Assignment 3  
Building ETL pipelines with Airflow

short line

Joanne Siarivita

StudentID:24714673  
19/04/2024

94693 – Big Data Engineering

Master of Data Science and Innovation

University of Technology of Sydney

Table of Contents

[1. Executive Summary 2](#_Toc164427984)

[2. Business Understanding 3](#_Toc164427985)

[3. Data Understanding 4](#_Toc164427986)

[4. Data Engineering 5](#_Toc164427987)

[a. Environment setup 5](#_Toc164427988)

[b. Part 1: Use Airflow to load raw data into Postgres 6](#_Toc164427989)

[c. Part 2: Design a data warehouse using dbt 6](#_Toc164427990)

[d. Part 3: Ad-hoc analysis 6](#_Toc164427991)

[5. Appendices 8](#_Toc164427992)

# 

Table of Figures

[Figure 1: Screenshot of airflow storage bucket in cloud storage where datasets were loaded into. 9](#_Toc164414916)

[Figure 2: Raw schema on dbeaver postgres with relevant raw tables 10](#_Toc164414917)

[Figure 3: One-off airflow docker dag code 10](#_Toc164414918)

# Executive Summary

In today's dynamic real estate market, leveraging big data analytics offers unparalleled opportunities for informed decision-making and strategic planning. This report outlines the potential for utilizing comprehensive datasets, including Airbnb activity and census data, to optimize real estate investment and property management practices in Sydney, Australia. By harnessing the power of big data, companies can gain valuable insights into market demand, demographic trends, and property performance metrics, driving sustainable growth and maximizing returns on investment.

The proliferation of digital platforms and the availability of vast datasets have transformed the real estate landscape, offering new avenues for market analysis and predictive modeling. This report explores the integration of Airbnb data, spanning from May 2020 to April 2021, with census data from the 2016 census of Australia, to inform strategic decision-making for real estate investment and property management companies operating in Sydney. The focus is on harnessing the full potential of these datasets to extract actionable insights and drive business success.

The Airbnb dataset comprises 12 months of data, providing insights into listing performance, pricing trends, and geographical distribution of short-term rental properties in Sydney. With 22 columns in each monthly dataset, including attributes such as listing ID, neighborhood, price, and occupancy rates, this dataset offers a granular view of market dynamics over time.

Complementing the Airbnb dataset, the census data offers a comprehensive snapshot of demographic characteristics in Sydney, encompassing variables such as population density, income levels, household composition, and socioeconomic indicators. Split into two sets – selected characteristics by sex and the general community profile pack – the census data provides valuable insights into the preferences and behaviors of target demographics.

The integration of Airbnb and census data presents a unique opportunity for real estate investment and property management companies to optimize their operations and capitalize on market demand. By analyzing Airbnb listing performance alongside demographic trends, companies can identify lucrative investment opportunities, tailor marketing strategies, and optimize pricing models to maximize revenue generation.

Furthermore, insights derived from the datasets can inform strategic decision-making across various facets of the business, including property development initiatives, portfolio management, and customer segmentation. By harnessing the power of big data analytics, companies can gain a competitive edge in the dynamic real estate market of Sydney, driving sustainable growth and delivering exceptional value to stakeholders. short dash

# Business Understanding

Real estate investment and property management companies can utilize the integrated Airbnb and census data to identify untapped market segments and strategically target neighborhoods with high demand and favorable demographic profiles.

Firstly, by analyzing the Airbnb dataset, companies can identify areas within Sydney where short-term rental demand is consistently high throughout the year. They can leverage metrics such as occupancy rates, average nightly rates, and booking trends to pinpoint neighborhoods that offer the greatest potential for profitability. Additionally, they can assess the performance of existing listings in terms of guest satisfaction, reviews, and repeat bookings to understand what factors contribute to success in specific areas.

Secondly, integrating census data allows companies to overlay demographic insights onto their analysis of Airbnb activity. They can identify neighborhoods with demographics that align with their target market, such as young professionals, families, or tourists. By examining variables like income levels, household size, and age distribution, companies can tailor their property investment and marketing strategies to meet the preferences and needs of their desired clientele.

With these insights, real estate investment and property management companies can make informed decisions about where to acquire new properties, how to price them competitively, and how to market them effectively to maximize occupancy and revenue. This approach not only optimizes the return on investment for property owners but also enhances the guest experience by providing accommodations that cater to their specific needs and preferences. Overall, leveraging insights from the integrated datasets enables companies to strategically expand their short-term rental portfolio in Sydney, driving growth and profitability in the competitive real estate market.

short dash

# Data Understanding

The datasets provided for this assignment offer a robust foundation for analyzing the dynamics of Airbnb listings and demographic characteristics in NSW particularly in Sydney, Australia. The Airbnb dataset, spanning from May 2020 to April 2021, is sourced from Airbnb's database with minimal data preparation. Each monthly dataset consists of 22 columns capturing various attributes related to Airbnb listings, providing insights into trends and patterns over the specified period. Complementing this, the census data sourced from the 2016 census of Australia offers a comprehensive snapshot of the population, encompassing close to 10 million dwellings and approximately 24 million people. Split into two sets, the census data provides detailed demographic information segmented by gender, as well as medians and averages of socio-economic indicators through the General Community Profile Pack. The integration of these datasets is facilitated by the inclusion of NSW Local Government Area (LGA) codes, allowing for seamless joining based on common identifiers. Furthermore, a mapping between LGAs and suburbs aids in contextualizing the geographical distribution of data points. These datasets will be leveraged to delve into the intersection between Airbnb activity and demographic characteristics, offering valuable insights into housing dynamics and socio-economic trends within Sydney, Australia.

There were minimal attempts to prepare data, as the objective of this assignment was to build production-ready data pipelines with Airflow.

short dash

# Data Engineering

The objective of this project is to transform a single entity dataset into dimension tables, thereby creating a comprehensive data warehouse for efficient data management and analysis. The process entails developing raw data into a unified table and subsequently splitting it into multiple dimension tables, including listing, host, suburb, and LGA. The data warehouse architecture will consist of four layers: raw, staging, warehouse, and datamart. Additionally, census data tables will be integrated as dimension tables within the warehouse layer. By linking all dimensions together, the data warehouse will facilitate the handling of slow-changing dimensions (SCD), enabling easy access to historical values for analytical purposes. This structured approach ensures coherence and accessibility of data for informed decision-making and historical analysis.

## Environment setup

The initial steps of creating and configuring the necessary infrastructure on Google Cloud Platform (GCP) involves creating a GCP project, enabling essential APIs, setting up a Cloud Composer environment for Airflow orchestration, and deploying a Cloud SQL instance for PostgreSQL database storage. This phase establishes the foundation for the data pipeline and analytics environment. Once the infrastructure is in place, the focus shifts to integrating and configuring the tools required for data pipeline orchestration and analytics. This includes configuring Airflow connections to the PostgreSQL instance, installing and configuring dbt within the Cloud Composer environment, and setting up a private GitHub repository to store dbt project code securely. Proper integration and configuration ensure seamless communication and operation between the various components of the environment. With the tools integrated and configured, the next phase involves developing the workflows and automation logic using Airflow Directed Acyclic Graphs (DAGs). DAGs are created to orchestrate dbt runs, clone the dbt repository, and execute dbt commands for data transformation and modeling tasks. This phase ensures that the data pipeline operates efficiently and reliably according to the defined workflow logic.

Security and access control are critical aspects of the environment setup, ensuring that sensitive data and resources are protected from unauthorized access or misuse. This phase involves setting up IAM roles and permissions to control access to GCP resources, managing encryption keys for data security, and implementing best practices for securing database credentials and API keys. Robust security measures are essential to maintaining the integrity and confidentiality of the data environment. Finally testing, validation, and monitoring activities ensure the reliability, performance, and health of the environment. Testing involves validating the setup through unit testing and integration testing to identify and address any issues or inconsistencies. Monitoring and logging are configured to track the performance and health of the environment, providing visibility into Airflow DAG execution, dbt runs, and overall system metrics. Continuous monitoring and validation are essential for maintaining the stability and efficiency of the data pipeline and analytics workflows.

## Part 1: Use Airflow to load raw data into Postgres

Airflow was utilized solely for data extraction and loading into a PostgreSQL database, focusing on the Extract (E) and Load (L) phases, while data transformation (T) was performed using dbt. The project datasets were initially loaded into a storage bucket within the Airflow environment on Google Cloud Platform (GCP). Subsequently, a one-off Airflow Directed Acyclic Graph (DAG) was developed to retrieve the data from the storage bucket and ingest it into the designated raw schema and tables created in PostgreSQL. All 12 files were incrementally loaded into a single unified table. As per the assignment brief instructions, the scheduling interval was deliberately set to None, indicating a one-time execution of the DAG.

## Part 2: Design a data warehouse using dbt

In the data transformation process, data moves through distinct layers, namely raw, staging, warehouse, and datamart, adhering to naming conventions specified in the assignment instructions. Within the raw layer, snapshot tables are generated for dimensions based on timestamp strategies, residing in the raw layer of the data warehouse. The staging layer hosts cleaned snapshots where field names, data types, and number formats are standardized, addressing cleaning issues such as formatting inconsistencies in the listing date column. In the warehouse layer, a star schema is established with dimensions and fact tables, where dimension IDs are replaced with actual descriptions in fact tables, and models are materialized as tables, incorporating slowly changing dimensions (SCD). The datamart layer encompasses three views created to address assignment queries, including listing neighborhood, property type, and host neighborhood. Notably, host neighborhood data is transformed into Local Government Areas (LGAs) and mapped to corresponding suburbs for consistency, facilitated by provided KPI definitions aiding estimated revenue calculations. Furthermore, dbt is integrated with a private GitHub repository, with the repository [link](https://github.com/JoanneSiarivita/BDE_AT3) included in the appendices for reference.

## Part 3: Ad-hoc analysis

The analysis requires exploring various facets of Airbnb listings data, examining population dynamics, listing types, host behaviors, and financial implications. Through SQL queries, the analysis would reveal insights such as demographic variations between best and worst-performing neighborhoods, optimal listing types for maximizing stays in top neighborhoods, tendencies of hosts with multiple listings to cluster in their residential LGAs, and whether hosts with unique listings generate sufficient revenue to cover median mortgage repayments in their neighborhoods. This comprehensive assessment would provide actionable insights for Airbnb hosts and investors aiming to optimize revenue and understand market dynamics.

The SQL code for the analysis is included in the zipped file submission.

* 1. What are the main differences from a population point of view (i.g. higher population of under 30s) between the best performing “listing\_neighbourhood” and the worst (in terms of estimated revenue per active listings) over the last 12 months?
  2. What will be the best type of listing (property type, room type and accommodates for) for the top 5 “listing\_neighbourhood” (in terms of estimated revenue per active listing) to have the highest number of stays?
  3. Do hosts with multiple listings are more inclined to have their listings in the same LGA as where they live?
  4. For hosts with a unique listing, does their estimated revenue over the last 12 months can cover the annualised median mortgage repayment of their listing’s “listing\_neighbourhood”?

short dash

# Appendices

* [Link to github repo](https://github.com/JoanneSiarivita/BDE_AT3)
* [Link to airflow storage bucket](storage_bucket_assignment3)
* Part 1 Screenshots

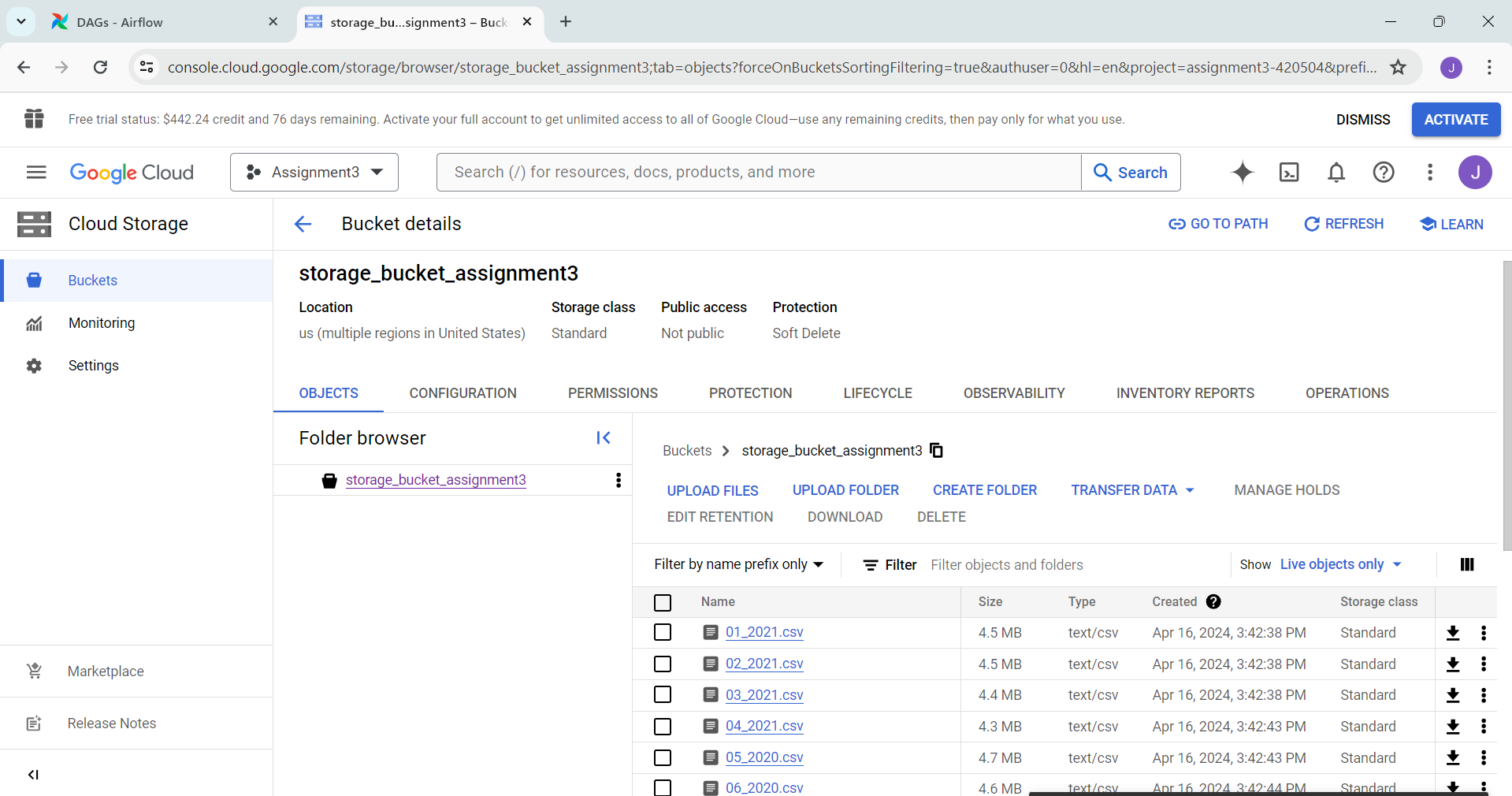


Figure 1: Screenshot of airflow storage bucket in cloud storage where datasets were loaded into.

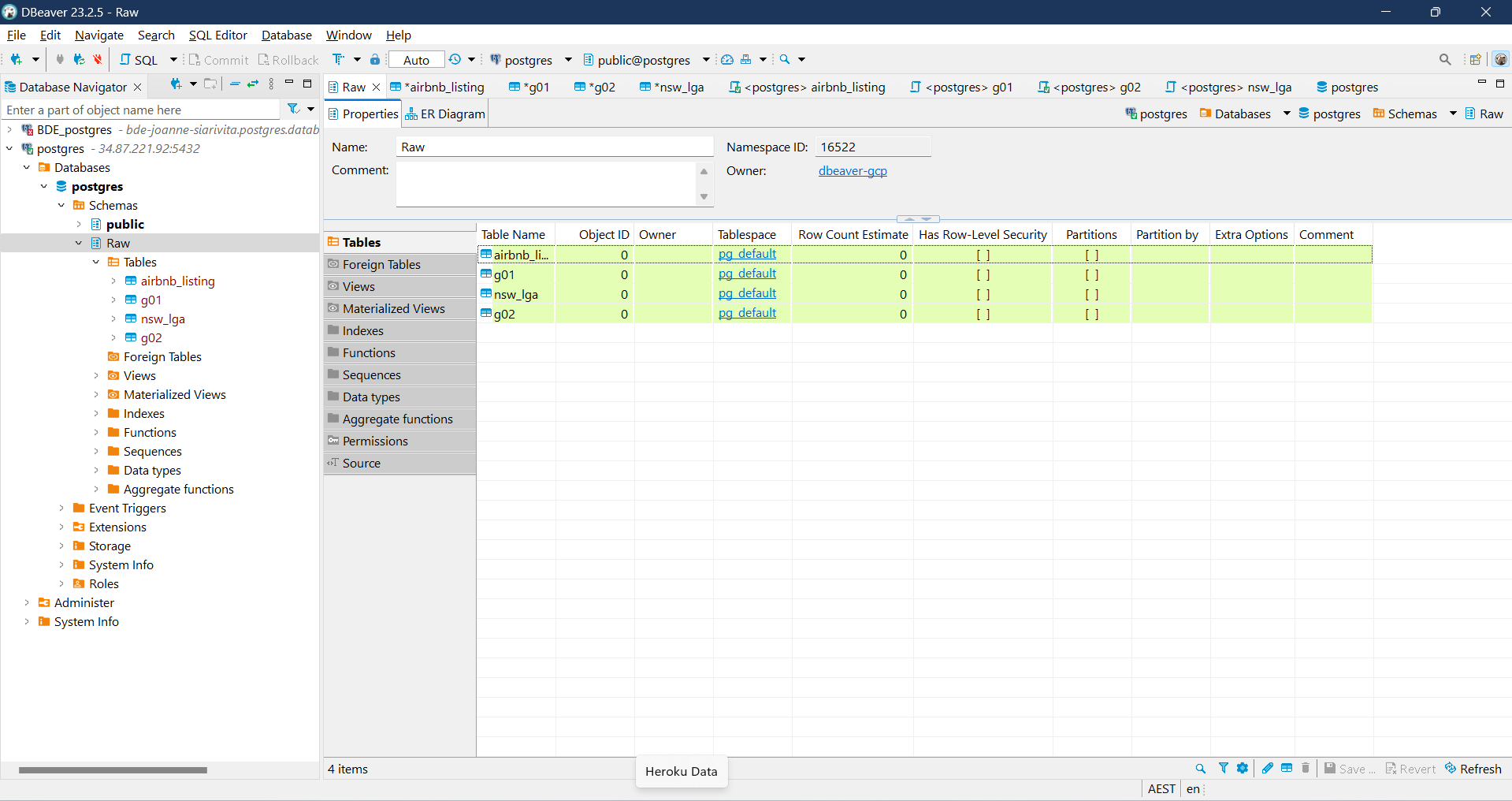


Figure 2: Raw schema on dbeaver postgres with relevant raw tables

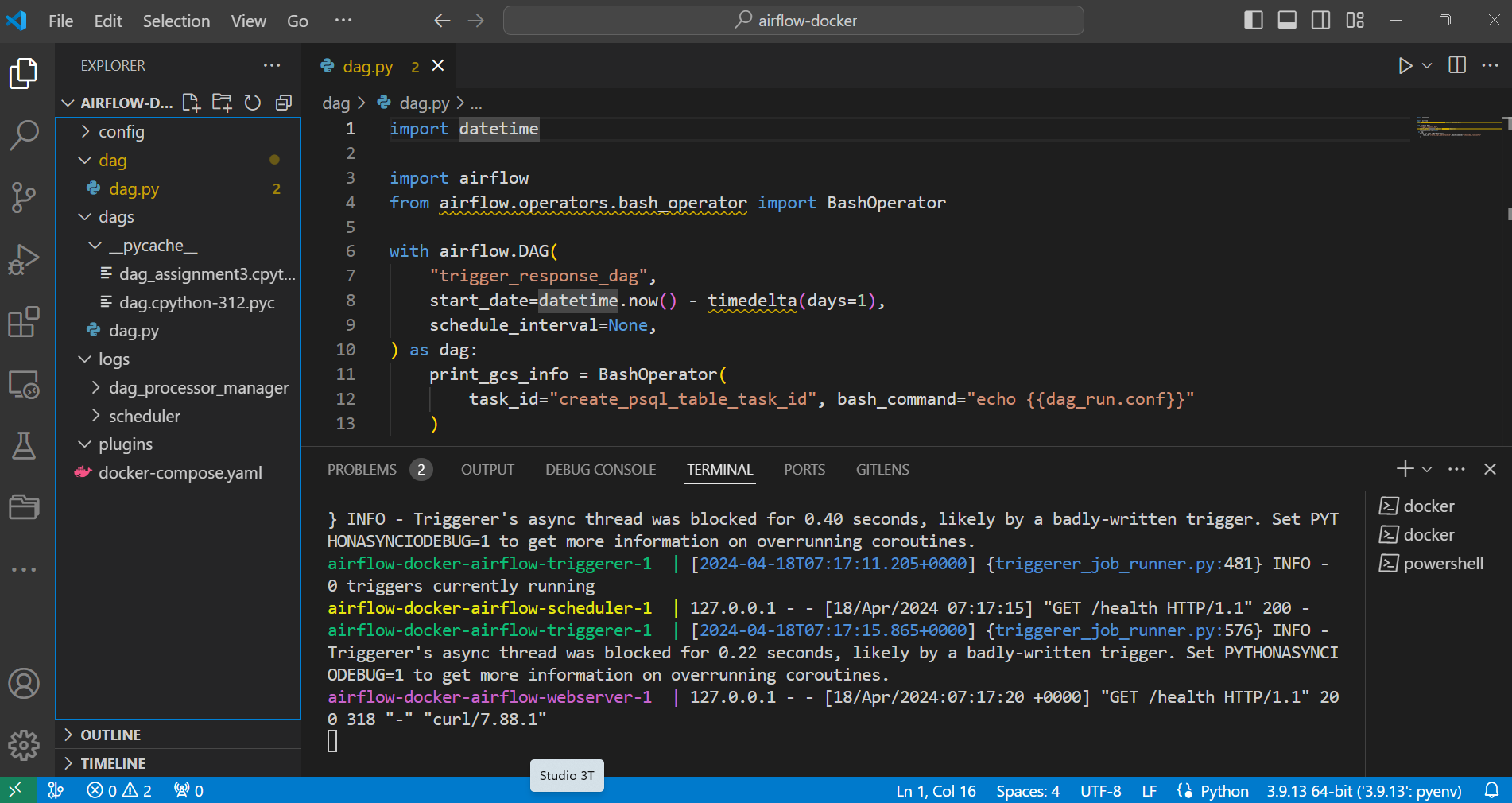


Figure 3: One-off airflow docker dag code