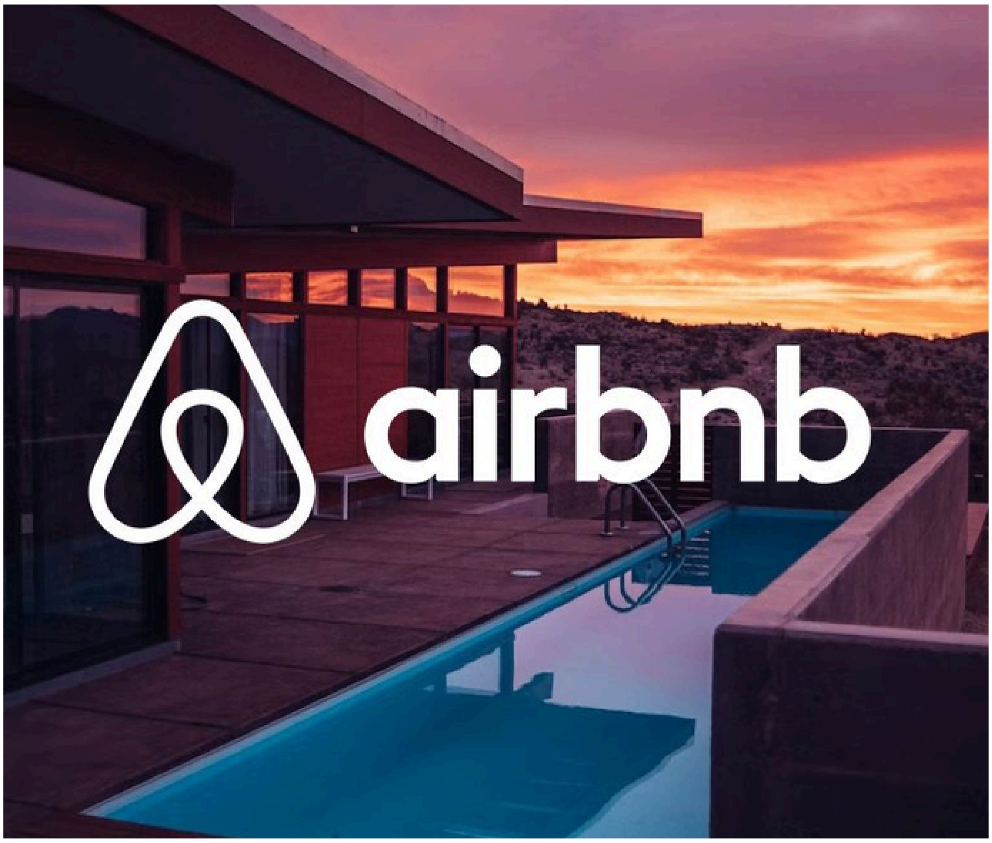


2025

Airbnb ELT Data Pipeline with Apache Airflow

Big Data Engineering | Data Orchestration | Analytics Automation

Executive Summary

This report outlines the development of a production-ready ELT pipeline using Airflow and dbt Cloud, built to process Airbnb and Census data for Sydney. The solution implements a Medallion architecture within a PostgreSQL data warehouse to support scalable data ingestion, transformation, and analytics. Key components include workflow orchestration, layered modelling, and business-focused analysis on revenue and demographic trends.

Main Topics

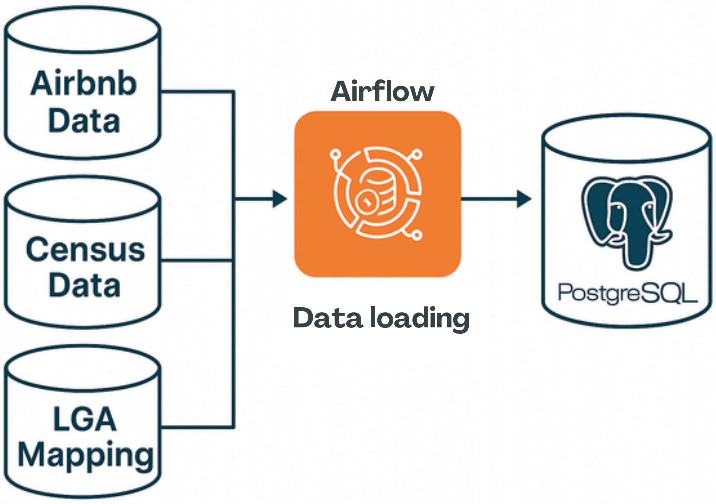
DATA INGESTION & STORAGE

WORKFLOW ORCHESTRATION

ANALYTICS & INSIGHT GENERATION

Airbnb ELT Pipeline

End-to-End Data Engineering for Short-Term Rental Insights



KEY DATA SOURCES

- Airbnb Listings
- May 2020 – April 2021: Includes pricing, host details, availability, and reviews.
- Census 2016 (G01 & G02)
- Covers demographics, income, rent, household size by LGA.
- NSW LGA Mapping
- Enables geographic linking between Airbnb and Census data.

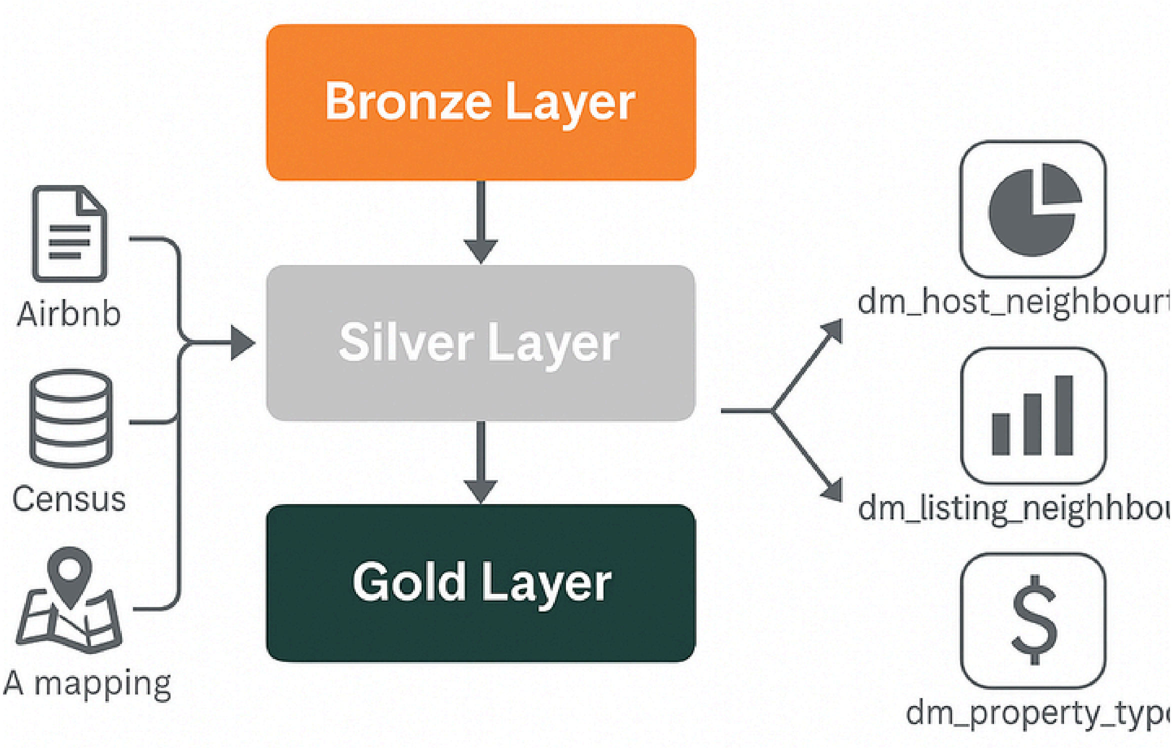
DATA LOADING OVERVIEW

Pipeline Layer: Bronze
Platform: PostgreSQL
Orchestrator: Apache Airflow (airbnb_bronze_ingestion DAG)
Each task automates the ingestion of raw CSV files into staging tables:

- load_raw_facts → Airbnb listings
- load_census_g01 → Demographic data
- load_census_g02 → Household & income data
- load_nsw_lga_code, load_nsw_lga_suburb → LGA structure

Upon completion, files are archived to avoid reprocessing.

Data Warehouse Strategy



End-to-End Orchestration

Airflow + dbt Cloud Integration

Airflow orchestrates the ingestion of monthly Airbnb datasets into the bronze layer, with each file processed sequentially and archived post-load. Once ingestion is complete, the DAG triggers a dbt Cloud job to launch transformations across the silver and gold layers, ensuring smooth end-to-end pipeline execution.

ARCHITECTURE OVERVIEW

Bronze Layer

Raw Airbnb, Census, and LGA data ingested and stored untransformed for traceability.

Silver Layer

Cleansed and modelled data with SCD Type 2 snapshots for historical tracking.

Gold Layer

Well-structured dimension and fact tables for reporting and analytics.

DATA MARTS

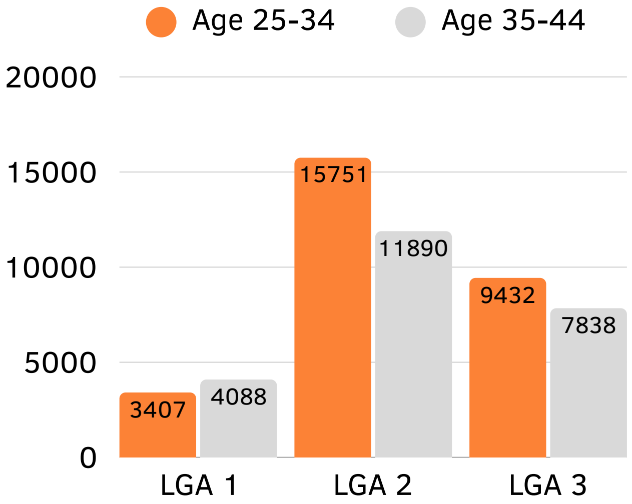
dm_host_neighbourhood: Host performance by suburb and LGA

dm_listing_neighbourhood: Listing activity, pricing, and trends

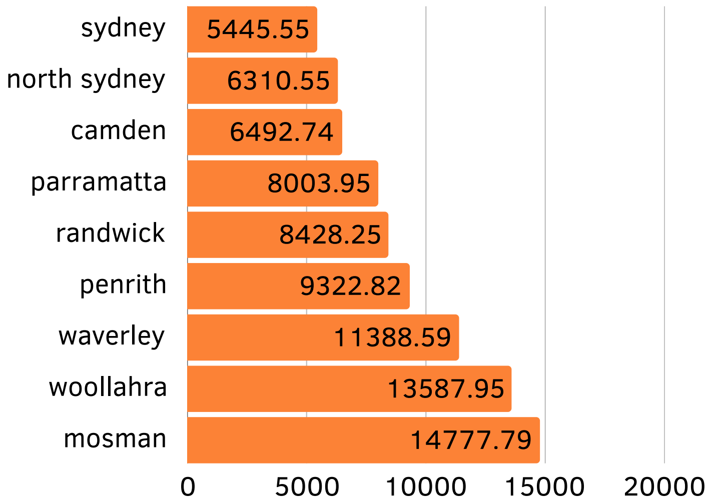
dm_property_type: Revenue by property and room type

Key insights

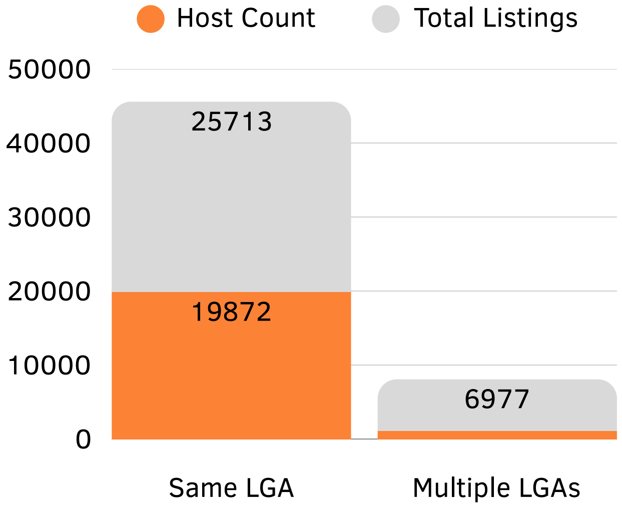
Key Demographics Driving Airbnb Revenue



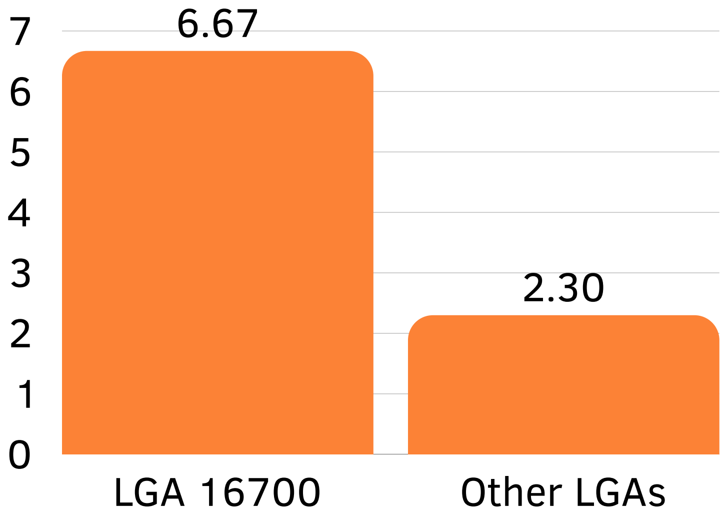
Highest Airbnb Revenue LGAs



Host Distribution by LGA Type



Percentage of Hosts Covering Mortgage by LGA



Airbnb Data Pipeline Report

Play a vital role in Business Partnerships

PROPERLY PREPARED BY:

OLIVER XIAO

This report outlines the design and orchestration of ELT data pipelines built with Apache Airflow, dbt, and PostgreSQL, analysing Airbnb and Census datasets to uncover revenue trends and demographic impacts at the LGA level.

FOR MORE INFORMATION

☎ oliverxiao.r@gmail.com

🌐 www.github.com/oli7794

📍 Sydney, NSW