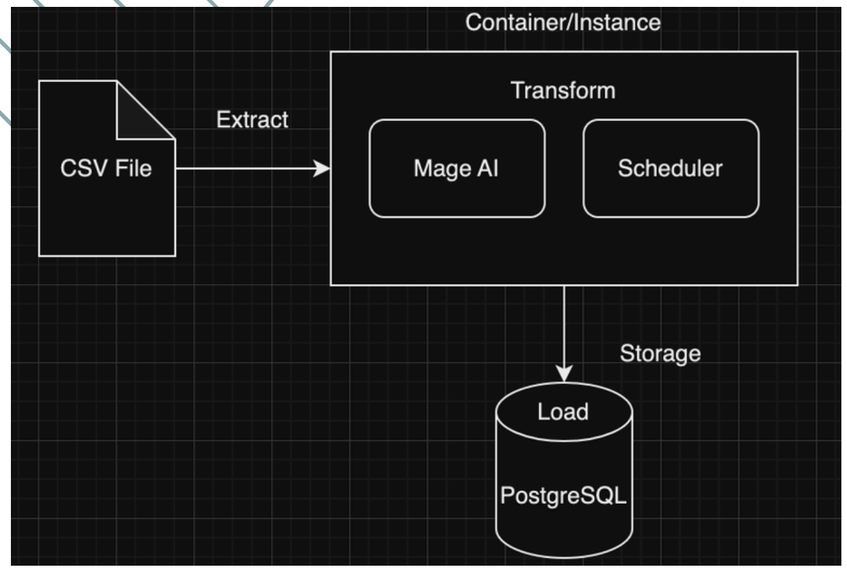
TECHNICAL TEST DATA ANALYST - MNC MEDIA

Taqiyudin Muhammad Khalil

TABLE OF CONTENT

ARCHITECTURE PLAN



Components:

1. Docker-Compose: Containerized Mage AI and PostgreSQL services.
2. Mage AI: Orchestration tool to manage ETL/ELT pipelines.
3. PostgreSQL: Data storage.
4. CSV File: Input data for extraction.

ETL PROCESS
EXTRACTION

PY DATA LOADER load_data_csv Edit parents

```python
import io
import pandas as pd
import requests


@data_loader
def load_data_from_api(*args, **kwargs):
    path = 'magic/data_loaders/dataset_user_behavior_for_test.csv'
    return pd.read_csv(path, sep=',')


@test
def test_row_count(df, *args) -> None:
    assert len(df.index) >= 1000, 'The data does not have enough rows.'
```

Process:

1. Run Docker Compose
2. Upload the provided CSV file to the first pipeline in Mage AI.
3. Use Mage AI's connectors to read the dataset.
4. Extract data from the CSV and prepare it for transformation.

Tools Used: Mage AI, Docker container for Mage AI service.

ETL PROCESS
TRANSFORM

PY TRANSFORMER transform_data_csv 1 parent

```python
def transform(df, *args, **kwargs):
    # Add number of meals for each user
    df_new_column = number_of_rows_per_key(df, 'Iduser', 'total_row')
    df = df.join(df_new_column, on='Iduser')

    # Clean column names
    df.columns = [clean_column(col) for col in df.columns]
    df = preprocess(df)
    df['start_watching'] = pd.to_datetime(df['start_watching'], format='%m/%d/%Y %

    return df


@test
def test_number_of_columns(df, *args) -> None:
    assert len(df.columns) < 11, 'There needs to be at least 11 columns.'
```

1/1 tests passed.

OUTPUT 0

iduser	start_watching	device_id	province	city

Process:

1. Apply transformations like data cleaning, normalization, or aggregation.
2. Handle missing values, duplicate data, and ensure consistency.
3. Convert date format, standardize columns.

Tools Used: Mage AI (Python).

Quality Checks:

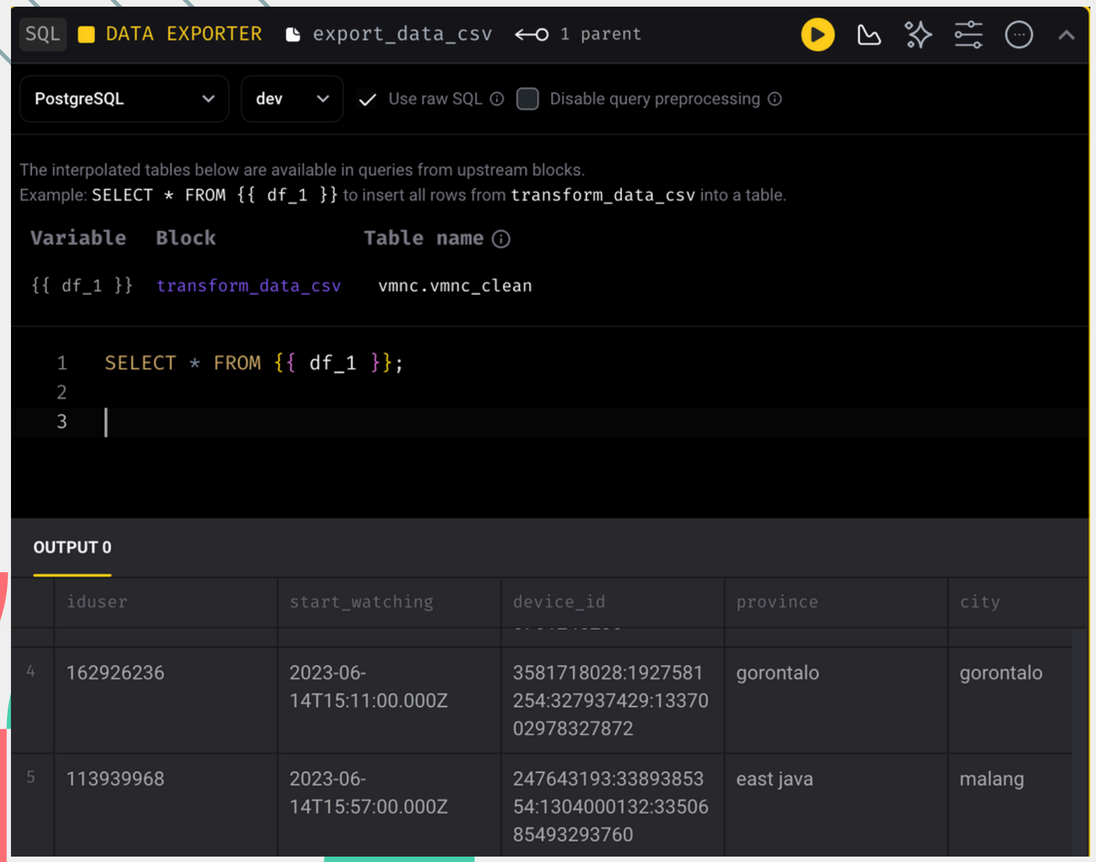
1. Missing Data: Ensure there are no missing or NULL values in critical columns.
2. Data Types: Validate that the data types are consistent (e.g., integers, dates).
3. Duplicates: Identify and remove any duplicate rows.

Validation Steps:

1. Verify data integrity before loading into PostgreSQL.
2. Validate data types and formats post-transformation.

ETL PROCESS

LOAD



Process:

1. Once the data is transformed, load it into PostgreSQL.
2. Use Mage AI's PostgreSQL connector to insert the cleaned data into the database.

Tools Used: PostgreSQL container (via Docker), Mage AI.

TABLE RESULT

SQL DATA LOADER load_data_contenttype Edit parents

PostgreSQL dev Use raw SQL Disable query preprocessing

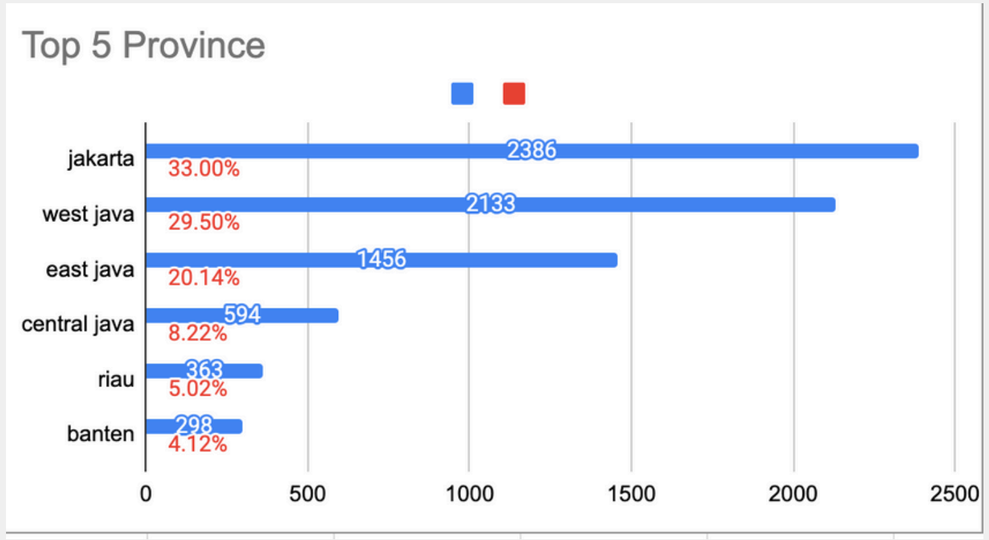
```sql
DROP TABLE IF EXISTS vmnc.vmnc_content_type;

CREATE TABLE vmnc.vmnc_content_type (
    province VARCHAR(255) PRIMARY KEY,
    total_user INT
);


INSERT INTO vmnc.vmnc_content_type
SELECT
    content_type,
    COUNT(DISTINCT iduser) AS total_user
FROM vmnc.vmnc_clean
GROUP BY 1
ORDER BY 2 DESC;
```

OUTPUT 0

province	total_user



SQL DATA LOADER load_data_province Edit parents

PostgreSQL dev Use raw SQL Disable query preprocessing

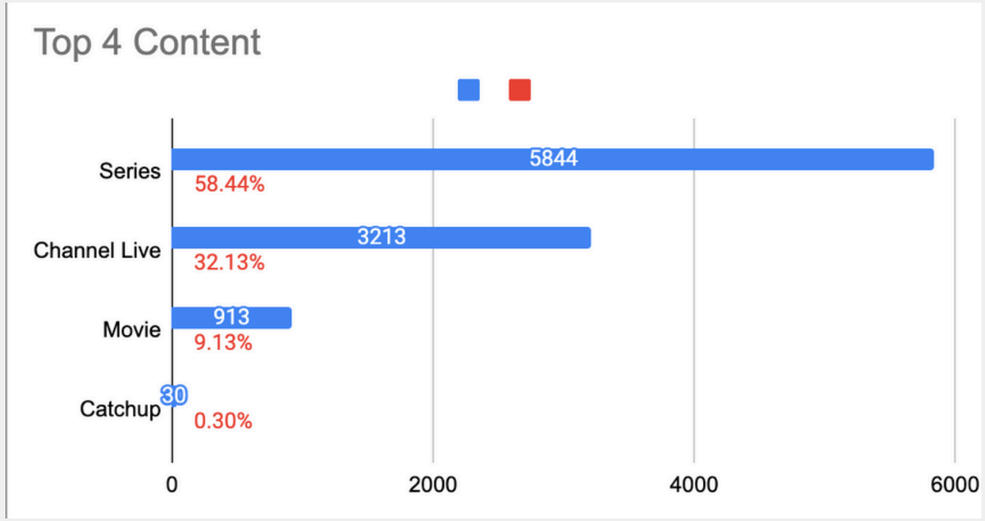
```sql
DROP TABLE IF EXISTS vmnc.vmnc_province;

CREATE TABLE vmnc.vmnc_province (
    province VARCHAR(255) PRIMARY KEY,
    total_user INT
);


INSERT INTO vmnc.vmnc_province
SELECT
    province,
    COUNT(DISTINCT iduser) AS total_user
FROM vmnc.vmnc_clean
GROUP BY 1
ORDER BY 2 DESC;
```

OUTPUT 0

	province	total_user
0	jakarta	2386



The top 5 provinces with the highest number of users are still dominated by Java Island, accounting for 94.98%, with Jakarta leading at 33%. Riau is the only province outside Java Island that ranks in the top 5. This presents a potential opportunity to increase users around that region.

The top content is still dominated by series, with 58.44%. Movie and catch-up content have user numbers below 10%. This needs to be evaluated, and innovations should be sought to boost users in these content categories.

THANK YOU