Assignment 3

Building ELT data pipelines with Airflow

Van Dung Do
Student ID: 25039177
28/10/2024

94693 – Big Data Engineering
Master of Data Science and Innovation
University of Technology of Sydney

Table of Contents

1. Project Overview

Airbnb connects travelers with hosts renting out properties, reshaping hospitality with a vast network of global stays and rich data on rentals, pricing, and guest reviews. This project aims to create production-ready ELT data pipelines using Apache Airflow and dbt Cloud to process and transform Sydney's Airbnb and Census data. Following the Medallion architecture (Bronze, Silver, Gold), the data is structured to support a data mart for analytical insights, along with ad-hoc analyses to address key business questions.

1.1 Project overview

The project aims to leverage Airflow, Postgres' environment using GCP and dbt Cloud for data warehousing to provide valuable insights. Objectives include setting up a data warehouse architecture, transforming data through layers, and creating a data mart for business analysis.

Specific Tasks:

- **Part 0: Data Download**
  Download the Airbnb listing data (May 2020–April 2021) and Census datasets from the Australian Bureau of Statistics, including LGA mapping.
- **Part 1: Data Ingestion with Airflow**
  Set up an Airflow DAG to load initial raw data (Airbnb and Census) into Postgres, establishing a Bronze schema for the data.
- **Part 2: Data Warehouse Design with dbt**
  Design a data warehouse using Medallion architecture (Bronze, Silver, Gold) in Postgres.
  - **Bronze:** Store raw data.
  - **Silver:** Create cleaned tables with consistent naming conventions and snapshots for dimensions.
  - **Gold:** Implement a star schema and create data mart views (e.g., dm_listing_neighbourhood, dm_property_type, dm_host_neighbourhood) to address key metrics such as active listing rate, revenue per listing, and demographic insights.
- **Part 3: End-to-End Orchestration**
  Update the Airflow DAG to include a dbt transformation step and load Airbnb data month-by-month sequentially.
- **Part 4: Ad-Hoc Analysis**
  Perform SQL-based analyses to answer business questions, such as demographic differences in high/low-performing LGAs, correlation between median age and revenue, optimal listing types, and revenue comparisons against mortgage repayments.

1.2 Data Exploration

The Airbnb is public on Inside Airbnb website: https://insideairbnb.com/get-the-data/
The Census and NSW_LGA is on Australian Bureau of Statistics:
https://www.abs.gov.au/census/find-census-data/datapacks

Table 1.1 Summary of dataset

File name	**File type**	**Number of files**	**Describe**
Listing	CSV	12 (from 5/2020 to 4/2021)	This dataset contains monthly snapshots of Airbnb listings, capturing key information about each property listed on the platform during this period.
Census data	CSV	2	This dataset includes demographic and socio-economic data from the New South Wales (NSW) Census
NSW_LGA	CSV	2	This dataset contains geographic and administrative information about Local Government Areas in New South Wales

1.3 Tools Used

- Google Cloud Platform (store data)
- Airflow (Load data)
- dbt Cloud (create data pipeline)
- DBeaver (PostgreSQL)

1.4 Challenges

Issue	Solution
Couldn't finished project because of dealing with pipeline and snapshot.	

2. Part 1 – Data Ingestion & Preparation

Corresponding file: part_1.sql and dag_1.py

Step 1: Establish Connection between GCP and DBeaver

- Set Up Google Cloud Platform:

In Google Cloud Composer, configure the environment to support connections with Cloud SQL and Airflow UI.

Enable IP Allowlisting for your local machine's IP in the Google Cloud SQL instance, allowing access over a Public/Private IP.

- Configure PostgreSQL on DBeaver:

Create a new PostgreSQL connection.

Enter the Public IP of the Google Cloud SQL instance, along with the necessary credentials.

Test the connection to ensure access to the Google Cloud-hosted PostgreSQL database.

- Create the Bronze Schema:

Using DBeaver or an SQL script, create a schema named bronze. This schema will store raw, unprocessed data directly imported from sources.

- Create Raw Data Tables:

Define 5 tables within the Bronze schema to store the raw data as follows:

listing: Store data for May 2020, representing Airbnb listings.

nsw_census_01 and nsw_census_02: Store data from the two NSW Census CSV files.

nsw_lga_01 and nsw_lga_02: Store Local Government Area (LGA) data from two CSV files.

Structure each table with columns matching the CSV headers, ensuring each field aligns with the data type in the CSV files.

- Ingest Data into Bronze Tables:

Use Airflow to automate data ingestion by creating DAGs that pull data from the source files and load them into the corresponding Bronze tables in PostgreSQL.

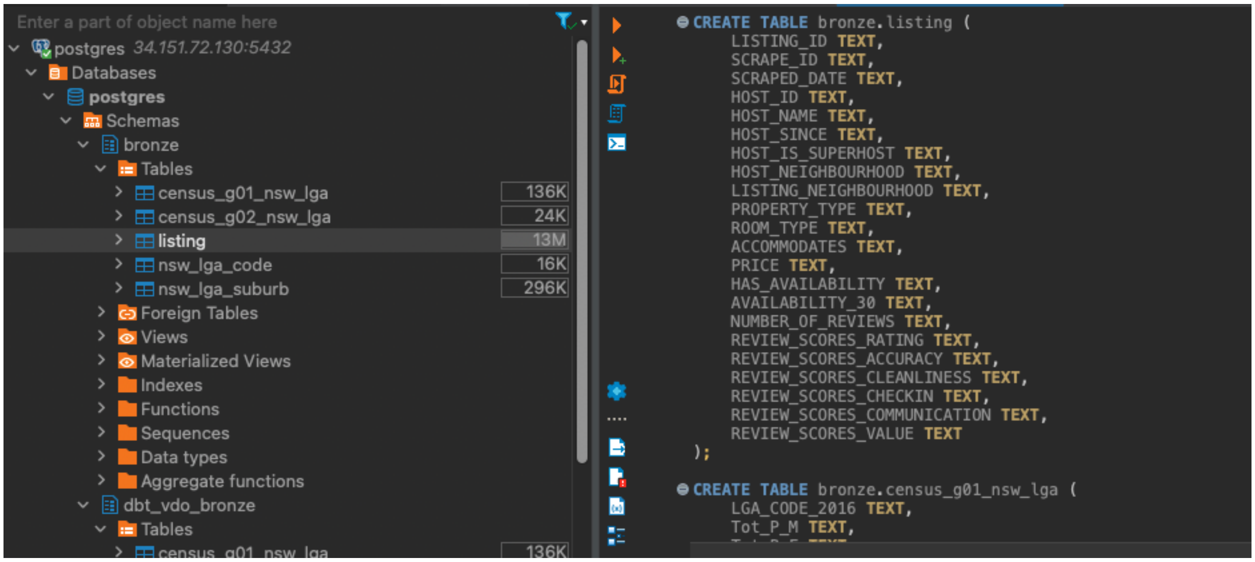


Figure 1.1: Create tables to store raw data

Step 2: Upload dag_1.py to dag folder in Google Cloud Storage (GCS) and 5 csv files in step 1 to data folder. After that, go to Airflow UI to trigger dag. This will help to load raw data from Google Cloud Storage to DBeaver.

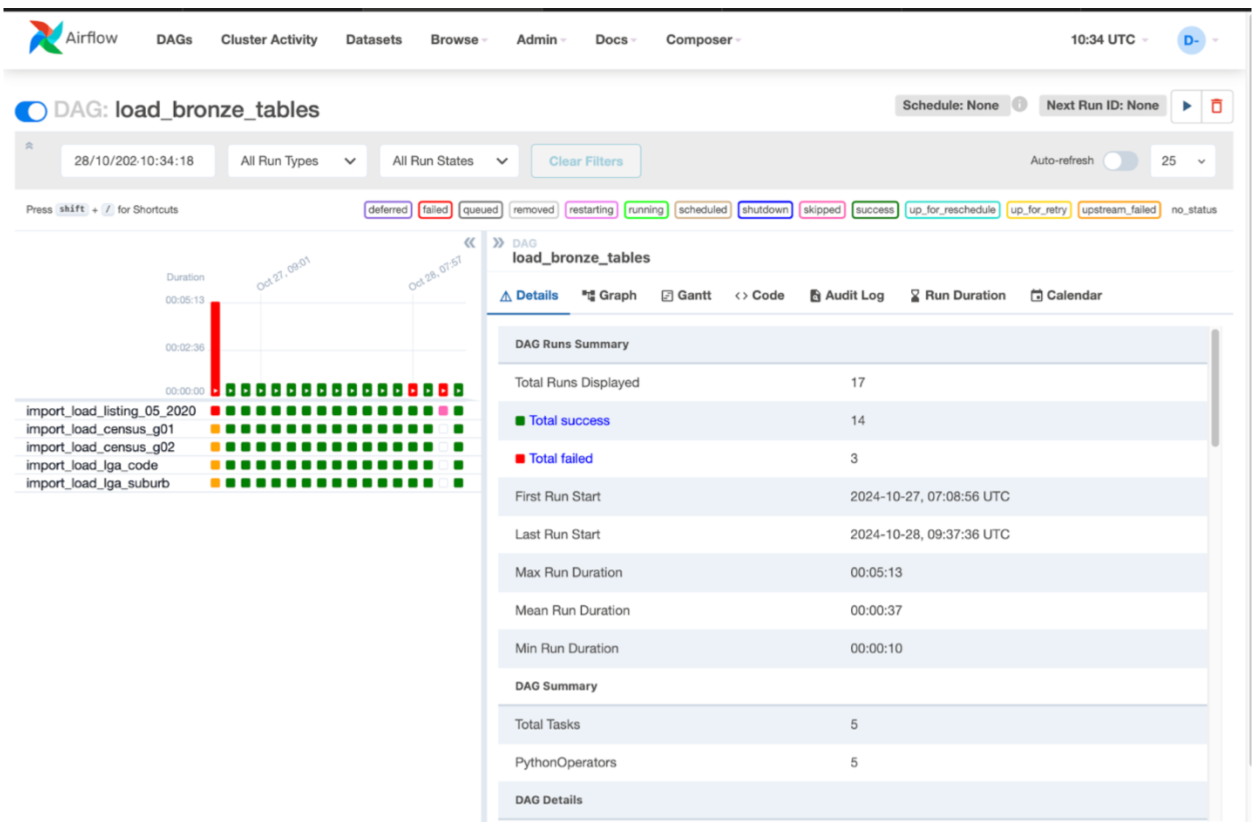


Figure 1.2: Load raw data from GCS to DBeaver

3. Part 2 – Build data warehouse with dbt

Corresponding file: dbt Cloud Files

In this part, I design and implement data warehouse architecture using dbt (data build tool) on Postgres, following the Medallion architecture pattern. This approach includes three layers:

- **Bronze** (raw data)
- **Silver** (cleaned and transformed data)
- **Gold** (curated data for analysis)

3.1 Bronze Layer

Corresponding file: dbt Cloud Files/bronze

This layer contains unprocessed and uncleaned data to serve as a historical reference.

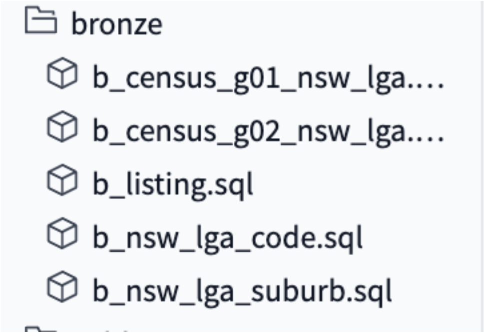


Figure 3.1 List of files in bronze layer

3.2 Silver Layer

Corresponding file: dbt Cloud Files/silver and dbt Cloud Files/snapshot

In the Silver layer, data from the Bronze layer is refined through straightforward cleaning processes, including setting the appropriate data types for columns and managing missing values.

A snapshot approach with timestamps is employed to track changes in key Airbnb features such as host and neighborhood information, property type, and room type. This approach, based on Slowly Changing Dimensions (SCD) Type 2, enables tracking of updates over time, with multiple versions of records maintained as changes occur. The snapshots acts as silver layer (cleaning and transforming data, prepare for Gold Layer).

The Silver layer queries ensure only the latest version of each record is used, organizing data into dimensions and fact tables, and loading these from the most recent snapshot version.

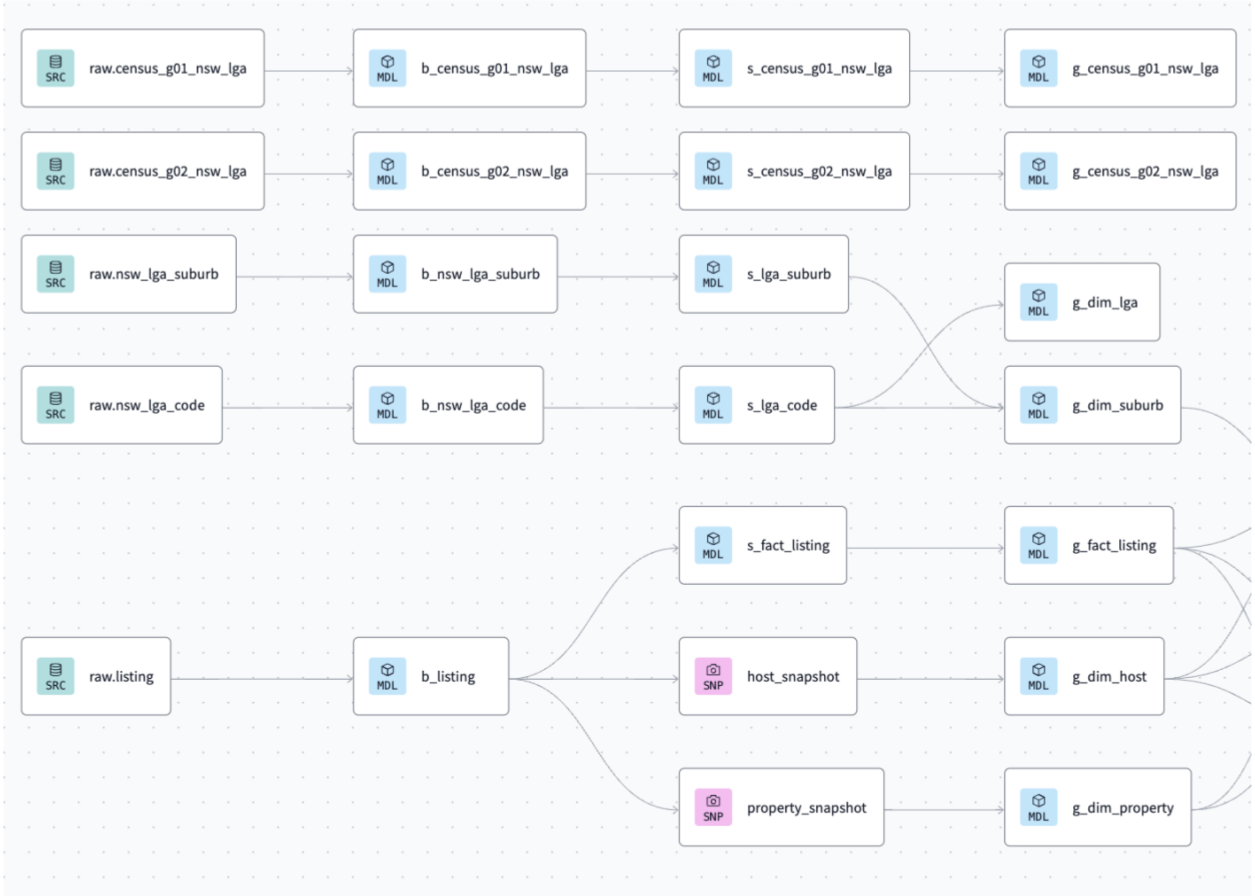


Figure 3.2 Workflow of silver layer

3.3 Gold Layer

The Gold layer is designed to hold data that is fully cleaned, aggregated, and ready for analysis. This layer is structured to align with business objectives, enabling insightful decision-making. It includes the following components:

3.3.1 Star Schema

Corresponding file: dbt Cloud Files/gold/star

Data is organized in a star schema format, with fact tables containing key metrics (like prices and counts) and dimension tables holding descriptive attributes (such as host and property information). This setup makes data retrieval more efficient and suitable for analysis.

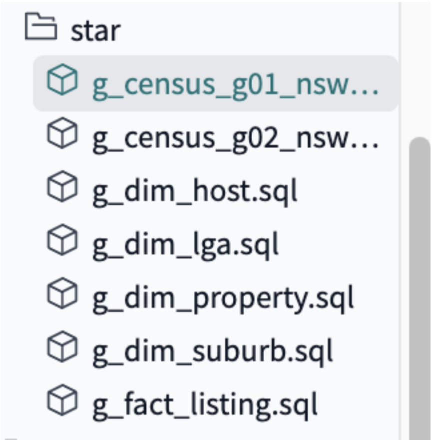


Figure 3.3 List of files in star schema

This project's star schema includes one fact table and six dimension tables for analyzing Airbnb data. The "g_fact_listing" table captures key metrics like price and review scores, as well as IDs for joining tables. Airbnb-specific details are organized into "g_dim_host" (host-related info) and "g_dim_property" (property-related info). Additionally, two dimension tables for Census data and two for LGAs mapping data provide context for broader analysis. These tables are linked through ID columns, enabling efficient querying to address business needs.

3.3.2 Data Mart

Corresponding file: dbt Cloud Files/gold/mart

In Data Mart, this step includes make 3 views with metrics defined below:

Table 3.1 Metrics definition

Metrics	How to calculate
Active listings	Listings where "has_availability" = "t".
Active Listing Rate	(total Active listings / total listing) * 100
Superhost Rate	(total distinct hosts with "host_is_superhost" = 't' / total distinct hosts) * 100
Percentage change (month to month)	((final value - original value) / original value) * 100
Number of stays (only for active listings)	30 - availability_30
Estimated revenue per active listings	for each active listing per period: number of stays * price
Estimated revenue per host	Total Estimated revenue per active listings/ total distinct hosts

3.3.3.1 dm_listing_neighbourhood

This view provides insights per listing neighbourhood and month/year with the following metrics:

- Active listings rate
- Minimum, maximum, median and average price for active listings
- Number of distinct hosts
- Superhost rate
- Average of review_scores_rating for active listings
- Percentage change for active listings
- Percentage change for inactive listings
- Total Number of stays
- Average Estimated revenue per active listings

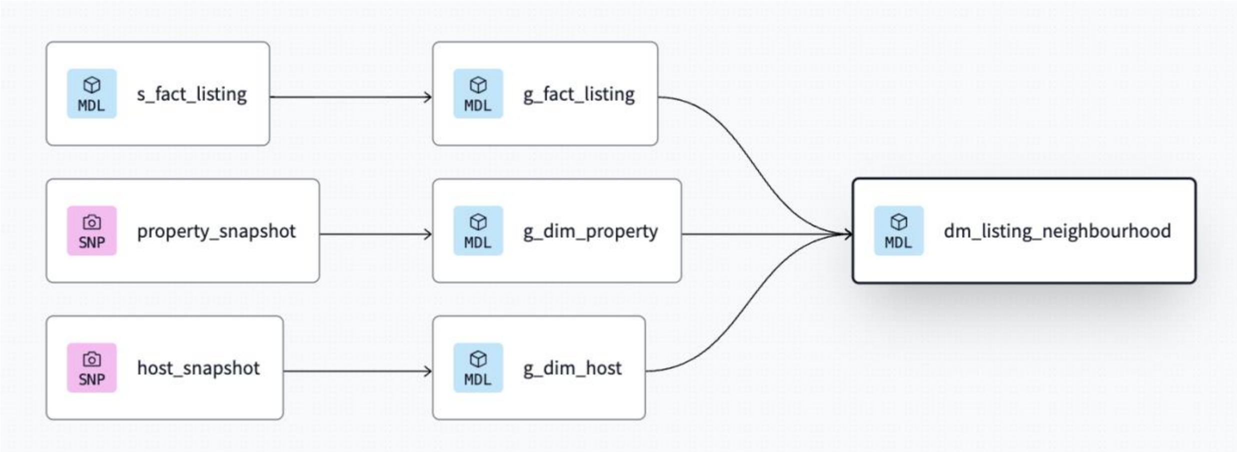
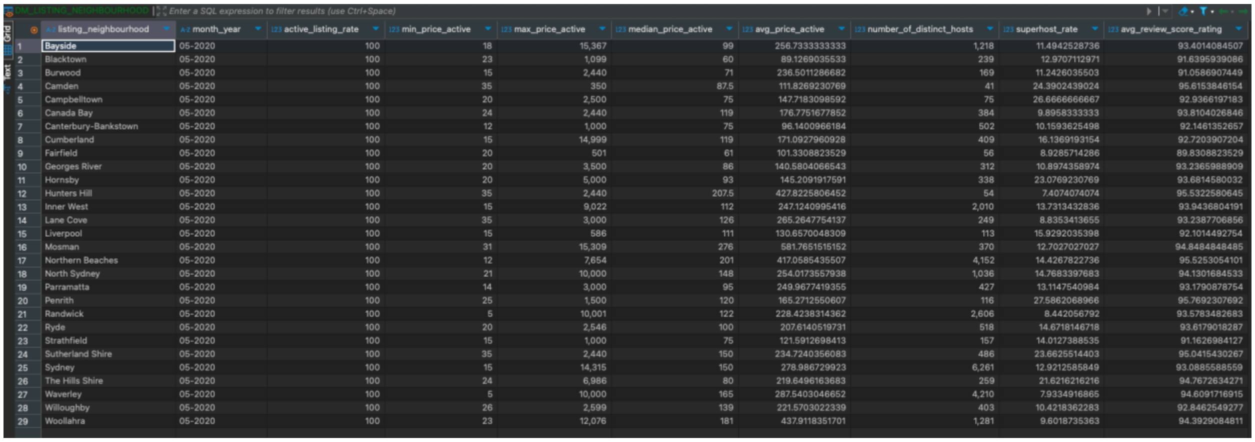


Figure 3.4 Pipeline to create dm_listing_neighbourhood



DM_LISTING_NEIGHBOURHOOD | Enter a SQL expression to filter results (use Ctrl+Space)

	listing_neighbourhood	month_year	active_listing_rate	min_price_active	max_price_active	median_price_active	avg_price_active	number_of_distinct_hosts	superhost_rate	avg_review_score_rating
1	Bayside	05-2020	100	18	15,367	99	256.7333333333	1,218	11.4942528736	93.4014084507
2	Blacktown	05-2020	100	23	1,099	60	89.1269035533	239	12.9707112971	91.6395939086
3	Burwood	05-2020	100	15	2,440	71	236.5011286682	169	11.2426035503	91.0586907449
4	Camden	05-2020	100	35	350	87.5	111.8269230769	41	24.3902439024	95.6153846154
5	Campbelltown	05-2020	100	20	2,500	75	147.7183098592	75	26.6666666667	92.9366197183
6	Canada Bay	05-2020	100	24	2,440	119	176.7751677852	384	9.8958333333	93.8104026846
7	Canterbury-Bankstown	05-2020	100	12	1,000	75	96.1400966184	502	10.1593625498	92.1461352657
8	Cumberland	05-2020	100	15	14,999	119	171.0927960928	409	16.1369193154	92.7203907204
9	Fairfield	05-2020	100	20	501	61	101.3308823529	56	8.9285714286	89.8308823529
10	Georges River	05-2020	100	20	3,500	86	140.5804066543	312	10.8974358974	93.2365988909
11	Hornsby	05-2020	100	20	5,000	93	145.2091917591	338	23.0769230769	93.6814580032
12	Hunters Hill	05-2020	100	35	2,440	207.5	427.8225806452	54	7.4074074074	95.5322580645
13	Inner West	05-2020	100	15	9,022	112	247.1240995416	2,010	13.7313432836	93.9436804191
14	Lane Cove	05-2020	100	35	3,000	126	265.2647754137	249	8.8353413655	93.2387706856
15	Liverpool	05-2020	100	15	586	111	130.6570048309	113	15.9292035398	92.1014492754
16	Mosman	05-2020	100	31	15,309	276	581.7651515152	370	12.7027027027	94.8484848485
17	Northern Beaches	05-2020	100	12	7,654	201	417.0585435507	4,152	14.4267822736	95.5253054101
18	North Sydney	05-2020	100	21	10,000	148	254.0173557938	1,036	14.7683397683	94.1301684533
19	Parramatta	05-2020	100	14	3,000	95	249.9677419355	427	13.1147540984	93.1790878754
20	Penrith	05-2020	100	25	1,500	120	165.2712550607	116	27.5862068966	95.7692307692
21	Randwick	05-2020	100	5	10,001	122	228.4238314362	2,606	8.442056792	93.5783482683
22	Ryde	05-2020	100	20	2,546	100	207.6140519731	518	14.6718146718	93.6179018287
23	Strathfield	05-2020	100	15	1,000	75	121.5912698413	157	14.0127388535	91.1626984127
24	Sutherland Shire	05-2020	100	35	2,440	150	234.7240356083	486	23.6625514403	95.0415430267
25	Sydney	05-2020	100	15	14,315	150	278.986729923	6,261	12.9212585849	93.0885588559
26	The Hills Shire	05-2020	100	24	6,986	80	219.6496163683	259	21.6216216216	94.7672634271
27	Waverley	05-2020	100	5	10,000	165	287.5403046652	4,210	7.9334916865	94.6091716915
28	Willoughby	05-2020	100	26	2,599	139	221.5703022339	403	10.4218362283	92.8462549277
29	Woollahra	05-2020	100	23	12,076	181	437.9118351701	1,281	9.6018735363	94.3929084811

Figure 3.5 view dm_listing_neighbourhood

3.3.3.2 dm_property_type

This view present information per property_type, room_type, accommodates, and month/year including:

- Active listings rate
- Minimum, maximum, median and average price for active listings
- Number of distinct hosts
- Superhost rate
- Average of review_scores_rating for active listings
- Percentage change for active listings
- Percentage change for inactive listings
- Total Number of stays
- Average Estimated revenue per active listings

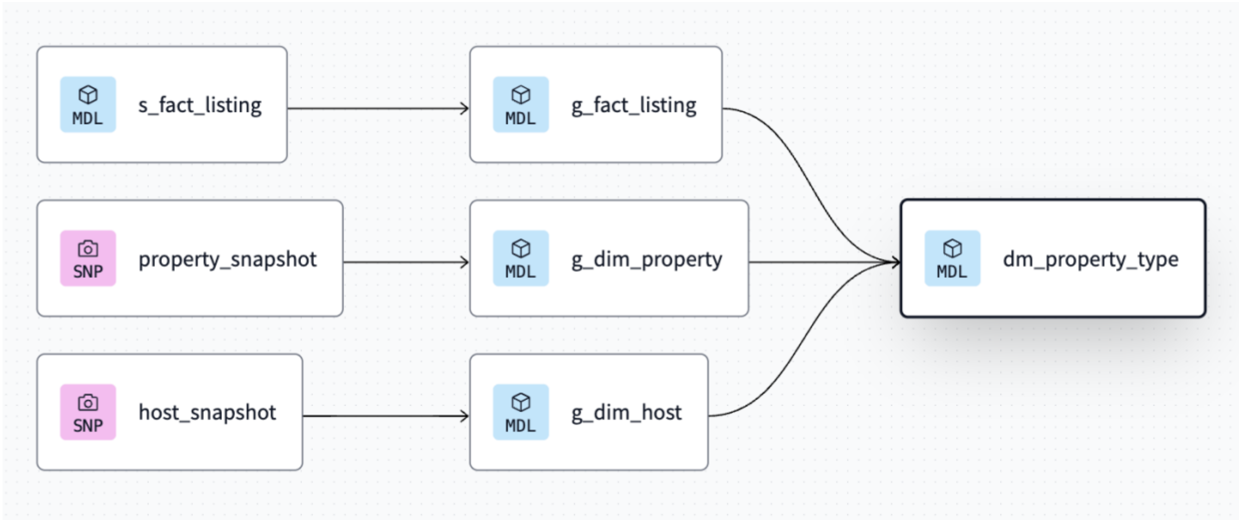
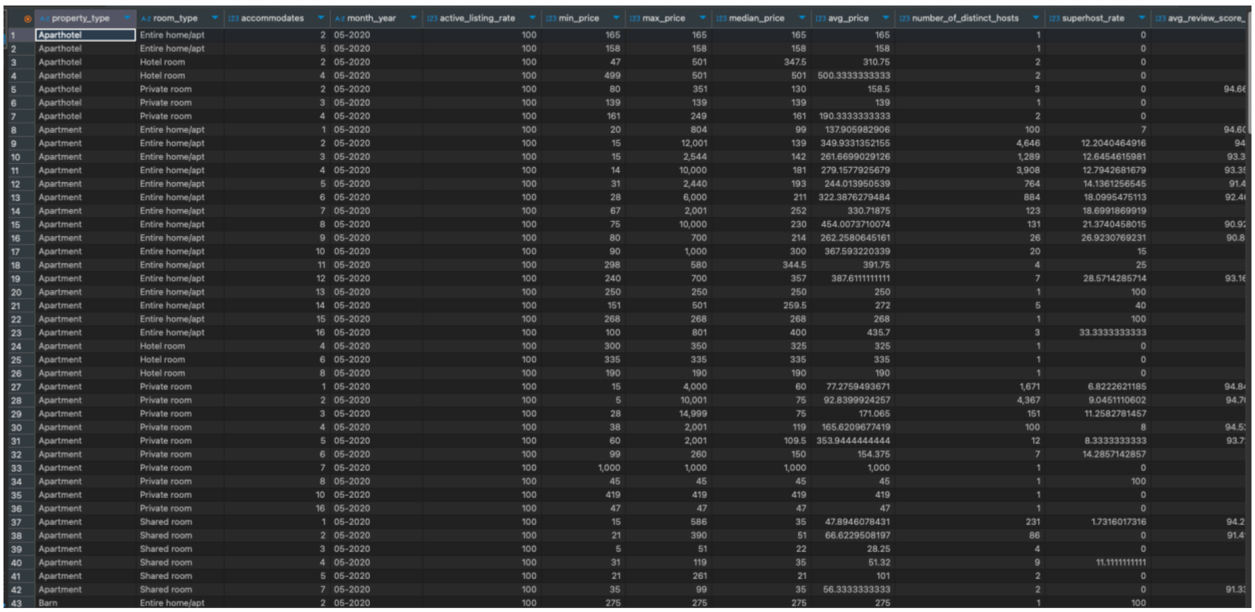


Figure 3.6 Pipeline to create dm_property_type



	property_type	room_type	accommodates	month_year	active_listing_rate	min_price	max_price	median_price	avg_price	number_of_distinct_hosts	superhost_rate	avg_review_score_
1	Aparthotel	Entire home/apt	2	05-2020	100	165	165	165	165	1	0	
2	Aparthotel	Entire home/apt	5	05-2020	100	158	158	158	158	1	0	
3	Aparthotel	Hotel room	2	05-2020	100	47	501	347.5	310.75	2	0	
4	Aparthotel	Hotel room	4	05-2020	100	499	501	501	500.3333333333	2	0	
5	Aparthotel	Private room	2	05-2020	100	80	351	130	158.5	3	0	94.66
6	Aparthotel	Private room	3	05-2020	100	139	139	139	139	1	0	
7	Aparthotel	Private room	4	05-2020	100	161	249	161	190.3333333333	2	0	
8	Apartment	Entire home/apt	1	05-2020	100	20	804	99	137.905982906	100	7	94.6
9	Apartment	Entire home/apt	2	05-2020	100	15	12,001	139	349.9331352155	4,646	12.2040464916	94
10	Apartment	Entire home/apt	3	05-2020	100	15	2,544	142	261.6699029126	1,289	12.6454615981	93.3
11	Apartment	Entire home/apt	4	05-2020	100	14	10,000	181	279.1577925679	3,908	12.7942681679	93.3
12	Apartment	Entire home/apt	5	05-2020	100	31	2,440	193	244.013950539	764	14.1361256545	91.4
13	Apartment	Entire home/apt	6	05-2020	100	28	6,000	211	322.3876279484	884	18.0995475113	92.4
14	Apartment	Entire home/apt	7	05-2020	100	67	2,001	252	330.71875	123	18.6991869919	
15	Apartment	Entire home/apt	8	05-2020	100	75	10,000	230	454.0073710074	131	21.3740458015	90.9
16	Apartment	Entire home/apt	9	05-2020	100	80	700	214	262.2580645161	26	26.9230769231	90.8
17	Apartment	Entire home/apt	10	05-2020	100	90	1,000	300	367.593220339	20	15	
18	Apartment	Entire home/apt	11	05-2020	100	298	580	344.5	391.75	4	25	
19	Apartment	Entire home/apt	12	05-2020	100	240	700	357	387.6111111111	7	28.5714285714	93.1
20	Apartment	Entire home/apt	13	05-2020	100	250	250	250	250	1	100	
21	Apartment	Entire home/apt	14	05-2020	100	151	501	259.5	272	5	40	
22	Apartment	Entire home/apt	15	05-2020	100	268	268	268	268	1	100	
23	Apartment	Entire home/apt	16	05-2020	100	100	801	400	435.7	3	33.3333333333	
24	Apartment	Hotel room	4	05-2020	100	300	350	325	325	1	0	
25	Apartment	Hotel room	6	05-2020	100	335	335	335	335	1	0	
26	Apartment	Hotel room	8	05-2020	100	190	190	190	190	1	0	
27	Apartment	Private room	1	05-2020	100	15	4,000	60	77.2759493671	1,671	6.8222621185	94.8
28	Apartment	Private room	2	05-2020	100	5	10,001	75	92.8399924257	4,367	9.0451110602	94.7
29	Apartment	Private room	3	05-2020	100	28	14,999	75	171.065	151	11.2582781457	
30	Apartment	Private room	4	05-2020	100	38	2,001	119	165.6209677419	100	8	94.5
31	Apartment	Private room	5	05-2020	100	60	2,001	109.5	353.9444444444	12	8.3333333333	93.7
32	Apartment	Private room	6	05-2020	100	99	260	150	154.375	7	14.2857142857	
33	Apartment	Private room	7	05-2020	100	1,000	1,000	1,000	1,000	1	0	
34	Apartment	Private room	8	05-2020	100	45	45	45	45	1	100	
35	Apartment	Private room	10	05-2020	100	419	419	419	419	1	0	
36	Apartment	Private room	16	05-2020	100	47	47	47	47	1	0	
37	Apartment	Shared room	1	05-2020	100	15	586	35	47.8946078431	231	1.7316017316	94.2
38	Apartment	Shared room	2	05-2020	100	21	390	51	66.6229508197	86	0	91.4
39	Apartment	Shared room	3	05-2020	100	5	51	22	28.25	4	0	
40	Apartment	Shared room	4	05-2020	100	31	119	35	51.32	9	11.1111111111	
41	Apartment	Shared room	5	05-2020	100	21	261	21	101	2	0	
42	Apartment	Shared room	7	05-2020	100	35	99	35	56.3333333333	2	0	91.3
43	Barn	Entire home/apt	2	05-2020	100	275	275	275	275	1	100	

Figure 3.7 view dm_property_type

3.3.3.3 dm_host_neighbourhood

This view provides data per host neighbourhood lga (derived from transforming host neighbourhood to the corresponding LGA) and month/year with the following metrics:

- Number of distinct host
- Estimated Revenue

- Estimated Revenue per host (distinct)

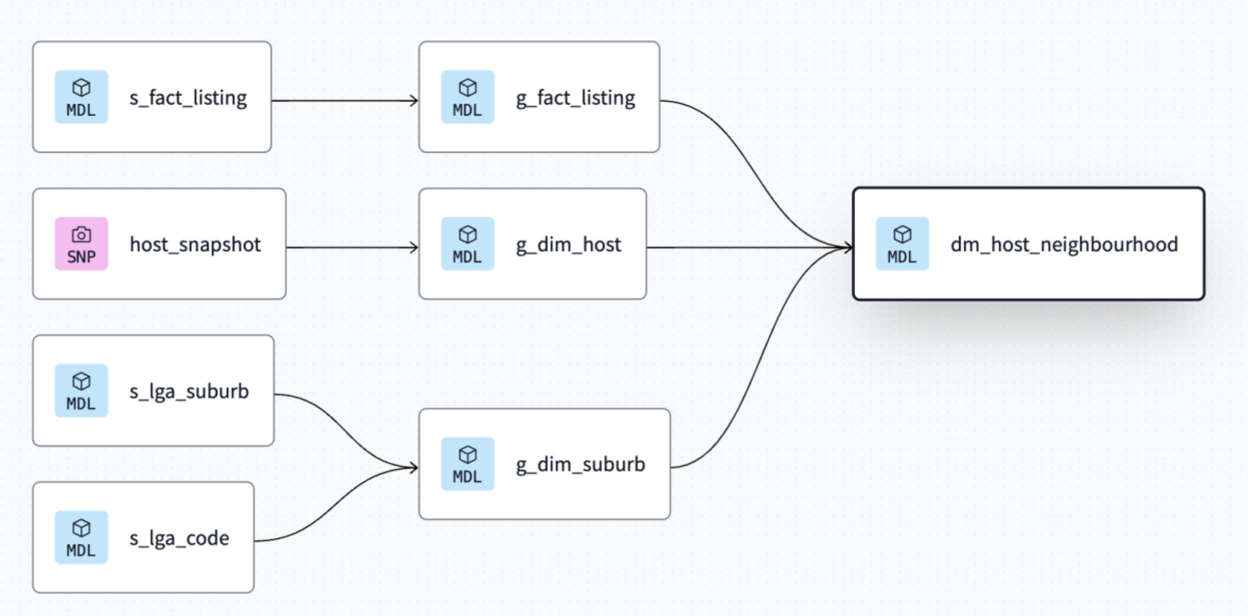


Figure 3.7 Pipeline to create dm_host_neighbourhood

	host_neighbourhood_lga	month_year	number_of_distinct_hosts	estimated_revenue	estimated_revenue_per_host
1	Armidale Regional	05-2020	1	1,800	1,800
2	Bathurst Regional	05-2020	10	188,177	18,817.7
3	Bayside	05-2020	881	6,331,557	7,186.7843359818
4	Burwood	05-2020	169	702,144	4,154.6982248521
5	Canada Bay	05-2020	204	1,672,186	8,196.9901960784
6	Canterbury-Bankstown	05-2020	280	931,121	3,325.4321428571
7	Cumberland	05-2020	4	162,655	40,663.75
8	Dungog	05-2020	49	176,442	3,600.8571428571
9	Georges River	05-2020	132	822,362	6,230.0151515152
10	Hawkesbury	05-2020	5	10,470	2,094
11	Hunters Hill	05-2020	65	337,270	5,188.7692307692
12	Inner West	05-2020	1,676	25,023,625	14,930.5638424821
13	Lake Macquarie	05-2020	1	6,522	6,522
14	Lane Cove	05-2020	223	1,540,483	6,907.9955156951
15	Mosman	05-2020	267	2,883,909	10,801.1573033708
16	Northern Beaches	05-2020	1,347	24,929,176	18,507.1833704529
17	North Sydney	05-2020	724	4,416,331	6,099.9046961326
18	Overseas	05-2020	91	6,682,163	73,430.3626373626
19	Parramatta	05-2020	63	126,730	2,011.5873015873
20	Randwick	05-2020	1,798	32,482,131	18,065.7013348165
21	Ryde	05-2020	17	46,868	2,756.9411764706
22	Singleton	05-2020	72	251,935	3,499.0972222222
23	Strathfield	05-2020	58	344,802	5,944.8620689655
24	Sutherland Shire	05-2020	121	581,189	4,803.2148760331
25	Sydney	05-2020	4,317	34,274,828	7,939.5015056752
26	Unknown	05-2020	10,674	49,583,646	4,645.2731871838
27	Upper Hunter Shire	05-2020	118	700,950	5,940.2542372881
28	Waverley	05-2020	3,096	35,115,381	11,342.1773255814
29	Willoughby	05-2020	222	2,491,680	11,223.7837837838
30	Woollahra	05-2020	585	7,677,108	13,123.2615384615

Figure 3.8 view dm_host_neighbourhood

4. Part 3 – End to end orchestration

Corresponding file: dag1_3.py and snapshot_update.sql

In this section, I built and trained a baseline model as well as two machine learning models: Linear Regression and Decision Tree. The steps followed are as outlined below:

Step 1: In updating the Airflow pipeline, I started by modifying the existing DAG to add a new task that initiates a dbt job. This integration ensures that dbt models are executed as part of the DAG run, allowing the data to flow seamlessly through the various layers of the data warehouse and transforming it as needed.

Step 2: I extended the DAG to load the remaining Airbnb datasets one month at a time, in chronological order. Processing each month's data sequentially maintains the correct order and ensures data integrity throughout the pipeline.

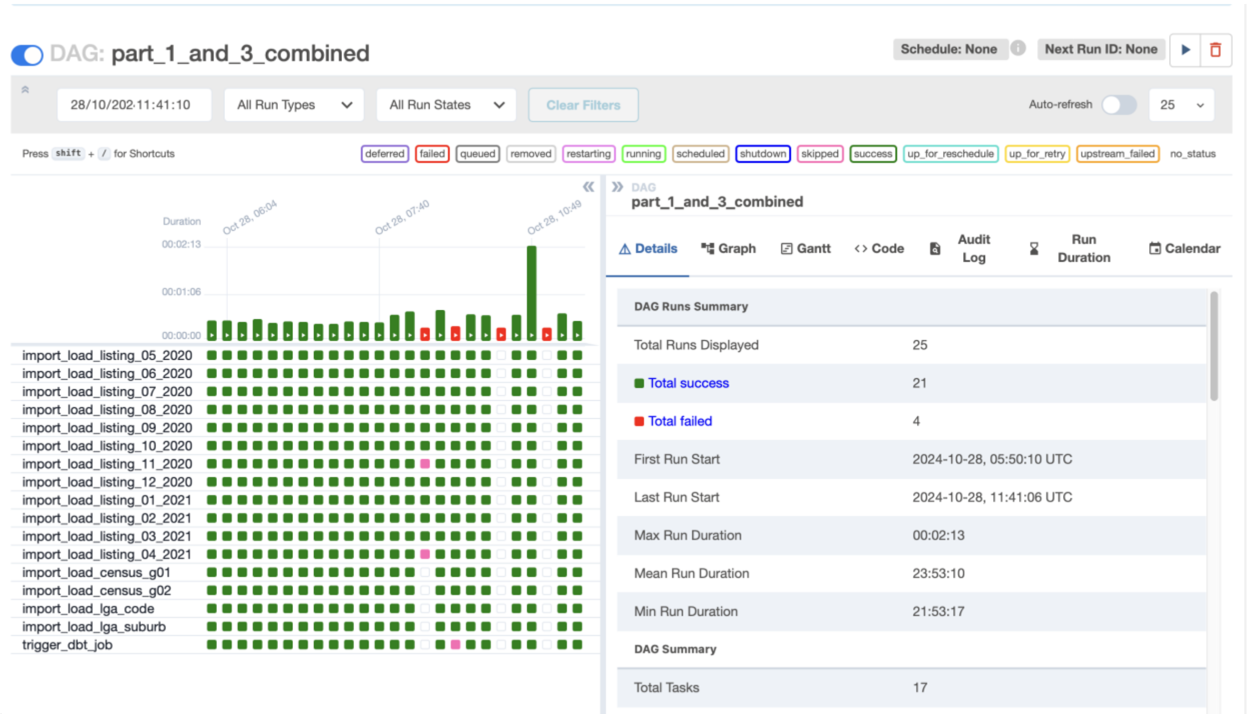


Figure 4.1 Load new data and trigger dbt job

Step 3: After triggering the DAG in Airflow, I proceed by running the SQL script in "snapshot_update.sql" to ensure the snapshots are updated with the latest data. This step is crucial as it captures any changes or updates in the existing data before moving on to load the next month's dataset. By updating the snapshot first, I ensure that each batch of data loaded sequentially remains consistent, accurately reflecting any historical changes, and maintaining data integrity as we progress through each month's records.

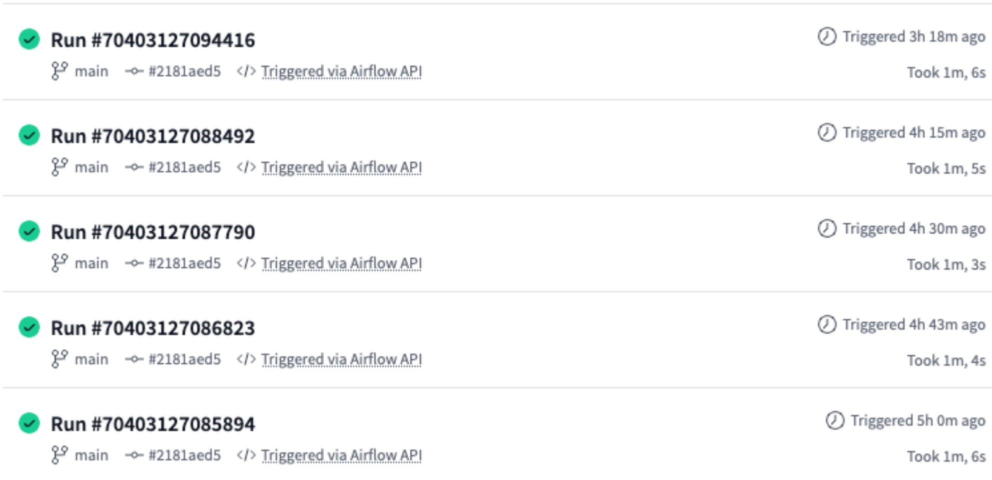


Figure 4.2 Trigger dbt job from Airflow

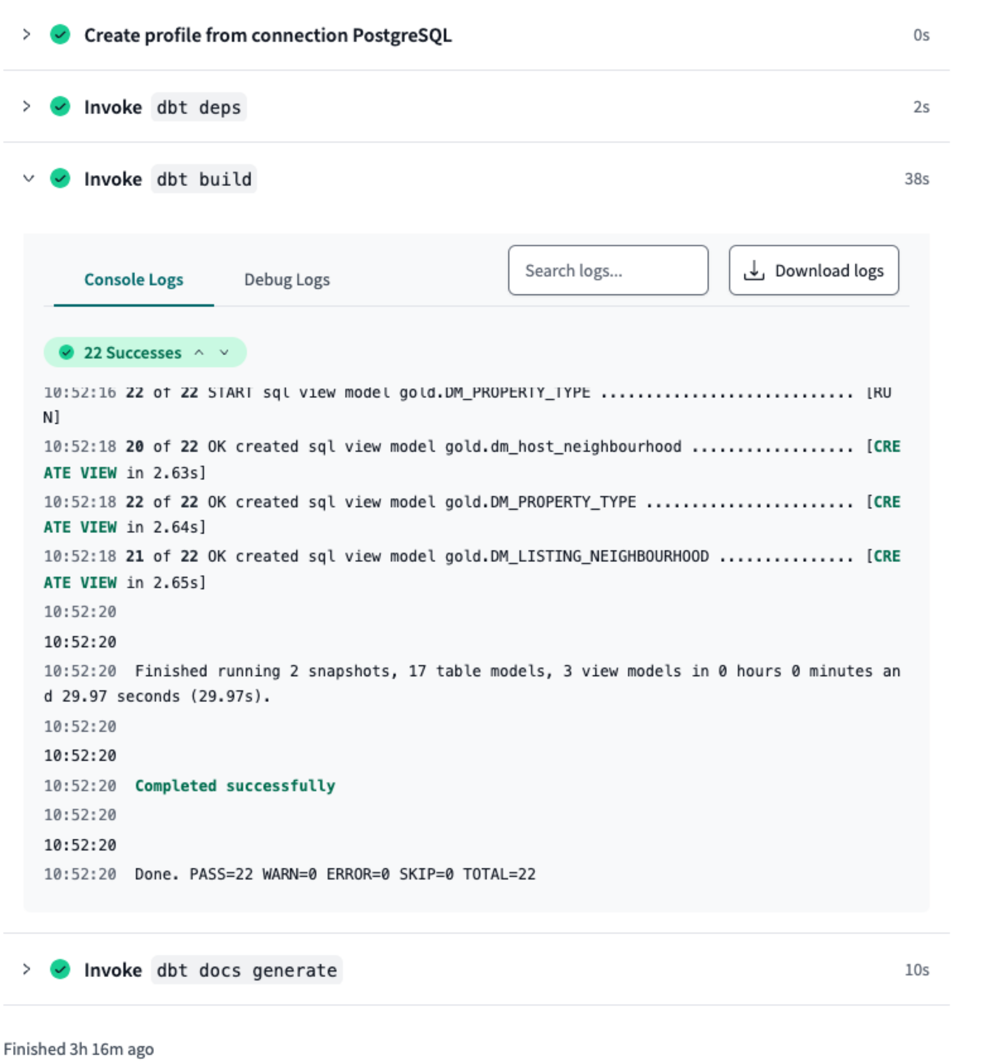
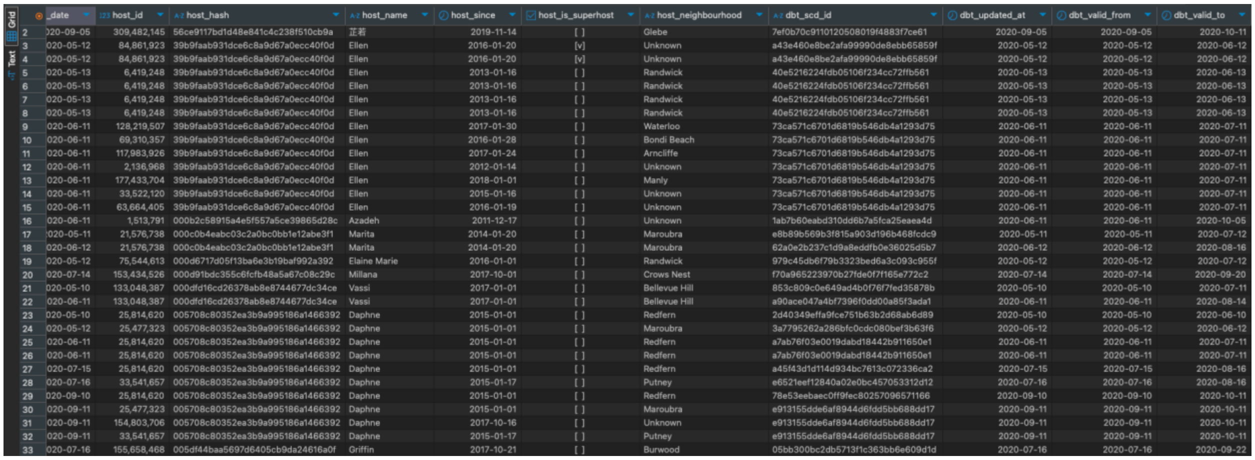
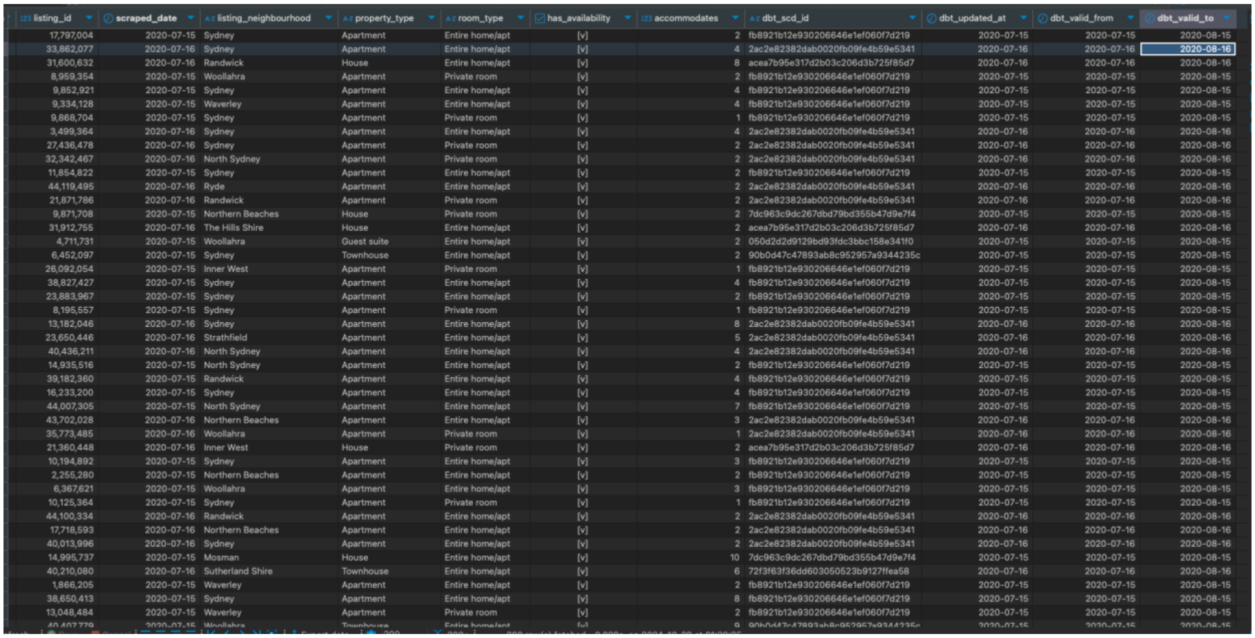


Figure 4.3 Trigger dbt job



	_date	host_id	host_hash	host_name	host_since	host_is_superhost	host_neighbourhood	dbt_scd_id	dbt_updated_at	dbt_valid_from	dbt_valid_to
2	020-09-05	309,482,145	56ce9117bd1d48e841c4c238f510cb9a	芷若	2019-11-14	[ ]	Glebe	7ef0b70c9110120508019f4883f7ce61	2020-09-05	2020-09-05	2020-10-11
3	020-05-12	84,861,923	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2016-01-20	[v]	Unknown	a43e460e8be2afa99990de8ebb65859f	2020-05-12	2020-05-12	2020-06-12
4	020-05-12	84,861,923	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2016-01-20	[v]	Unknown	a43e460e8be2afa99990de8ebb65859f	2020-05-12	2020-05-12	2020-06-12
5	020-05-13	6,419,248	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2013-01-16	[ ]	Randwick	40e5216224fdb05106f234cc72ffb561	2020-05-13	2020-05-13	2020-06-13
6	020-05-13	6,419,248	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2013-01-16	[ ]	Randwick	40e5216224fdb05106f234cc72ffb561	2020-05-13	2020-05-13	2020-06-13
7	020-05-13	6,419,248	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2013-01-16	[ ]	Randwick	40e5216224fdb05106f234cc72ffb561	2020-05-13	2020-05-13	2020-06-13
8	020-05-13	6,419,248	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2013-01-16	[ ]	Randwick	40e5216224fdb05106f234cc72ffb561	2020-05-13	2020-05-13	2020-06-13
9	020-06-11	128,219,507	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2017-01-30	[ ]	Waterloo	73ca571c6701d6819b546db4a1293d75	2020-06-11	2020-06-11	2020-07-11
10	020-06-11	69,310,357	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2016-01-28	[ ]	Bondi Beach	73ca571c6701d6819b546db4a1293d75	2020-06-11	2020-06-11	2020-07-11
11	020-06-11	117,983,926	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2017-01-24	[ ]	Arncliffe	73ca571c6701d6819b546db4a1293d75	2020-06-11	2020-06-11	2020-07-11
12	020-06-11	2,136,968	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2012-01-14	[ ]	Unknown	73ca571c6701d6819b546db4a1293d75	2020-06-11	2020-06-11	2020-07-11
13	020-06-11	177,433,704	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2018-01-01	[ ]	Manly	73ca571c6701d6819b546db4a1293d75	2020-06-11	2020-06-11	2020-07-11
14	020-06-11	33,522,120	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2015-01-16	[ ]	Unknown	73ca571c6701d6819b546db4a1293d75	2020-06-11	2020-06-11	2020-07-11
15	020-06-11	63,664,405	39b9faab931dce6c8a9d67a0ecc40f0d	Ellen	2016-01-19	[ ]	Unknown	73ca571c6701d6819b546db4a1293d75	2020-06-11	2020-06-11	2020-07-11
16	020-06-11	1,513,791	000b2c58915a4e5f557a5ce39865d28c	Azadeh	2011-12-17	[ ]	Unknown	1ab7b60eabd310dd6b7a5fca25eaea4d	2020-06-11	2020-06-11	2020-10-05
17	020-05-11	21,576,738	000c0b4eabc03c2a0bc0bb1e12abe3f1	Marita	2014-01-20	[ ]	Maroubra	e8b89b569b3f815a903d196b468fcdc9	2020-05-11	2020-05-11	2020-07-12
18	020-06-12	21,576,738	000c0b4eabc03c2a0bc0bb1e12abe3f1	Marita	2014-01-20	[ ]	Maroubra	62a0e2b237c1d9a8eddfb0e36025d5b7	2020-06-12	2020-06-12	2020-08-16
19	020-05-12	75,544,613	000d6717d05f13ba6e3b19baf992a392	Elaine Marie	2016-01-01	[ ]	Randwick	979c45db6f79b3323bed6a3c093c955f	2020-05-12	2020-05-12	2020-07-12
20	020-07-14	153,434,526	000d91bdc355c6fcfb48a5a67c08c29c	Millana	2017-10-01	[ ]	Crows Nest	f70a965223970b27fde0f7f165e772c2	2020-07-14	2020-07-14	2020-09-20
21	020-05-10	133,048,387	000dfd16cd26378ab8e8744677dc34ce	Vassi	2017-01-01	[ ]	Bellevue Hill	853c809c0e649ad4b0f76f7fed35878b	2020-05-10	2020-05-10	2020-07-11
22	020-06-11	133,048,387	000dfd16cd26378ab8e8744677dc34ce	Vassi	2017-01-01	[ ]	Bellevue Hill	a90ace047a4bf7396f0dd00a85f3ada1	2020-06-11	2020-06-11	2020-08-14
23	020-05-10	25,814,620	005708c80352ea3b9a995186a1466392	Daphne	2015-01-01	[ ]	Redfern	2d40349effa9fce751b63b2d68ab6d89	2020-05-10	2020-05-10	2020-06-10
24	020-05-12	25,477,323	005708c80352ea3b9a995186a1466392	Daphne	2015-01-01	[ ]	Maroubra	3a7795262a286bfc0cdc080bef3b63f6	2020-05-12	2020-05-12	2020-06-12
25	020-06-11	25,814,620	005708c80352ea3b9a995186a1466392	Daphne	2015-01-01	[ ]	Redfern	a7ab76f03e0019dabd18442b911650e1	2020-06-11	2020-06-11	2020-07-11
26	020-06-11	25,814,620	005708c80352ea3b9a995186a1466392	Daphne	2015-01-01	[ ]	Redfern	a7ab76f03e0019dabd18442b911650e1	2020-06-11	2020-06-11	2020-07-11
27	020-07-15	25,814,620	005708c80352ea3b9a995186a1466392	Daphne	2015-01-01	[ ]	Redfern	a45f43d1d114d934bc7613c072336ca2	2020-07-15	2020-07-15	2020-08-16
28	020-07-16	33,541,657	005708c80352ea3b9a995186a1466392	Daphne	2015-01-17	[ ]	Putney	e6521eef12840a02e0bc457053312d12	2020-07-16	2020-07-16	2020-08-16
29	020-09-10	25,814,620	005708c80352ea3b9a995186a1466392	Daphne	2015-01-01	[ ]	Redfern	78e53eebaec0ff9fec80257096571166	2020-09-10	2020-09-10	2020-10-11
30	020-09-11	25,477,323	005708c80352ea3b9a995186a1466392	Daphne	2015-01-01	[ ]	Maroubra	e913155dde6af8944d6fdd5bb688dd17	2020-09-11	2020-09-11	2020-10-11
31	020-09-11	154,803,706	005708c80352ea3b9a995186a1466392	Daphne	2017-10-16	[ ]	Unknown	e913155dde6af8944d6fdd5bb688dd17	2020-09-11	2020-09-11	2020-10-11
32	020-09-11	33,541,657	005708c80352ea3b9a995186a1466392	Daphne	2015-01-17	[ ]	Putney	e913155dde6af8944d6fdd5bb688dd17	2020-09-11	2020-09-11	2020-10-11
33	020-07-16	155,658,468	005df44baa5697d6405cb9da24616a0f	Griffin	2017-10-21	[ ]	Burwood	05bb300bc2db5713f1c363bb6e609d1d	2020-07-16	2020-07-16	2020-09-22

Figure 4.4 Update host_snapshot using SCD type 2



listing_id	scraped_date	listing_neighbourhood	property_type	room_type	has_availability	accommodates	dbt_scd_id	dbt_updated_at	dbt_valid_from	dbt_valid_to
17,797,004	2020-07-15	Sydney	Apartment	Entire home/apt	[v]	2	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
33,862,077	2020-07-16	Sydney	Apartment	Entire home/apt	[v]	4	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
31,600,632	2020-07-16	Randwick	House	Entire home/apt	[v]	8	acea7b95e317d2b03c206d3b725f85d7	2020-07-16	2020-07-16	2020-08-16
8,959,354	2020-07-15	Woollahra	Apartment	Private room	[v]	2	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
9,852,921	2020-07-15	Sydney	Apartment	Entire home/apt	[v]	4	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
9,334,128	2020-07-15	Waverley	Apartment	Entire home/apt	[v]	4	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
9,868,704	2020-07-15	Sydney	Apartment	Private room	[v]	1	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
3,499,364	2020-07-16	Sydney	Apartment	Entire home/apt	[v]	4	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
27,436,478	2020-07-16	Sydney	Apartment	Private room	[v]	2	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
32,342,467	2020-07-16	North Sydney	Apartment	Private room	[v]	2	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
11,854,822	2020-07-15	Sydney	Apartment	Entire home/apt	[v]	2	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
44,119,495	2020-07-16	Ryde	Apartment	Entire home/apt	[v]	2	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
21,871,786	2020-07-16	Randwick	Apartment	Private room	[v]	2	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
9,871,708	2020-07-15	Northern Beaches	House	Private room	[v]	2	7dc963c9dc267dbd79bd355b47d9e7f4	2020-07-15	2020-07-15	2020-08-15
31,912,755	2020-07-16	The Hills Shire	House	Entire home/apt	[v]	2	acea7b95e317d2b03c206d3b725f85d7	2020-07-16	2020-07-16	2020-08-16
4,711,731	2020-07-15	Woollahra	Guest suite	Entire home/apt	[v]	2	050d2d2d9129bd93fdc3bbc158e341f0	2020-07-15	2020-07-15	2020-08-15
6,452,097	2020-07-15	Sydney	Townhouse	Entire home/apt	[v]	2	90b0d47c47893ab8c952957a9344235c	2020-07-15	2020-07-15	2020-08-15
26,092,054	2020-07-15	Inner West	Apartment	Private room	[v]	1	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
38,827,427	2020-07-15	Sydney	Apartment	Entire home/apt	[v]	4	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
23,883,967	2020-07-15	Sydney	Apartment	Entire home/apt	[v]	2	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
8,195,557	2020-07-15	Sydney	Apartment	Private room	[v]	1	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
13,182,046	2020-07-16	Sydney	Apartment	Entire home/apt	[v]	8	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
23,650,446	2020-07-16	Strathfield	Apartment	Entire home/apt	[v]	5	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
40,436,211	2020-07-16	North Sydney	Apartment	Entire home/apt	[v]	4	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
14,935,516	2020-07-15	North Sydney	Apartment	Entire home/apt	[v]	2	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
39,182,360	2020-07-15	Randwick	Apartment	Entire home/apt	[v]	4	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
16,233,200	2020-07-15	Sydney	Apartment	Entire home/apt	[v]	4	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
44,007,305	2020-07-15	North Sydney	Apartment	Entire home/apt	[v]	7	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
43,702,028	2020-07-16	Northern Beaches	Apartment	Entire home/apt	[v]	3	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
35,773,485	2020-07-16	Woollahra	Apartment	Private room	[v]	1	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
21,360,448	2020-07-16	Inner West	House	Private room	[v]	2	acea7b95e317d2b03c206d3b725f85d7	2020-07-16	2020-07-16	2020-08-16
10,194,892	2020-07-15	Sydney	Apartment	Entire home/apt	[v]	3	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
2,255,280	2020-07-15	Northern Beaches	Apartment	Entire home/apt	[v]	2	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
6,367,621	2020-07-15	Woollahra	Apartment	Entire home/apt	[v]	3	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
10,125,364	2020-07-15	Sydney	Apartment	Private room	[v]	1	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
44,100,334	2020-07-16	Randwick	Apartment	Entire home/apt	[v]	2	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
17,718,593	2020-07-16	Northern Beaches	Apartment	Entire home/apt	[v]	2	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
40,013,996	2020-07-16	Sydney	Apartment	Entire home/apt	[v]	2	2ac2e82382dab0020fb09fe4b59e5341	2020-07-16	2020-07-16	2020-08-16
14,995,737	2020-07-15	Mosman	House	Entire home/apt	[v]	10	7dc963c9dc267dbd79bd355b47d9e7f4	2020-07-15	2020-07-15	2020-08-15
40,210,080	2020-07-16	Sutherland Shire	Townhouse	Entire home/apt	[v]	6	72f3f63f36dd603050523b9127ffea58	2020-07-16	2020-07-16	2020-08-16
1,866,205	2020-07-15	Waverley	Apartment	Entire home/apt	[v]	2	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
38,650,413	2020-07-15	Sydney	Apartment	Entire home/apt	[v]	8	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
13,048,484	2020-07-15	Waverley	Apartment	Private room	[v]	2	fb8921b12e930206646e1ef060f7d219	2020-07-15	2020-07-15	2020-08-15
40,407,779	2020-07-15	Woollahra	Townhouse	Entire home/apt	[v]	9	90b0d47c47893ab8c952957a9344235c	2020-07-15	2020-07-15	2020-08-15

Figure 4.5 Update property_snapshot using SCD type 2

5. Part 4 – Ad-hoc analysis

I wasn't able to complete this part as I encountered challenges in loading all the data for Part 3. I faced significant delays in Part 2, particularly with managing snapshots, and struggled to keep up with the workload. In the end, I managed to get the pipeline running for Part 2, but I didn't have enough time to load all the data for Part 3. I was only able to load data from May 2020 to July 2020. I feel disappointed with myself for not meeting my expectations in this task.

6. Conclusion

In conclusion, this project showcases the integration of Apache Airflow and dbt Cloud to create a robust ELT pipeline for Airbnb and Census data, following the Medallion architecture. Despite challenges with pipeline and snapshot management, significant progress was made in structuring the data across Bronze, Silver, and Gold layers. These layers support a well-defined data mart for comprehensive analytical insights, aiding business decision-making with key metrics like active listing rates and estimated revenues. Although some data remains to be processed, the established framework provides a strong foundation for further analysis, ensuring scalability and data integrity for future enhancements.