Scraping of Google News Top Stories Using Airflow

DA5402: MLOPS

Assignment – 2 Report

Rajat Abhijit Kambale [DA24M014]

DEPARTMENT OF DATA SCIENCE AND ARTIICIAL INTELLIGENCE

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

CHENNAI – 600036

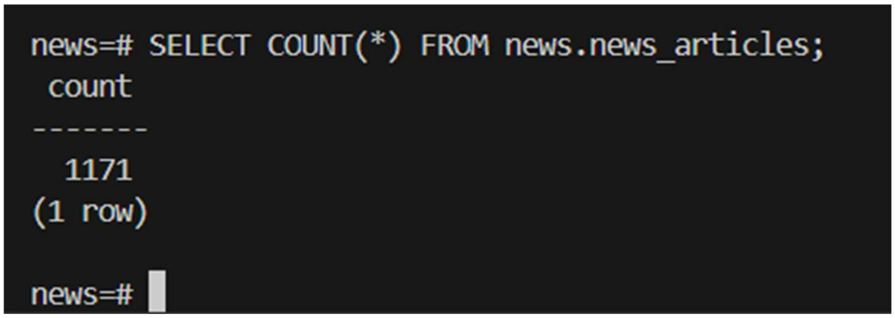
DAGS

1. Scraper DAG:

The news_scraper DAG is in charge of automatically extracting the most important news items from Google News, saving them in a PostgreSQL database, and sending out an email alert whenever new content is published. The first step involves making sure the database is configured properly using a PostgresOperator, which generates a table (news.news_articles) to hold news articles that have been scraped. The PythonOperator, which is at the heart of this DAG, uses a script called scrap.py to extract the most recent news headlines and related metadata. After that, the scraped data is momentarily saved in XComs so that it may be accessed by further tasks.

After the data is available, a different PythonOperator runs a function from dbload.py to add the news stories to the PostgreSQL database. A UNIQUE constraint on headlines and timestamps is used to prevent duplicate entries. As a trigger for the send_email_dag, the DAG then creates a status file with the quantity of recently added articles. In order to enable automated email notifications, the workflow's last task, a TriggerDagRunOperator, initiates the send_email_dag. As a scalable system for automated news monitoring, this structured process guarantees effective and dependable news aggregation, storage, and reporting.

Data stored in postgres



```
news=# SELECT COUNT(*) FROM news.news_articles;
 count
-------
  1171
(1 row)

news=# 
```

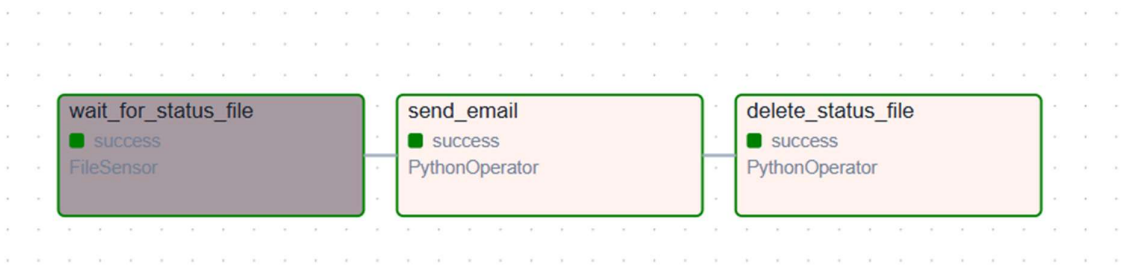
```
-----------------------------------------------------------------------------------------------------------------------------+----------------------------
 1277 | ‘It was for everyone’s safety’: US deportees speak about handcuffing, chains and emotional struggles | UklGRiwEAABXRUJQVlA4TCAEAAAvd8ACELVQrm1bta30MKjS3n7kn+P9ONTH1zu/uxu
ERpIcScEfVduZ8/+QQgIkSaZtnWfbtm3btm3btm3btvVt27Z5fe9MQNWZrP4kXIzzEcP+Thq5TXl4I91/4enEk3CO5h0AblDNU75QydHgpHX9Ybx4DYp9FP8QpgB68I2bf4oYcPzm7zOIuL55Q1QoQAKDAYTv/OM7e3Bhfff7g58cIRE/+
PP+GWgOohT8Rm+RB3ZjqDrX9hzDk7Cz9hVOPC9InAkuIG5w64/F9vrjFKwIQexPkbQ94/+QZ2F8AmF+TCKy8WZ7i/3G221B3H3+g2ML30FiYyMT9+3B51HUeXOcgpuVOcf2BJyE0zHFFMzBDVBedxuVEUhThQHdFFx840rwE4d6pJFTIU8
HWQ94CIIy+ZAV3vTigz6uYU5GqNKNUCSOL1XwbThuj5ADVht1W22YgR5uFUd7HCF8Z/LeqO76k/nfhv5clZOws0Y3K3kK3p6tCg6QcAewKOhQH3L0sgygGSPMxBeqC9ZRA8XDwckq51cU8KZ3+47K+F7coQuEtMFL4mjDj4HR9HuxsY2DU
Nogb6PjAtywBDU8QYq58RWziuHxE6vq9NWbNr1d6xPw9nTY7T/bSeJJ+PUIVbQCUFYUj3yYSRgIAZBqJBXOwyFmeU11d/kXriuUh3mjv8fcK3pBhfCoegr3cIvEbWOoENqGwgX+tGyEb5SyuEEBbtu6bhsdIbVt9HYSdrumNyt1Ck7AmbJ
fnhS3ja7iNqgqLoBHM67wNAgUwMg5BisaAJjjP5/hAIHETKO5rgvp4QTHMAUJYClSR3b3B30kDiwbHH0sbLd9vNhGDBbb9gqFII6bbHO+EGZpG9PB92kbE7drdwJOwqnQ82Hd23PczkPwAmDrYpyqUQ336CAK5ioufwuu0RTiwx3loVXje
aQMKopWr0dk7K+4wusicXztkrfjxkCH4oYuSoXzPj/fkO7zcKlghqFo3bpf1A1cZzROwgm4nbJOwblmVYgBZFaoDNuwGKqoD72CsRo7ETL4I2uRMNiibLTECV9BWANTXBw8jZEa9Rd8uXg4frQRheQGZReLk7EH8nm4V/wZCWE0UNwHyA+
NgavzFDtJPAFnUm/XokvChnpFwaCOymFPwJgOCXoLoZEWi2CLEz7hSvMwoPzhCCcJEhNksTKYGWNg2BAP4AaC1gilGOTlBvbNHvJkpMfM+IXkBl2FyxY6G0wQdhfcXb6RubHH/zH0cnvDnzFZ3VyGk/9OjXRJHa+uHCBZrHC9CB4VET1wj
BaO3LhyFckfDBYdwyO+A9hCua/xEDRCKIwfsJIxYvGG+DYExvkikn98vV75zTh/P0Ae4k8fUPb+/cVDUHwrrj+ghhfX6/UnH6MZ9MZYUcv/jekuzwk8QX+H/1k= | 2025-02-16 14:00:07.922528
 1278 | "Treacherous Routes, Phones Confiscated": Indians Recount Deportation Horror                         | iVBORw0KGgoAAAANSUhEUgAAADkAAAA5CAYAAACMGIOFAAADGElEQVRoge2WW0gUURi
Av3Fd0TDLVoU0KtsStQyikkILKrpYRA9BERJlBUYlUlmQQkX40EVSyAclyaio6E26K2pKVuhmKZJEbpZplu5qV1fLdXowRyxzx+hCp/PBwO755798nN0zo2RPDqpTVdWMoCiKYnUTWRBAVVWz298e4k8gJUVBSoqClBQFKSkKUlIUpKQoS
ElRkJKiICVFQUqKgpQUBXdXN8TXNQHQ43TS/uQx5WmHabhVqMU3WGrwHO1LVU4WprBwxkXN/6UD3s9Mx5KR9t366rwb+E2NoOluGVfWrxmyhu6ddDMYMIWGszQ7l8A5UUPe293p4MPLJu3qaG0ZEHfYbVrs46vm/vW2Nm29s70NgOmb4vE
y+Q3In7h4GX5TIwCoSD/qcnaXO9nHs4KbjAqehO/kKUQfSOVSzIIBcY/mFpSxQQA0FBdSkBA/ID4uej4rTl8AoOzQfqxX83rzRvoQ96AWgJJ9u3lemA+A5xgT64ru4OHtzcyEndw+mKLVmpW4G4AXpbd4XWlxObvunbTV1lCSnASA75QQg
pfEaLGus+cJOHkWz10pqC2tg+Y33i7F+fnzVwFfl/062+xU5WQBELo2Fp/xEwAIXrocU2g4ABUZx3TNPqyD53WlhcayUgBmbEvU1p31z3o/OBz02Gwu6zi7unT1qz6VTYetFYPRyOxde4H+XXxeVEBr9UNddYZ9ut4/kQ6A/7QIog6kYjB
64JW4g/fRkXzauhlDeNigeUFzozEYjQC8a2jQ1au7o4PKzAwAzCtWEbknmTEhoaiqSsUgh9GP0P2f7OOVpRzrtcuYl69k2vo4bf3twnmYwnp/RuMXLCK2tFyLuRmNjPAPAODN0zqaK+7p7ld78RwRG7cwamIwM+K3A1Cffx37oxrdNX7qO
VmclEjNmVztBPwWd08vvAODtGuEfwAOu536/BtcjYtFdTp19+rp7qb8+BHtu6qqgz5ShkLJMgeqw8r4B/kv3nikpChISVGQkqIgJUVBSoqClBQFKSkKUlIUpKQoSElR+D8kFUWx/u0hfieKoli/AImA+9jNnsP1AAAAAElFTkSuQmCC

                                                                                                                             | 2025-02-16 14:00:07.922528
```

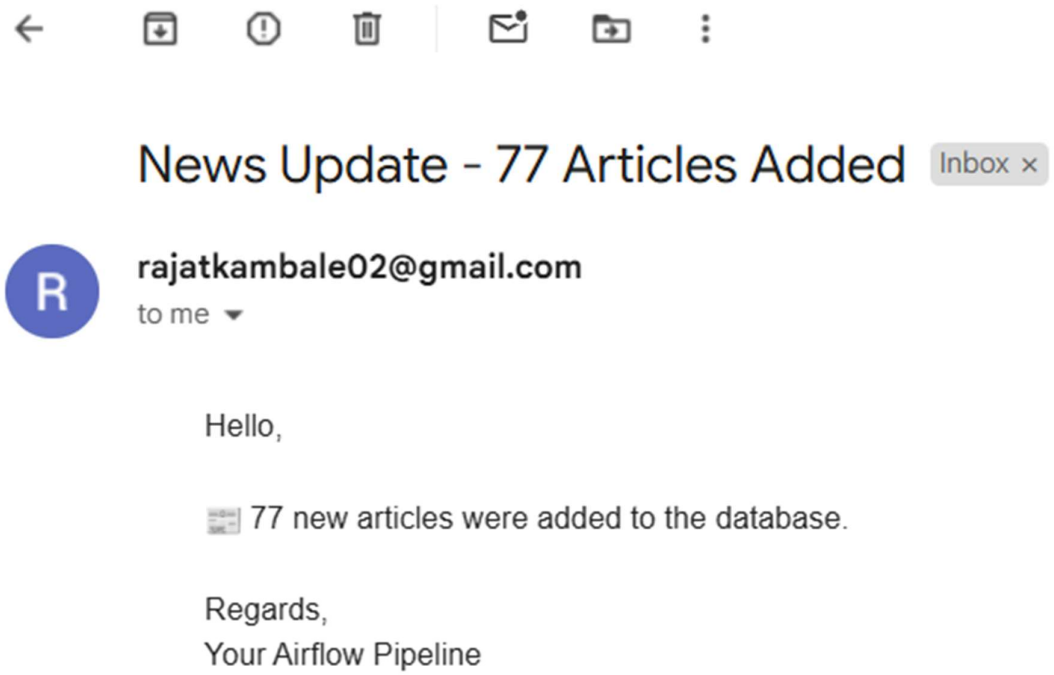
2. Send Email DAG:

When new articles are added to a database, the send_email_dag is intended to automatically send out email notifications. It starts with a FileSensor that keeps an eye on a specified status file (status) all the time. It signifies the addition of new articles if the file is present. After that, a PythonOperator is triggered by the DAG, which creates an email notice after reading the file's number of new articles. Using login credentials that are safely kept in Airflow Variables, this email is delivered via SMTP. The quantity of newly discovered articles is dynamically reflected in the subject and text of the email. The job logs the occurrence and ends without sending an email if no new entries are discovered.

After the email is sent, the status file is deleted by another PythonOperator as the last step in the DAG to guarantee the system stays clean. By doing this, redundant alerts are avoided, and notifications are only triggered by new data in subsequent runs. Because the DAG is set up with error handling and recording, it is possible to monitor problems (such SMTP authentication errors). Furthermore, Airflow's retry procedures make sure that the pipeline as a whole is not disrupted by transitory failures (like SMTP outages).




```
1f0a697775c8
 ▶ Log message source details
[2025-02-16, 19:31:41 IST] {local_task_job_runner.py:123} ▶ Pre task execution logs
[2025-02-16, 19:31:45 IST] {sendmail_dag.py:70} INFO - ✅ Email successfully sent: 77 new articles.
[2025-02-16, 19:31:45 IST] {python.py:240} INFO - Done. Returned value was: None
[2025-02-16, 19:31:45 IST] {taskinstance.py:340} ▶ Post task execution logs
```


News Update - 77 Articles Added
Inbox ×
R
rajatkambale02@gmail.com
to me
Hello,
77 new articles were added to the database.
Regards,
Your Airflow Pipeline