Building ELT data pipelines with Airflow:

Objective:

The objective of this project is to build production-ready ELT data pipelines using Apache Airflow. The pipeline processes and transforms Airbnb and Census data for Sydney, load it into PostgreSQL-based data warehouse following the **Medallion Architecture** (Bronze, Silver, Gold). The final goal is to create a data mart that supports analytical insights. This also includes ad-hoc analyses to address key business questions.

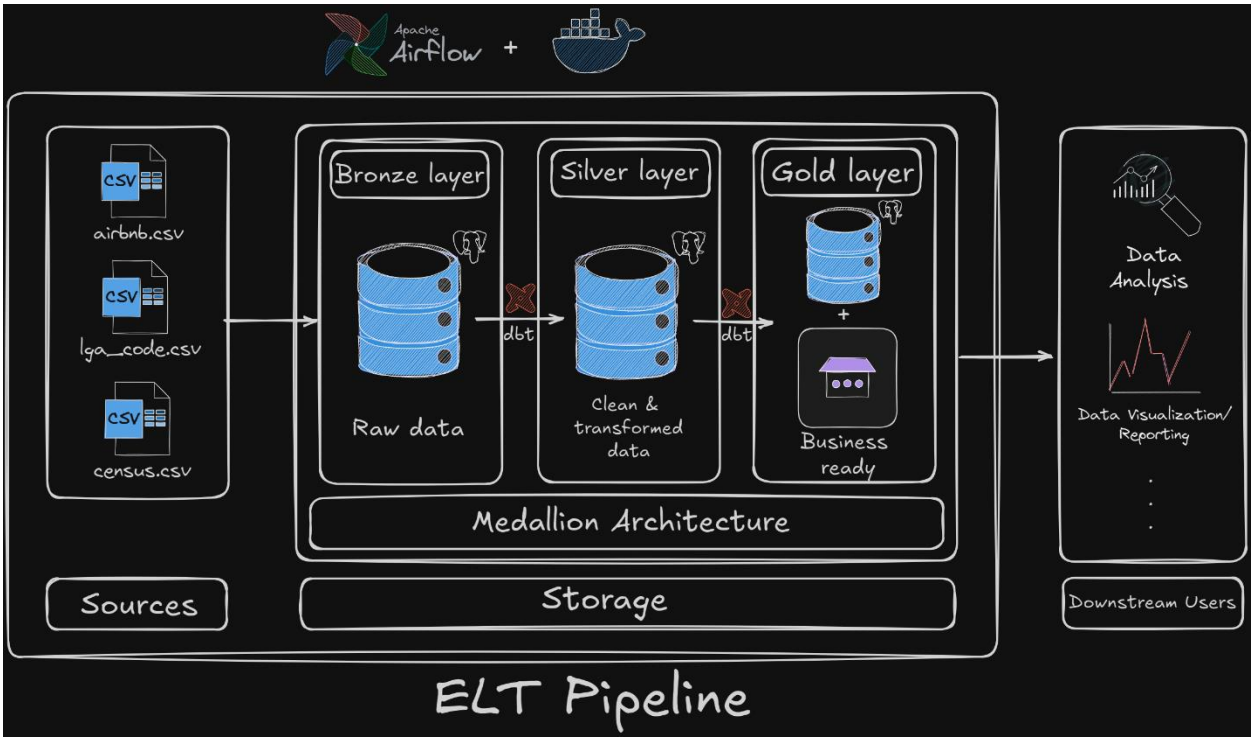
Scope:

1. Airbnb dataset: Airbnb is an online-based marketing company that connects people looking for accommodates (Airbnb guests) to people looking to rent their properties (Airbnb hosts). This project utilizes Airbnb data for **Sydney**, covering the period from **May 2020 to April 2021**.
2. Census dataset: The Census of population and Housing (Census) is Australia's largest statistical collection undertaken by the Australian Bureau of Statistics (ABS). The aim of Census is to accurately collet data on key characteristics of people of Australia on demographic, social and economics characteristics and the dwellings in which they live.
3. LGA dataset: Local Government Area (LGA) is a specific geographic area administered by a local council. This dataset is used to link Airbnb and Census data geographically for more localized analysis.

Architecture Overview:

The project includes PostgreSQL-based Medallion Architecture that has three-layer Bronze, Silver, and Gold.

- Bronze layer: Contains the raw ingested data from Airbnb, Census and LGA sources.
- Silver layer: Stores cleaned, structured and standardized data for further transformation.
- Gold layer: Includes fact and dimension tables designed using dbt that will be used to answer business key question.



The diagram illustrates the end-to-end data pipeline architecture and processing stages from ingestion to reporting.

Tools & Technologies:

The project includes:

- **Programming Language**: Python – Used for scripting data ingestion and pipeline logic.
- **Data Analytics**: SQL – Used for data exploration, analysis, and querying across different layers of the warehouse.
- **Orchestration**: Apache Airflow – Managed and scheduled ETL/ELT workflows to ensure reliable data movement.
- **Transformation**: dbt (Data Build Tool) – Used to build modular, testable SQL-based models in the Silver and Gold layers.
- **Schema Design & Exploration**: DBeaver – Utilized for managing database schemas, running queries, and visually exploring data in PostgreSQL.

Data Ingestion:

The Airbnb dataset was sourced from airbnb.com, where monthly CSV files were downloaded for the period of **May 2020 to April 2021**. Each file contains comprehensive information about Airbnb listings, including details about hosts, properties, availability, pricing, and estimated activity metrics.

These files were ingested into the **Bronze layer** of the data warehouse as-is (raw format). The ingestion process involved:

- Reading each monthly CSV file using Python's pandas.
- Loading the data into a **PostgreSQL** table under the **bronze.airbnb** schema.

The ingestion was automated using **Apache Airflow**, which schedules and manages the monthly ingestion DAGs (Directed Acyclic Graphs).

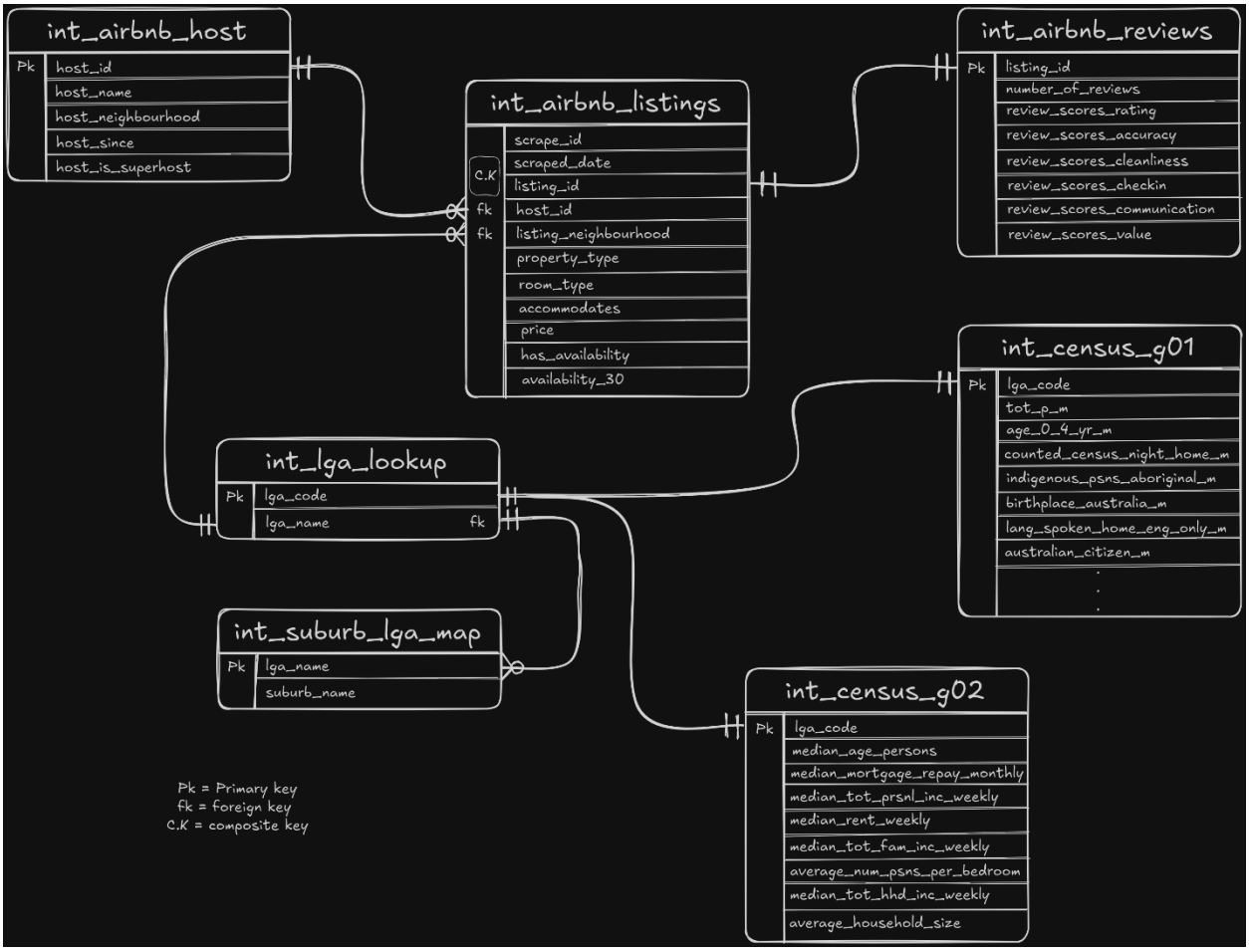
Medallion Architecture (Silver Layer):

The **Silver Layer** is the second stage in the Medallion Architecture. It builds upon the raw ingested data from the **Bronze Layer** and focuses on creating **structured, cleaned, and transformed datasets** ready for advanced analytics.

Key tasks performed in the Silver Layer include:

- Cleaning null values and handling missing or inconsistent data.
- Standardizing column names and enforcing naming conventions.
- Joining related tables on common keys to prepare for snapshotting and dimensional modeling.
- Implementing **Slowly Changing Dimension Type 2 (SCD2)** logic using dbt snapshots:
    - SCD2 allows tracking historical changes in dimension data over time.
    - Instead of updating existing records, new rows are inserted with valid time ranges (valid_from, valid_to), preserving a full history of changes.

This layer ensures that the data is accurate, consistent, and ready to be modeled in the Gold Layer for analytical purposes.



Gold Layer:

The Gold Layer represents the final stage of the Medallion Architecture. It contains fully transformed, business-ready data organized into a star schema to support reporting, dashboards, and ad-hoc business analyses. It contains a fact table and dimension tables.

Fact table: Stores quantitative metrics such as price, revenue, reviews etc. It contains foreign keys that link to dimension tables.

Dimension tables: Contains descriptive attributes that provide context for fact table.

Reference tables: Two Census tables are included to enrich insights with socio-economic context at the LGA.

This layer is used to answer a set of predefined business questions – providing strategic insights for decision-making.

Orchestration:

To automate and manage the end-to-end workflow, Apache Airflow was used as the orchestration tool. Airflow enables scheduling, monitoring, and dependency management of the entire ELT process.

Two Directed Acyclic Graphs (DAGs) were created:

1. **extracting_datasets_dag**:
   Responsible for the initial ingestion of data. It fetches and loads the raw Airbnb and Census datasets into the Airflow bucket.

2. **ingestion_datasets_dag**:
   This DAG executes the transformation logic that builds the Silver and Gold layers using dbt. It ensures the data flows through all transformation steps, including cleaning, normalization, and fact/dimension modeling.

By using Airflow, the pipelines can be triggered automatically on a **scheduled basis**, ensuring that data remains fresh and up to date with minimal manual intervention.

Data Mart:

The Data Mart is the final layer of the pipeline where key business questions are answered using structured and optimized data. It is built on top of the Gold Layer, which contains well-modeled fact and dimension tables.

Business logic is encapsulated in SQL views created using dbt, which are materialized to improve performance and simplify analysis.

Slowly Changing Dimensions (SCD2) are used in dimension tables to **track historical changes** over time. This is useful for maintaining historical accuracy when attributes like host neighbourhood or listing details change.

The Data Mart forms the analytical backbone of the project, enabling stakeholders to derive insights and make data-driven decisions.

Ad-hoc Questions:

All ad-hoc business questions were addressed using SQL queries defined in the ad-hoc.sql file located inside the models folder of the dbt project. These queries use data from the Gold Layer (fact and dimension tables) and reflect insights derived from the Airbnb and Census datasets.

Path: **airbnb_warehouse/models/ad-hoc.sql**

The supporting screenshots of query outputs are attached below to validate and visualize the answers.

1. What are the demographic differences (e.g., age group distribution, household size) between the top 3 performing and lowest 3 performing LGAs based on estimated revenue per active listing over the last 12 months?

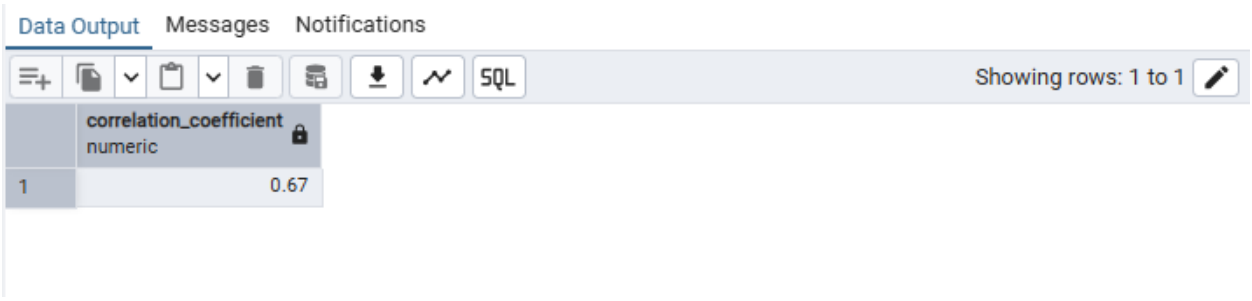

Data Output Messages Notifications

Showing rows: 1 to 1

	diff_revenue_percentage numeric
1	70.58

Output: So Top 3 lga's revenue is 70% more than the revenue of bottom 3 lga's.

2. Is there a correlation between the median age of a neighbourhood (from Census data) and the revenue generated per active listing in that neighbourhood?

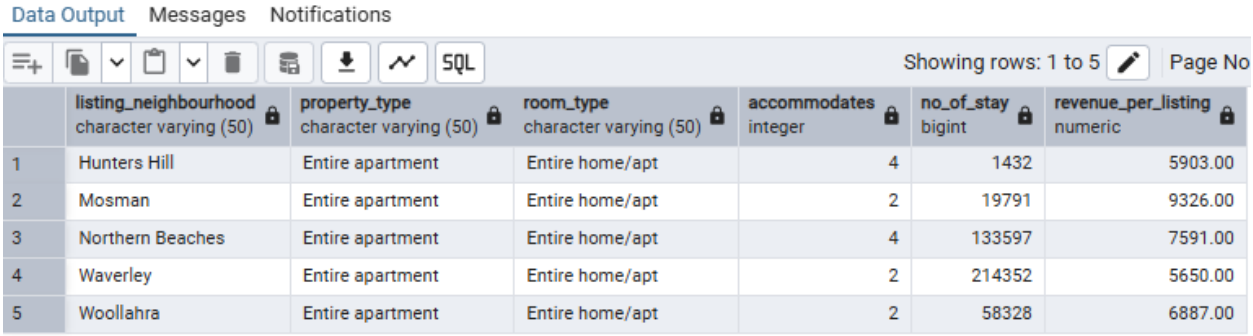

Data Output Messages Notifications

Showing rows: 1 to 1

	correlation_coefficient numeric
1	0.67

Output: The correlation is 0.67 which means strong.

3. 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?

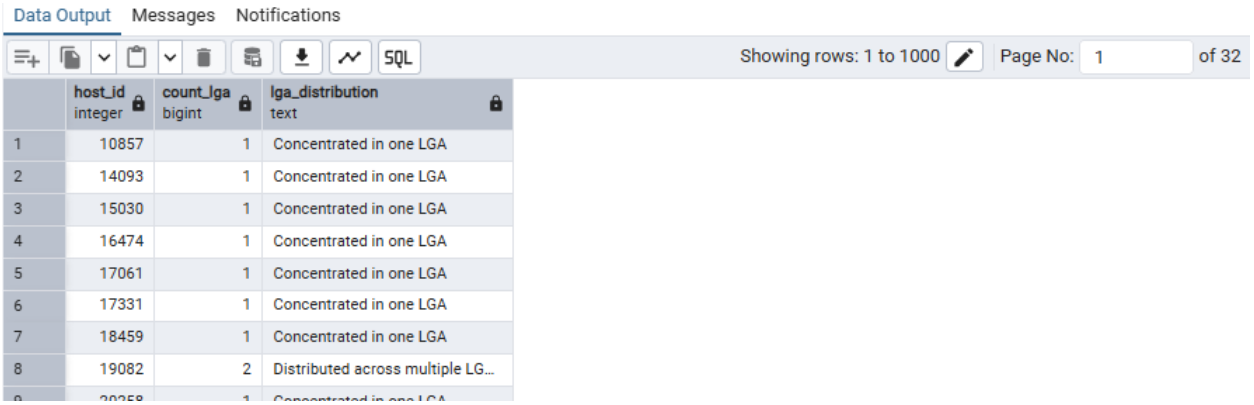

Data Output Messages Notifications

Showing rows: 1 to 5 Page No

	listing_neighbourhood character varying (50)	property_type character varying (50)	room_type character varying (50)	accommodates integer	no_of_stay bigint	revenue_per_listing numeric
1	Hunters Hill	Entire apartment	Entire home/apt	4	1432	5903.00
2	Mosman	Entire apartment	Entire home/apt	2	19791	9326.00
3	Northern Beaches	Entire apartment	Entire home/apt	4	133597	7591.00
4	Waverley	Entire apartment	Entire home/apt	2	214352	5650.00
5	Woollahra	Entire apartment	Entire home/apt	2	58328	6887.00

Mosman, Northern Beaches, Woollahra, Hunters Hill, and Waverley are the highest-performing LGAs in terms of revenue per listing.

4. For hosts with multiple listings, are their properties concentrated within the same LGA, or are they distributed across different LGAs?

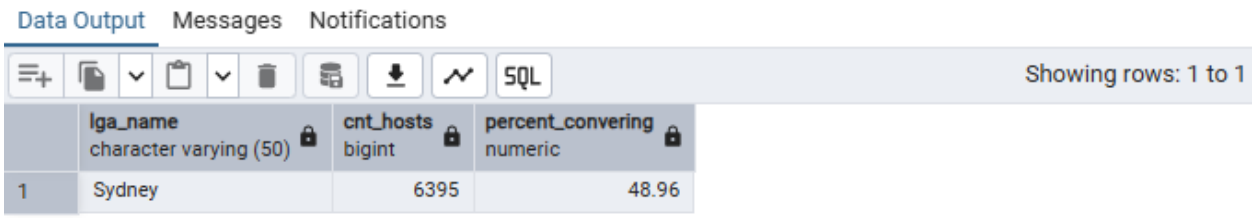


Data Output | Messages | Notifications

Showing rows: 1 to 1000 | Page No: 1 of 32

	host_id integer	count_lga bigint	lga_distribution text
1	10857	1	Concentrated in one LGA
2	14093	1	Concentrated in one LGA
3	15030	1	Concentrated in one LGA
4	16474	1	Concentrated in one LGA
5	17061	1	Concentrated in one LGA
6	17331	1	Concentrated in one LGA
7	18459	1	Concentrated in one LGA
8	19082	2	Distributed across multiple LG...

5. For hosts with a single Airbnb listing, does the estimated revenue over the last 12 months cover the annualized median mortgage repayment in the corresponding LGA? Which LGA has the highest percentage of hosts that can cover it?



Data Output | Messages | Notifications

Showing rows: 1 to 1

	lga_name character varying (50)	cnt_hosts bigint	percent_convering numeric
1	Sydney	6395	48.96

Output: The highest percent of host cover by "Sydney" with hosts counts = 6395 and percent_covering = 48.96