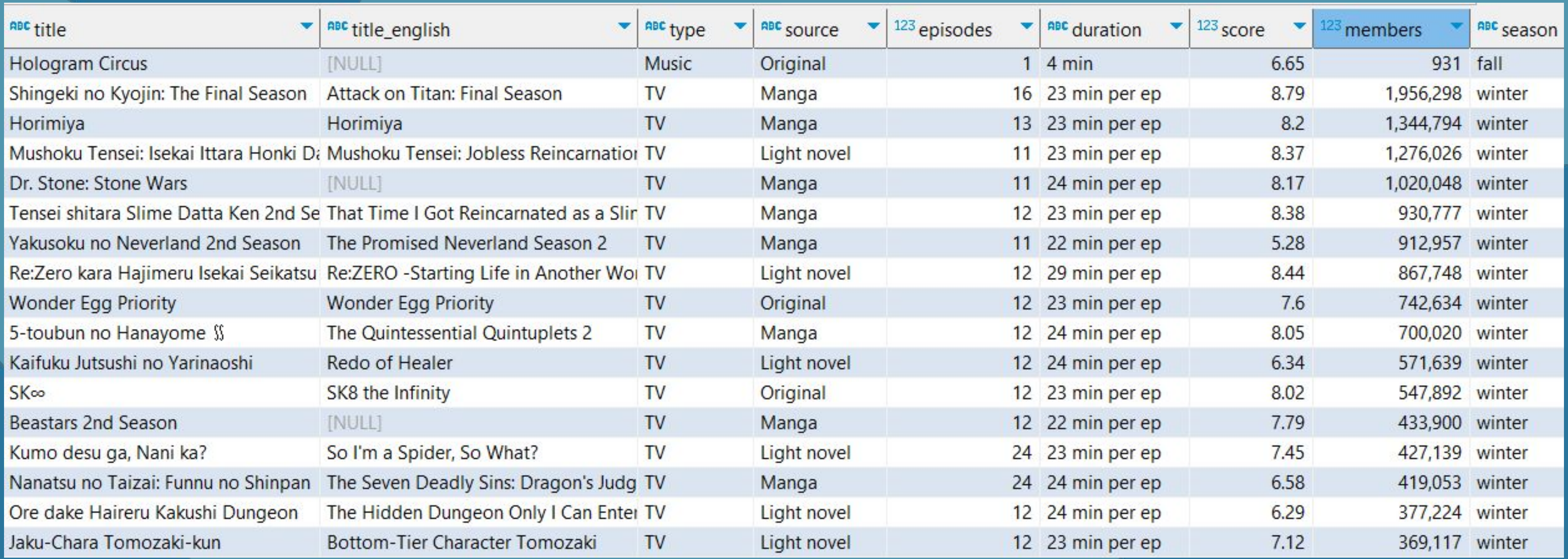
```
ti = context['ti']
df = ti.xcom_pull(key='myanimelist_df', task_ids='extract_api_task')

# Drop Unneeded Columns
drop_list = ['mal_id', 'url', 'approved', 'titles', 'title_japanese', 'title_synonyms', 'status', 'airing',
df.drop(drop_list, inplace=True, axis=1)
df.drop(df.iloc[:, 12:42], inplace=True, axis=1)

# Value Imputation on Null Columns (Using Median)
df['episodes'] = df['episodes'].fillna(df['episodes'].median())
df['score'] = df['score'].fillna(df['score'].median())

# Drop Placeholder Row
df.drop([0], axis=0, inplace=True)

# Extract Genres and Studios
df['genre_extracted'] = df['genres'].apply(lambda x: ", ".join([studio.get('name') for studio in x]))
df['studio_extracted'] = df['studios'].apply(lambda x: ", ".join([studio.get('name') for studio in x]))

# Drop Genres and Studios because Postgres cannot process numpy.ndarray
df.drop(['genres', 'studios'], inplace=True, axis=1)

# Return the filtered result
ti.xcom_push(key='cleaned_myanimelist_df', value=df)
```

Load



```
ti = context['ti']
df = ti.xcom_pull(key='cleaned_myanimelist_df', task_ids='data_transformation_task')
hook = PostgresHook(postgres_conn_id='personal_postgres_db')
df.to_sql('public.jikan_anime_2020', hook.get_sqlalchemy_engine(), if_exists='replace', chunksize=1000)
```

Postgres Table Result



title	title_english	type	source	episodes	duration	score	members	season
Hologram Circus	[NULL]	Music	Original	1	4 min	6.65	931	fall
Shingeki no Kyojin: The Final Season	Attack on Titan: Final Season	TV	Manga	16	23 min per ep	8.79	1,956,298	winter
Horimiya	Horimiya	TV	Manga	13	23 min per ep	8.2	1,344,794	winter
Mushoku Tensei: Isekai Ittara Honki Da	Mushoku Tensei: Jobless Reincarnatio	TV	Light novel	11	23 min per ep	8.37	1,276,026	winter
Dr. Stone: Stone Wars	[NULL]	TV	Manga	11	24 min per ep	8.17	1,020,048	winter
Tensei shitara Slime Datta Ken 2nd Se	That Time I Got Reincarnated as a Slir	TV	Manga	12	23 min per ep	8.38	930,777	winter
Yakusoku no Neverland 2nd Season	The Promised Neverland Season 2	TV	Manga	11	22 min per ep	5.28	912,957	winter
Re:Zero kara Hajimeru Isekai Seikatsu	Re:ZERO -Starting Life in Another Wo	TV	Light novel	12	29 min per ep	8.44	867,748	winter
Wonder Egg Priority	Wonder Egg Priority	TV	Original	12	23 min per ep	7.6	742,634	winter
5-toubun no Hanayome ∬	The Quintessential Quintuplets 2	TV	Manga	12	24 min per ep	8.05	700,020	winter
Kaifuku Jutsushi no Yarinaoshi	Redo of Healer	TV	Light novel	12	24 min per ep	6.34	571,639	winter
SK∞	SK8 the Infinity	TV	Original	12	23 min per ep	8.02	547,892	winter
Beastars 2nd Season	[NULL]	TV	Manga	12	22 min per ep	7.79	433,900	winter
Kumo desu ga, Nani ka?	So I'm a Spider, So What?	TV	Light novel	24	23 min per ep	7.45	427,139	winter
Nanatsu no Taizai: Funnu no Shinpan	The Seven Deadly Sins: Dragon's Judg	TV	Manga	24	24 min per ep	6.58	419,053	winter
Ore dake Haireru Kakushi Dungeon	The Hidden Dungeon Only I Can Ente	TV	Light novel	12	24 min per ep	6.29	377,224	winter
Jaku-Chara Tomozaki-kun	Bottom-Tier Character Tomozaki	TV	Light novel	12	23 min per ep	7.12	369,117	winter

Analysis Dashboard

Dashboard result of the analysis done on the anime table (done using Metabase):

https://drive.google.com/file/d/1PnklUncTggf2UPMe7UXF3OGH5vp_HDLR/view

Thank You