EXTRACT LOAD PIPELINE

IMDb Web Scraping to BigQuery

Beami MW

Padjadjaran University

Biology Major

Projects:

- Relationship of Leaf Characteristics of Bougainvillea Spectabilis Wild and Hibiscus Archeri Wats with Particulate and Pb (Lead) Accumulation at Cibeunying Park Area Bandung
- Study of Macroalgae Community at Pancur Beach West Zone of Alas Purwo National Park East Java

Purwadhika School

Data Science & Machine Learning

Projects:

- Crime in Boston Analysis
- E-Commerce Customer Churn Prediction
- Hotel Bookings Analysis and Cancellation Prediction

Job Experience

PT Eikon Technology
Data Analyst (April 2023-Aug 2024)

- Defined business problems for each case and identified best practice solutions for cost optimization, then implemented them.
- Analyzed data using SQL, defined and calculated metrics, created data visualization dashboard reports, and presented them.
- Transformed, cleaned, and structured data to make it more usable and useful for analytics or machine learning.
- Built machine learning models for forecasting.
- Presented teaching materials and conducted demonstrations at workshops and webinars.

Dibimbing Projects

Web Scraping

wikipedia website using request and BeautifulSoup

Data Modeling

create data warehouse star schema in postgreSQL using dummy data

create airflow DAGs with PySpark

Transformation data from aggregation and load to PostgreSQL

Streaming Processing

It streams data through Kafka, processes with Spark Streaming, and stores the aggregated results in PostgreSQL

Project Background

- Role:

Data Engineer

- Project:

EL pipeline: IMDb Web Scraping to BigQuery

- Goal:

Get data from external company website for data analyst team

Problem Statement

The data analyst team needs access to external data from IMDb to perform accurate and timely analysis. However, manually retrieving and processing data from external sources like IMDb is inefficient, error-prone, and time-consuming. The lack of an automated process for extracting and loading data into BigQuery creates delays in delivering insights.

Success Metric:
Ensuring the data analyst team has well-structured data for their visualization. Success will be measured by the pipeline's ability to accurately and consistently retrieve the required data, with the end goal of reducing manual work and improving the efficiency of the data extraction process.

Data Platform Understanding

Extract

Web scraping technique using selenium

Load

Stored in the staging area (data folder) in CSV and JSON formats then loaded into BigQuery

Data Understanding

Data Source

IMDb website : 20 Best Indonesian Movies

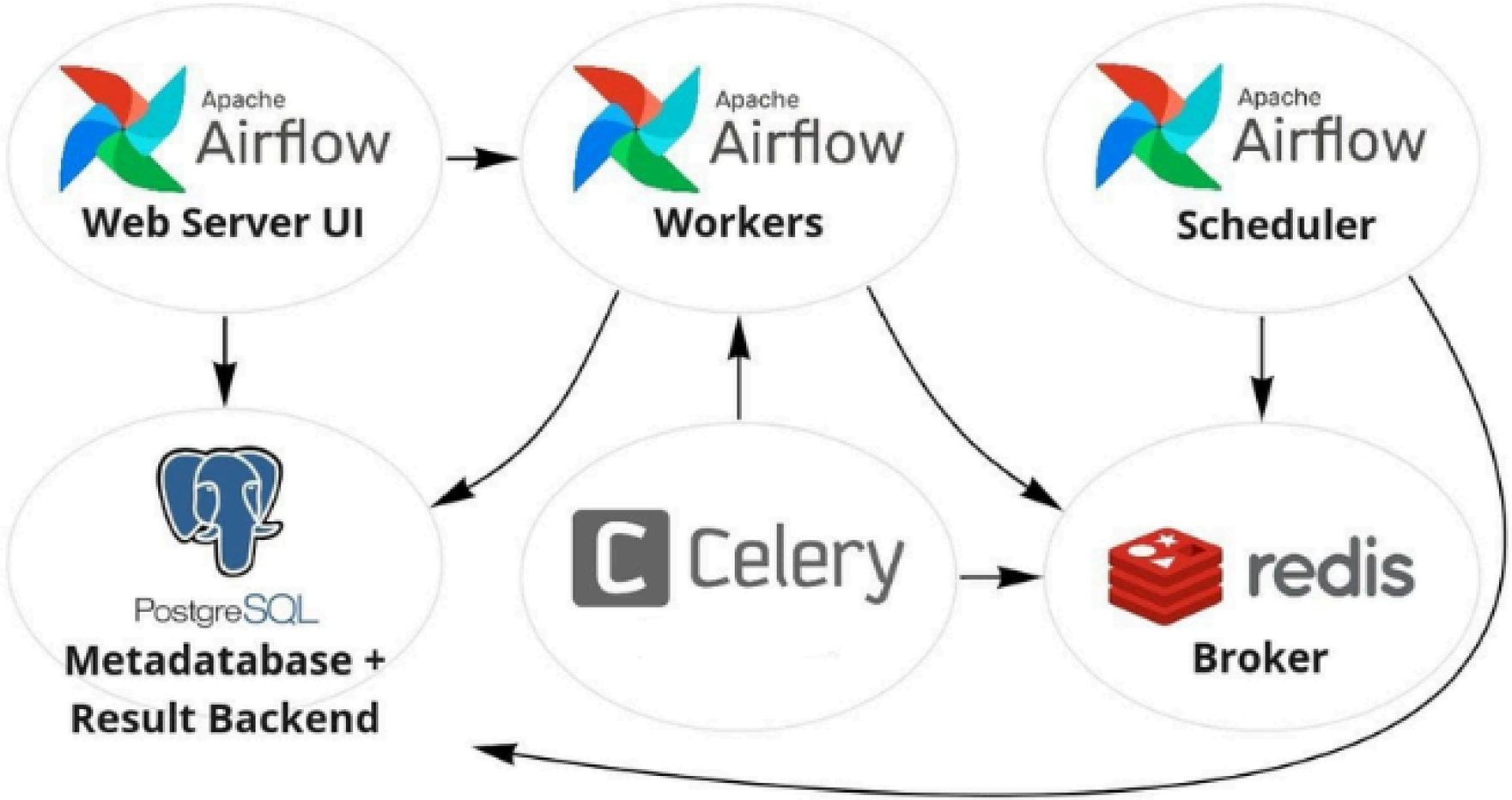
Data Type

structured data in HTML format

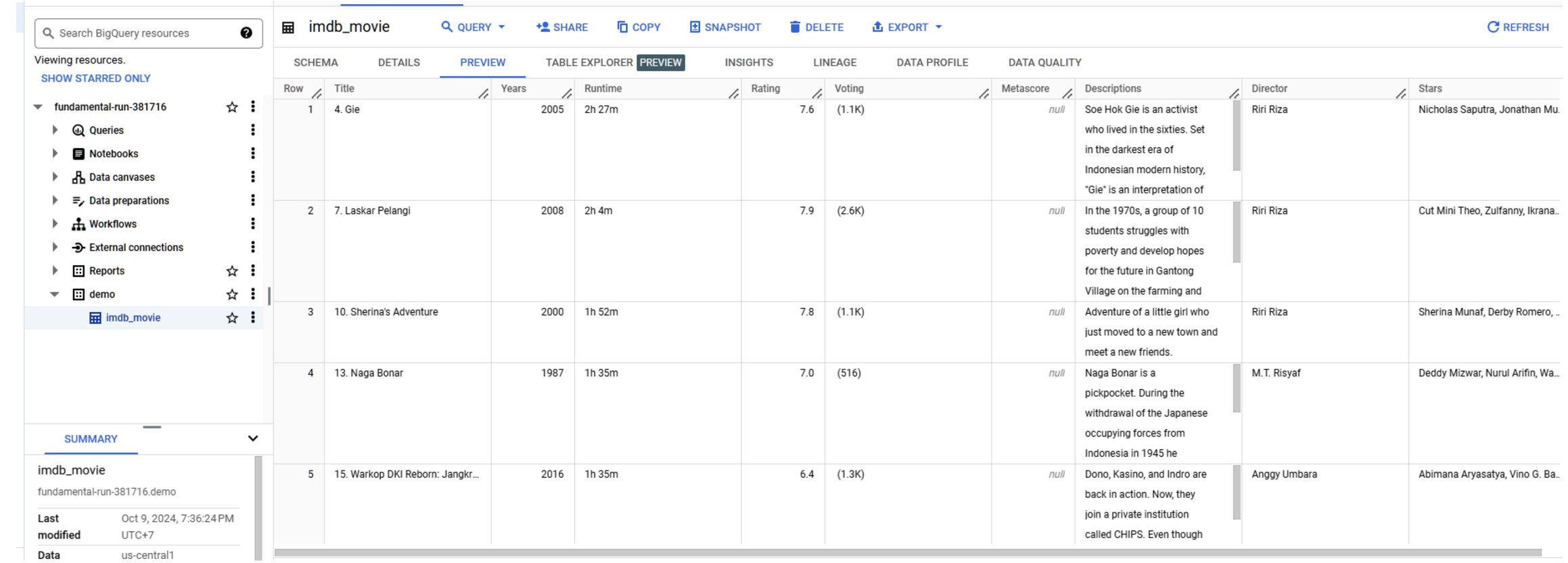
Data extraction method

Data was collected through a **web scraping** process using the **Python programming language** and using:

- **selenium** : for handle dynamic websites
- **Browser Automation**: Using Selenium to automatically control the browser.
- **Handling Dynamic Content**: Waiting for dynamically loaded elements using WebDriverWait.
- **Text Data Extraction**: Accessing HTML elements that contain data such as title, rating, runtime, and description ,etc via CSS Selectors.
- **Structured Data Storage:** Storing the collected data in JSON format and then converting it to CSV



source: https://elest.io/open-source/airflow-worker


Search BigQuery resources
Viewing resources.
SHOW STARRED ONLY
fundamental-run-381716
Queries
Notebooks
Data canvases
Data preparations
Workflows
External connections
Reports
demo
imdb_movie
SUMMARY
imdb_movie
fundamental-run-381716.demo
Last modified
Oct 9, 2024, 7:36:24 PM UTC+7
Data
us-central1
imdb_movie
QUERY
SHARE
COPY
SNAPSHOT
DELETE
EXPORT
REFRESH
SCHEMA
DETAILS
PREVIEW
TABLE EXPLORER PREVIEW
INSIGHTS
LINEAGE
DATA PROFILE
DATA QUALITY
Row
Title
Years
Runtime
Rating
Voting
Metascore
Descriptions
Director
Stars
1
4. Gie
2005
2h 27m
7.6
(1.1K)
null
Soe Hok Gie is an activist who lived in the sixties. Set in the darkest era of Indonesian modern history, "Gie" is an interpretation of
Riri Riza
Nicholas Saputra, Jonathan Mu.
2
7. Laskar Pelangi
2008
2h 4m
7.9
(2.6K)
null
In the 1970s, a group of 10 students struggles with poverty and develop hopes for the future in Gantong Village on the farming and
Riri Riza
Cut Mini Theo, Zulfanny, Ikrana..
3
10. Sherina's Adventure
2000
1h 52m
7.8
(1.1K)
null
Adventure of a little girl who just moved to a new town and meet a new friends.
Riri Riza
Sherina Munaf, Derby Romero, ..
4
13. Naga Bonar
1987
1h 35m
7.0
(516)
null
Naga Bonar is a pickpocket. During the withdrawal of the Japanese occupying forces from Indonesia in 1945 he
M.T. Risyaf
Deddy Mizwar, Nurul Arifin, Wa..
5
15. Warkop DKI Reborn: Jangkr...
2016
1h 35m
6.4
(1.3K)
null
Dono, Kasino, and Indro are back in action. Now, they join a private institution called CHIPS. Even though
Anggy Umbara
Abimana Aryasatya, Vino G. Ba..

Conclusion

The platform demonstrated its capability to process and present data effectively by automating the entire data extraction pipeline from IMDb to **BigQuery**. Using **selenium** for scraping and **Apache Airflow** orchestrated through **Docker**, the platform successfully extracts essential movie data and loads it into BigQuery, where it can be further analyzed and visualized by the data analyst team

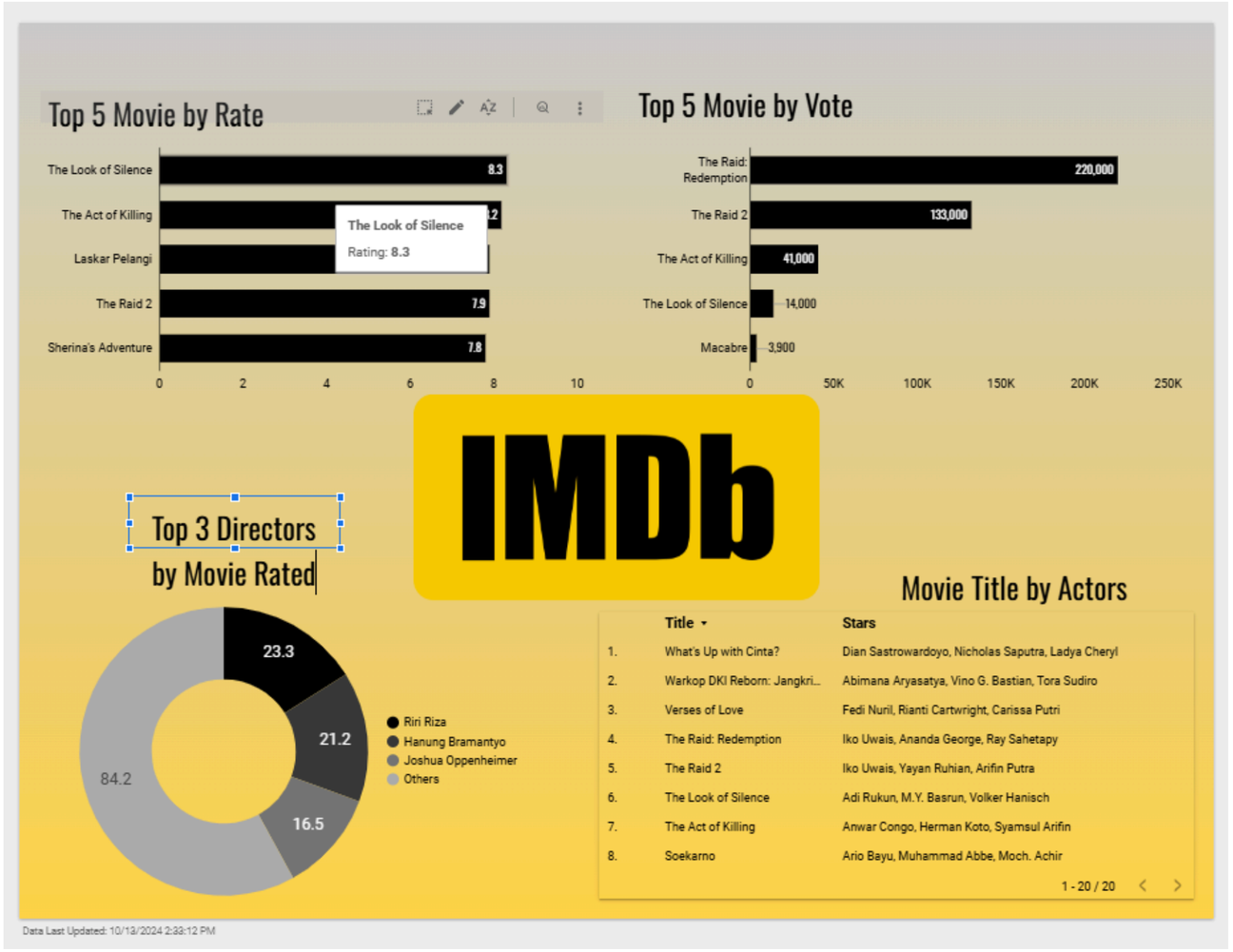
By automating the data extraction and loading process, the platform reduces manual effort, minimizes errors, and provides well-structured data for data analyst team, ultimately improving the efficiency of their work

Recommendation

Using IMDb API

Data transformation and cleaning in the pipeline

THANK YOU



link: https://lookerstudio.google.com/reporting/b458f6c0-c77b-45f8-ae3d-dde9462570b7

```sql
-- NonArray
SELECT
  -- Hapus angka dan titik di depan judul
  REGEXP_REPLACE(Title, r'^\d+\.\s', '') AS Title,

  -- Buang tanda kurung dan format ulang kolom Voting
  CASE
    -- Jika formatnya angka.titik.angka diikuti K, misal 1.1K menjadi 1100
    WHEN REGEXP_CONTAINS(Voting, r'\d\.\dK') THEN
      CAST(REGEXP_REPLACE(Voting, r'[^\d]', '') AS INT64) * 100
    -- Jika formatnya hanya angka diikuti K, misal 220K menjadi 220000
    WHEN REGEXP_CONTAINS(Voting, r'\d+K') THEN
      CAST(REGEXP_REPLACE(Voting, r'[^\d]', '') AS INT64) * 1000
    -- Jika hanya angka tanpa K atau titik
    ELSE
      CAST(REGEXP_REPLACE(Voting, r'[^\d]', '') AS INT64)
  END AS Vote,

  -- Kolom lain yang tidak dimodifikasi
  Years,
  Rating,
  Director,
  Stars

FROM `fundamental-run-381716.demo.imdb_movie`
```

Transformation