

MLOPS (DA5402) - Assignment 2

News Scraping Pipeline with Apache Airflow

Gokulakrishnan B
Roll No: DA24M007

February 10, 2025

1 Introduction

This report explains the implementation of the news scraping pipeline using apache airflow. This pipeline has two DAGs. One scraps the web, get the url of top stories, scrapes it and then store them in database after checking for duplicates, other sends email whenever new records are being added to the database.

2 Implementation Overview

The solution is implemented as two separate but interconnected DAGs:

- DAG 1: News Scraping Pipeline - Handles the core functionality of scraping, processing, and storing news articles
- DAG 2: Email Notification System - Manages state tracking and email notifications for new entries

The DAGs communicate through a status file, implementing a lightweight yet effective triggering mechanism.

3 DAG 1: News Scraping Pipeline

This DAG implements the core scraping functionality through four interconnected modules, each handling a specific aspect of the data pipeline:

3.1 Module 1: Homepage Scraping

This module initiates the scraping process by accessing the Google News homepage. It uses requests library for HTTP communication and implements proper error handling for network issues. The scraped content is passed between tasks using Airflow's XCom feature, ensuring reliable data transfer between modules.

```
1 def module_1(url,ti):
2     response = requests.get(url)
3     ti.xcom_push(key="home_page",value=response.text)
```

3.2 Module 2: Top Stories Extraction

The second module focuses on parsing the homepage content to locate the "Top stories" section. It employs BeautifulSoup for HTML parsing and implements pattern matching to reliably identify the correct link.

```
1 def module_2(pattern,ti):
2     response = ti.xcom_pull(key="home_page",task_ids="module_1")
3     soup = BeautifulSoup(response, 'html.parser')
4     links = soup.find_all('a')
5     for link in links:
6         if pattern.lower() in link.get_text().lower():
7             href = link.get('href')
8             if href:
9                 return f"https://news.google.com{href[1:-1]}"
```

3.3 Module 3: Article Extraction

This module handles the complex task of extracting article details from the Top Stories page. It handles lazyloading by refresing the page multiple times, configured by the user.

```
1 # Extract publication date and handle missing data
2 time_element = article.find('time')
3 publication_date = time_element.get('datetime') if time_element else
4     None
5
6 # Smart thumbnail extraction with lazy loading support
7 figure = article.find('figure')
8 if figure and figure.find('img'):
9     img = figure.find('img')
10    thumbnail_url = img.get('src') or img.get('data-src')
```

3.4 Module 4: Database Operations

This module manages data persistence and implements sophisticated duplicate detection:

- Uses Jaccard similarity for fuzzy matching of headlines
- Implements efficient batch processing
- Maintains referential integrity between images and headlines
- Tracks successful insertions for notification purposes

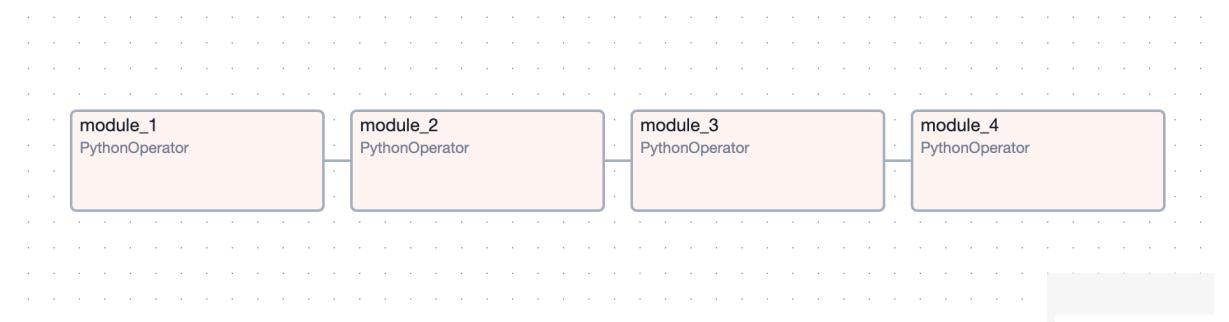


Figure 1: DAG 1 graph view

DBeaver 24.3.3 - news_headlines

Table: news_headlines

	headline_id	image_id	headline	thumbnail_url	article_url
1	1	1	Governor orders upcoming Manipur Assembly session	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
2	2	2	Trump plans to unveil 25% steel, aluminum tariffs on India	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
3	3	3	'World's biggest traffic jam': 300-km long snarl choke	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
4	4	4	Delhi CM Arvind Kejriwal submits resignation to LG	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
5	5	5	Tirupati laddu case: SIT led by CBI arrests four, say of	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
6	6	6	Who Will Be BJP's Delhi CM? Decoding Contenders, F	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
7	7	7	In one of biggest Naxal encounters yet, security forces	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
8	8	8	India planning tariff cuts ahead of PM Modi-Trump meet	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
9	9	9	Scooters, laptops at half price: How 26-year-old pulses	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
10	10	10	Grandson Stabs Industrialist 73 Times Over Property	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
11	11	11	Donald Trump hints minting of new pennies, calls coin	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
12	12	12	Prashant Bhushan, Yogendra Yadav hold Arvind Kejriwal	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
13	13	13	Revamping Yamuna riverfront top priority: BJP's Parvez	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
14	14	14	Stock Market Today: All You Need To Know Going Into	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
15	15	15	Delhi election results: Cong draws blank, loses secure	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
16	16	16	You lost due to curse of Yamuna, LG V K Sivaram tells A	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
17	17	17	Watch: India Head Coach Gautam Gambhir's Standing	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
18	18	18	'We live in an uncleancy,' say netizens after Bengaluru	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
19	19	19	'Shameless display...': Swati Malwal criticises Atishi	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
20	69	69	Opener Rohit Sharma surpasses Sachin Tendulkar in	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
21	20	20	'People are worried...': BJP's Mohan Singh Bishwakarma pledges	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
22	21	21	Israel PM says 'we'll do the job' of executing Trump	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
23	22	22	Maha Kumbh: Over 41 Crore Devotees Take Holy Dip	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
24	23	23	Swept Out Of Delhi Assembly, Here's Where AAP Chie	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
25	24	24	Delhi's new CM likely to be sworn in after PM Narendra	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
26	25	25	INDIA allies sharpening knives, Congress's 'all about A'	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
27	26	26	New Income Tax Bill to be concise with nearly 30 perc	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
28	27	27	Energy booster from US on menu in runup to PM Mod	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
29	28	28	Don't expect Centre to extend funds for programmes	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
30	29	29	Aero India 2025: Air Chief Marshal A. P. Singh, Army	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
31	30	30	Metro fare hike: Bengaluru MPs, netizens call it unfair,	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
32	31	31	Stocks to watch, February 10: Engineers India, L&T Fin	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com
33	32	32	Q3 results: Varun Beverages, Grasim Industries, Eicher	https://news.google.com/api/attachments/CC8l0Nn	https://news.google.com

Figure 2: Headlines table

DBeaver 24.3.3 - news_images

Table: news_images

	image_id	image_data	created_at
1	1	ÿØÿÙ JFIF ÿØ ... [12K]	2025-02-10 05:19:04.195
2	2	ÿØÿÙ JFIF ÿØ ... [11K]	2025-02-10 05:19:04.195
3	3	ÿØÿÙ JFIF ÿØ ... [24K]	2025-02-10 05:19:04.195
4	4	ÿØÿÙ JFIF ÿØ ... [17K]	2025-02-10 05:19:04.195
5	5	ÿØÿÙ JFIF ÿØ ... [36K]	2025-02-10 05:19:04.195
6	6	ÿØÿÙ JFIF ÿØ ... [18K]	2025-02-10 05:19:04.195
7	7	ÿØÿÙ JFIF ÿØ ... [26K]	2025-02-10 05:19:04.195
8	8	ÿØÿÙ JFIF ÿØ ... [19K]	2025-02-10 05:19:04.195
9	9	ÿØÿÙ JFIF ÿØ ... [7.1K]	2025-02-10 05:19:04.195
10	10	ÿØÿÙ JFIF ÿØ ... [8K]	2025-02-10 05:19:04.195
11	11	ÿØÿÙ JFIF ÿØ ... [13K]	2025-02-10 05:19:04.195
12	12	ÿØÿÙ JFIF ÿØ ... [18K]	2025-02-10 05:19:04.195
13	13	ÿØÿÙ JFIF ÿØ ... [21K]	2025-02-10 05:19:04.195
14	14	ÿØÿÙ JFIF ÿØ ... [16K]	2025-02-10 05:19:04.195
15	15	ÿØÿÙ JFIF ÿØ ... [13K]	2025-02-10 05:19:04.195
16	16	ÿØÿÙ JFIF ÿØ ... [9.9K]	2025-02-10 05:19:04.195
17	17	ÿØÿÙ JFIF ÿØ ... [8.6K]	2025-02-10 05:19:04.195
18	18	ÿØÿÙ JFIF ÿØ ... [16K]	2025-02-10 05:19:04.195
19	19	ÿØÿÙ JFIF ÿØ ... [17K]	2025-02-10 05:19:04.195
20	20	ÿØÿÙ JFIF ÿØ ... [18K]	2025-02-10 05:19:04.195
21	21	ÿØÿÙ JFIF ÿØ ... [11K]	2025-02-10 05:19:04.195
22	22	ÿØÿÙ JFIF ÿØ ... [31K]	2025-02-10 05:19:04.195
23	23	ÿØÿÙ JFIF ÿØ ... [15K]	2025-02-10 05:19:04.195
24	24	ÿØÿÙ JFIF ÿØ ... [21K]	2025-02-10 05:19:04.195
25	25	ÿØÿÙ JFIF ÿØ ... [14K]	2025-02-10 05:19:04.195
26	26	ÿØÿÙ JFIF ÿØ ... [15K]	2025-02-10 05:19:04.195
27	27	ÿØÿÙ JFIF ÿØ ... [9.8K]	2025-02-10 05:19:04.195
28	28	ÿØÿÙ JFIF ÿØ ... [8.1K]	2025-02-10 05:19:04.195
29	29	ÿØÿÙ JFIF ÿØ ... [8K]	2025-02-10 05:19:04.195
30	30	ÿØÿÙ JFIF ÿØ ... [16K]	2025-02-10 05:19:04.195
31	31	ÿØÿÙ JFIF ÿØ ... [10K]	2025-02-10 05:19:04.195
32	32	ÿØÿÙ JFIF ÿØ ... [13K]	2025-02-10 05:19:04.195
33	33	ÿØÿÙ JFIF ÿØ ... [12K]	2025-02-10 05:19:04.195

Figure 3: thumbnail(image) table

4 DAG 2: Email Notification System

Airflow is configured with the SMTP protocol to be able to send emails. Here, it is triggered by the successful execution of DAG1. Then it checks for the status file under the run directory. if exists, it checks the value in it, if not 0, it queries that much number of entries from database and headline id of latest of insert is also remembered to fetch only the new emails. Then the status file is deleted if the email is successfully sent, so that we can check for the successful execution of next iteration of dag 1.

Key components:

```
1 wait_for_status = FileSensor(  
2     task_id='wait_for_status_file',  
3     fs_conn_id='file',  
4     filepath='/opt/airflow/dags/run/status',  
5     poke_interval=30,  
6 )
```

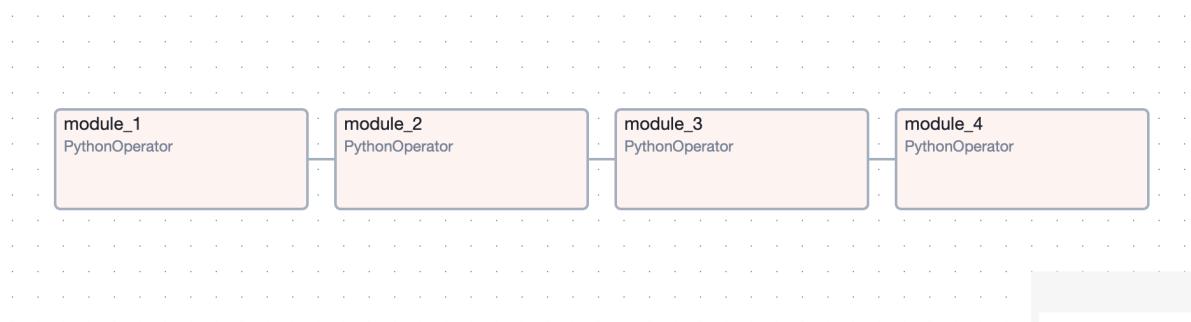


Figure 4: DAG 2 graph view

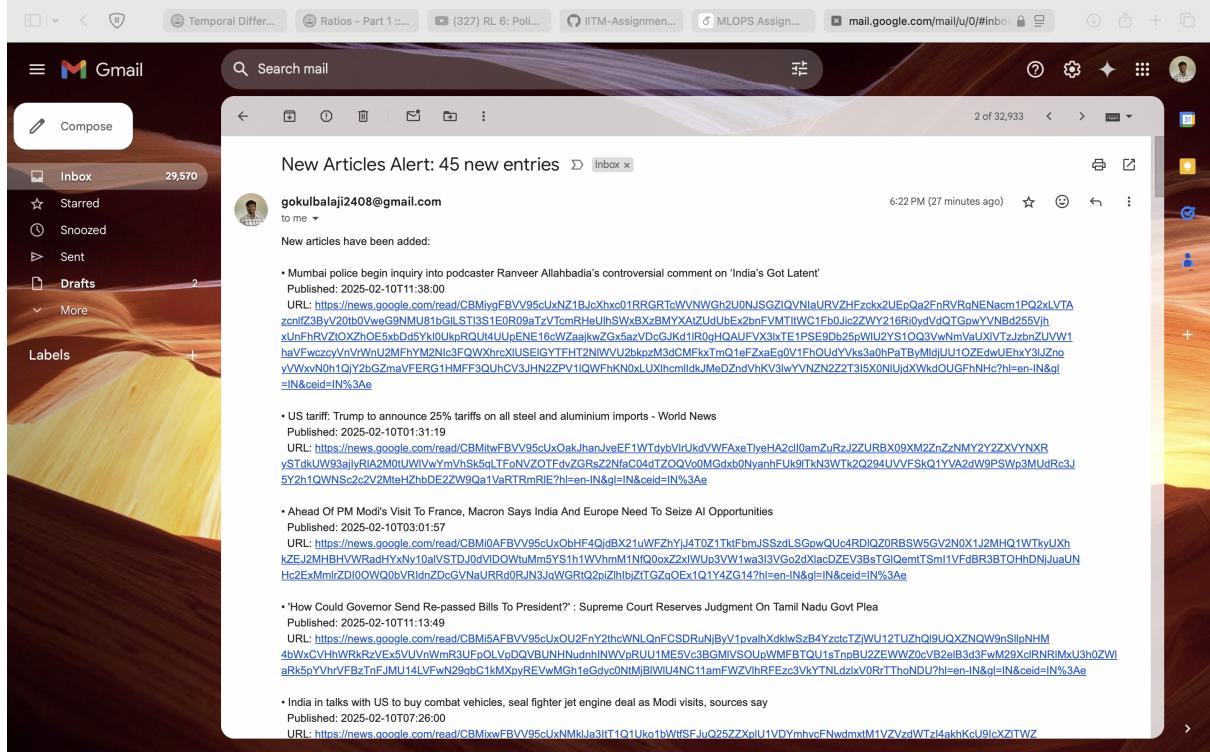


Figure 5: Email screenshot

5 Workflow Integration

The integration between the two DAGs is achieved through multiple mechanisms:

- File-based triggering using status files
- State persistence using JSON for tracking processed entries
- External Task Sensor for DAG synchronization
- Shared database access for data consistency

Version: v2.10.4
Git Version: .release:c083e456fa02c6cb32cdbe0c9ed3c3b2380beccd

Figure 6: Airflow Dashboard showing both DAGs

6 Database Implementation

Two main tables:

```

1 CREATE TABLE news_images (
2     image_id SERIAL PRIMARY KEY,
3     image_data BYTEA NOT NULL,
4     created_at TIMESTAMP DEFAULT CURRENT_TIMESTAMP
5 );
6
7 CREATE TABLE news_headlines (
8     headline_id SERIAL PRIMARY KEY,
9     image_id INTEGER REFERENCES news_images(image_id),
10    headline TEXT NOT NULL,
11    thumbnail_url TEXT,
12    article_url TEXT,
13    publication_date TIMESTAMP,
14    scrape_timestamp TIMESTAMP DEFAULT CURRENT_TIMESTAMP,
15    UNIQUE(headline, article_url)
16 );

```

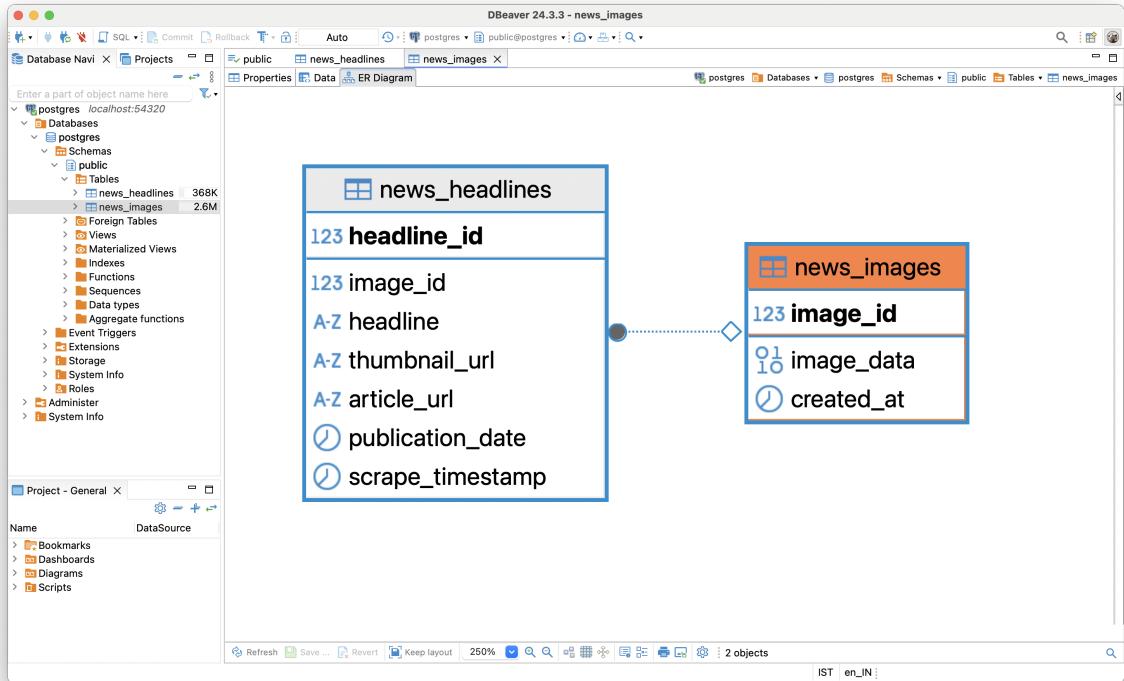


Figure 7: Database ER Diagram

7 Conclusion

The implementation successfully creates an automated news scraping and notification system using Apache Airflow. The system demonstrates the power of workflow orchestration and the benefits of using appropriate tools for automation tasks.