# BIG DATA ENGINEERING.

ASSIGNMENT 3: Building ELT data pipelines with Airflow.

NAME: AADITYA DESHMUKH.

STUDENT_ID: 14334864.

**INDEX**

## INTRODUCTION

This project embarks on the intricate journey of constructing robust ELT (Extract, Load, Transform) data pipelines, employing Apache Airflow within a cloudbased infrastructure. Central to this endeavor is the utilization of two diverse yet significant datasets: Airbnb listing data for Sydney and the 2016 Census of Population and Housing from the Australian Bureau of Statistics. The Airbnb dataset, a trove of information reflecting the dynamics of shortterm rental markets, offers insights into rental densities, price variations, and hostguest interactions. In contrast, the Census dataset provides a comprehensive snapshot of the Australian populace, crucial for strategic planning and resource allocation. Our task involves not only the meticulous extraction and loading of these datasets into a Postgres environment but also their sophisticated transformation and cleaning to derive meaningful insights. The ultimate goal is to seamlessly integrate this processed data into a meticulously architected data warehouse and a strategically designed data mart on Postgres, leveraging dbt for data transformation and modeling. This integration facilitates indepth analytical explorations, addressing key business questions and uncovering patterns that inform decisionmaking in the realms of accommodation and population dynamics.

## PROJECT SETUP

1. **GCP Composer Setup**
   a. Environment Configuration:
   - Name: bdeenvcc1.
   - Location: australiasoutheast1.
   - VM Disk Size: 30 GB.
   - Google Kubernetes Engine (GKE) Cluster: Located in australiasoutheast1b zone.

   b. Cloud SQL Machine:
   - Type: dbn1standard2.
   - Specifications: 2 vCPUs, 7.5 GB memory.

   c. Worker Nodes:
   - Configuration: 3 nodes.
   - Each with n1standard2 machine type and 30 GB disk size.

   d. Airflow Configuration:
   - Core Setting: enable_xcom_pickling set to True.

   e. Installed PyPI Packages:
   - pandas: For data analysis and manipulation.
   - apacheairflowproviderspostgres: To integrate Apache Airflow with PostgreSQL.

2. **PostgreSQL Instance Setup**

a. Instance Details:
   - ID: postgresql1.
   - Version: PostgreSQL 15.4.
   - Region: australiasoutheast1 (Sydney).

b. Hardware and Storage:
   - vCPUs: 2.
   - Memory: 8 GB.
   - Storage Capacity: 100 GB.
   - Data Cache: Disabled.

c. Networking and Security:
   - Connection Types: Private and Public IPs.
   - Public IP: 34.116.125.98.
   - Authorized Networks for Enhanced Security:
   - Broadway: 103.131.14.133.
   - UTS Building 11: 138.25.4.76.

d. Backup and Recovery:
   - Automated backups.
   - Single zone availability.
   - Pointintime recovery enabled.

3. **dbt Core installation and Connecting to PostgreSQL**

a. Development Environment and Tools:
   - Installed Visual Studio Code (VS Code) and Python 3.
   - Added VS Code extensions: "Python Environment Manager" and "Python".

b. Project Folder and Virtual Environment:
   - Created "dbtdev" folder in VS Code.
   - Established a Python virtual environment in "dbtdev".

c. dbt Core Installation and Configuration:
   - Installed dbt for PostgreSQL using pip install dbtpostgres.
   - Initialized a new dbt project named "bde".

d. Database Connection Setup:
   - Configured profiles.yml for dbt to connect with the PostgreSQL instance.
   - Settings included:
   - Host: 34.116.125.98 (Public IP of PostgreSQL instance).
   - Schema: raw_schema (Initial schema in PostgreSQL for raw data processing).
   - User, password, database name, and thread settings.

e. Version Control Integration:
   - Initialized a repository in VS Code for the dbt project.
   - Pushed the project to a GitHub repository.

f. Operational Workflow:
   - Established a structured workflow using dbt for data transformation.
   - Validated the setup and connectivity with dbt debug.

## DATA SOURCES AND PREPARATIONS

In this project, we analyze two datasets: Airbnb listings in Sydney and Australian Census data. The Airbnb data, from May 2020 to April 2021, reveals insights into the city's rental market, including property prices and guest-host interactions. The Census data provides a detailed view of Australia's demographics and housing. Together, these datasets allow us to examine the relationship between housing trends and population characteristics in a practical setting.

In our datasets, each column plays a crucial role in providing specific insights. To facilitate a comprehensive understanding, we present a detailed description of each column across our various datasets, ensuring clarity in the breadth and depth of the data they encompass.

1. Airbnb Listing Data
    a. Property Columns:
    - LISTING_ID: Unique identifier for the property listing.
    - SCRAPE_ID: Unique identifier for the scraping instance.
    - SCRAPED_DATE: Date when the data was scraped.
    - LISTING_NEIGHBOURHOOD: The neighborhood in which the property is located.
    - PROPERTY_TYPE: Type of property (e.g., apartment, house).
    - PRICE: Price of the property per night.
    - HAS_AVAILABILITY: Indicates if the property has availability.
    - AVAILABILITY_30: Availability of the property over the next 30 days.

    b. Host Columns:
    - HOST_ID: Unique identifier for the host.
    - HOST_NAME: Name of the host.
    - HOST_SINCE: Date since the host has been registered.
    - HOST_IS_SUPERHOST: Indicates if the host is classified as a "Superhost".
    - HOST_NEIGHBOURHOOD: The neighborhood of the host.

    c. Property Reviews Columns:
    - NUMBER_OF_REVIEWS: The total number of reviews for the property.
    - REVIEW_SCORES_RATING: Overall rating score of the property.
    - REVIEW_SCORES_ACCURACY: Rating score for the accuracy of the property's listing.
    - REVIEW_SCORES_CLEANLINESS: Cleanliness rating of the property.
    - REVIEW_SCORES_CHECKIN: Rating score for the check-in process.
    - REVIEW_SCORES_COMMUNICATION: Rating score for the host's communication.
    - REVIEW_SCORES_VALUE: Rating score for the value of the property.

2. New South Wales Local Government Area Data
    a. Local government area data
    - LGA_CODE: A unique code assigned to each Local Government Area (LGA) in New South Wales (NSW), Australia.
    - LGA_NAME: The name of the Local Government Area.

    b. Suburb data

- LGA_NAME: The name of the Local Government Area (LGA) in New South Wales (NSW), Australia.
- SUBURB_NAME: The name of the suburb within the corresponding LGA.

3. Census Data - Demographic and Socio-Economic Profile:
   This dataset has total 109 columns which can be further categorized among several categories such as
   a. Geographical Identifier: Unique code for each Local Government Area in NSW.
   b. Total Population By Gender: Population counts, split by male, female, and total.
   c. Population By Age And Gender: detail population by different age groups and gender.
   d. Census Night Location Count: record where people were on Census night, categorized by gender.
   e. Indigenous Population: provide data on Aboriginal and Torres Strait Islander populations.
   f. Birthplace: Differentiate between Australian-born and foreign-born residents.
   g. Language Spoken At Home: Columns like Lang_spoken_home_Eng_only_M indicate the primary language spoken at home.
   h. Australian Citizenship: Citizenship status.
   i. Educational Attendance By Age Group: Educational attendance data for various age groups
   j. Highest Year Of School Completed: Indicates the highest level of school education completed.
   k. Occupancy Count: Information on housing occupancy types.

4. Census Data - Income and Housing Statistics:
   - LGA_CODE_2016: Unique code identifying each Local Government Area in NSW.
   - Median_age_persons: Median age of persons within the LGA.
   - Median_mortgage_repay_monthly: Median monthly mortgage repayments in the LGA.
   - Median_tot_prsnl_inc_weekly: Median total personal income per week for individuals in the LGA.
   - Median_rent_weekly: Median weekly rent paid in the LGA.
   - Median_tot_fam_inc_weekly: Median total family income per week in the LGA.
   - Average_num_psns_per_bedroom: Average number of persons per bedroom in households in the LGA.
   - Median_tot_hhd_inc_weekly: Median total household income per week in the LGA.
   - Average_household_size: Average size of households in terms of the number of occupants in the LGA.

The Data Sources and Preparation section meticulously details the rich and diverse datasets at our disposal. It provides a clear understanding of each dataset's specific columns, underscoring their importance in painting a comprehensive picture of Sydney's Airbnb landscape and the demographic intricacies of New South Wales. This careful preparation and categorization of data lay a solid foundation for the insightful analyses that follow in our report, ensuring that the subsequent findings are both accurate and meaningful.

## DATA PIPELINE DESIGN WITH AIRFLOW.

Airflow serves as the backbone of our data engineering process, managing and automating the workflow with Directed Acyclic Graphs (DAGs). These DAGs efficiently orchestrate data operations, from initial extraction to final storage, ensuring seamless integration and processing of Airbnb and Census datasets for insightful analysis.

**Data Storage.**
In the GCP bucket australia-southeast1-bdeenv-29f99fa6-bucket/data, key datasets are stored, ready for processing. This includes Airbnb's monthly listings (05_2020.csv to 04_2021.csv), NSW Census profiles (2016Census_G01_NSW_LGA.csv, 2016Census_G02_NSW_LGA.csv), and geographical mappings (NSW_LGA_CODE.csv, NSW_LGA_SUBURB.csv). Precise organization of these files is pivotal for streamlined data retrieval and serves as the foundational step in our Airflow-driven pipeline.

**Raw Tables Creation.**
Using DBeaver, we initiated the creation of key raw tables within the GCP PostgreSQL database's raw_schema: airbnb_listing_all, Census_G01_NSW_LGA_2016, Census_G02_NSW_LGA_2016, nsw_lga_code, and nsw_lga_suburb. New tables were meticulously crafted, each structured to align with specific aspects of our datasets. After validating the schema through the insertion of a test record, we removed this entry, ensuring each table was pristine and primed for the actual data insertion process, thus establishing a solid foundation for our data analysis tasks.

**Airflow and Dags.**
Within our Airflow-managed data orchestration, a series of specialized DAGs efficiently governs the data pipeline. The load_listing_all_to_db DAG meticulously processes segmented Airbnb listing data, ingesting monthly segments from 05_2020.csv through 04_2021.csv into the raw_schema.airbnb_listing_all table. This ensures a chronological upload, preserving the temporal sequence essential for accurate trend analysis.

The load_nsw_lga_code_to_db DAG takes charge of the NSW_LGA_CODE.csv, populating the raw_schema.NSW_LGA_CODE table with Local Government Area codes, while its counterpart, load_nsw_lga_suburb_to_db, handles NSW_LGA_SUBURB.csv, filling the raw_schema.NSW_LGA_SUBURB table with the crucial links between LGAs and suburbs.

In parallel, the census datasets are deftly managed by two additional DAGs: load_census_G01_data_to_db ensures the 2016Census_G01_NSW_LGA.csv data is accurately reflected in raw_schema.Census_G01_NSW_LGA_2016, and load_census_G02_data_to_db commits the socio-economic details from 2016Census_G02_NSW_LGA.csv into raw_schema.Census_G02_NSW_LGA_2016.

These DAGs execute a well-coordinated dance of data logistics: downloading from the designated GCP bucket, careful truncation of target tables to remove stale data, and bulk-inserting the latest information. Such orchestrated precision across the data lifecycle not only enhances the reliability of the pipeline but also sets a clear path for downstream analytics, empowering decision-makers with the freshest and most organized data at their fingertips.

**Data Extraction, Loading, and Error Handling.**

Our data pipeline excels in extracting and loading data efficiently, with Airflow DAGs automating the retrieval from GCP buckets and insertion into PostgreSQL. Robust error handling ensures reliability; retries and alerts are in place for any hiccups during data transfer, guaranteeing data integrity and seamless pipeline continuity. This combination of systematic extraction, loading, and vigilant error monitoring forms the bedrock of dependable data analysis.

## DATA WAREHOUSE DESIGN WITH DBT

**Raw layer**

Our data warehouse, structured with dbt, features distinct layers—Raw, Staging, Warehouse, and Data Mart. Each serves a specific function, from initial raw data storage to refined transformation, leading to organized analysis-ready structures that drive informed business decisions.

From the airbnb_listing_all_view, we've derived specialized views—property, property_review, and host—each offering a focused lens on key data segments. These views facilitate targeted analysis, with property capturing listing details, host providing host-specific information, and property_review concentrating on customer feedback metrics, enriching our data landscape for nuanced insights. On this views we will do the analysis to understand what cleaning and transformation is needed in the staging layer.

Further employing dbt's snapshot functionality, we capture temporal states of our data with listing_snapshot of the Airbnb_listing_all_view view, property_snapshot, and host_snapshot. These snapshots, stored in raw_schema, track changes over time, keyed uniquely on listing_id and host_id, with scraped_date as the timestamp, enabling historical data analysis and trend observation.

**Staging layer**

In the staging layer, our meticulous examination of each table and view—spanning property, host, and property reviews—revealed null values. We've strategically filled these gaps, standardized column names, normalized text to lowercase, and imputed missing data with mode and median values for categorical and numerical fields, respectively, across seven transformative views.

In the property_stg view refines the airbnb_listing_all_view, transforming neighborhood and property type fields to lowercase for consistency. It introduces estimated_revenue, calculated only for available properties, enhancing the dataset with a vital financial metric. This staging layer view is pivotal for aligning disparate data types and preparing for advanced analytical functions.

The host_stg view is crafted through a strategic process utilizing the MODE() function within groups to identify the most frequent values for host attributes. By cross-joining with the mode_values subquery, it imputes missing values in the airbnb_listing_all_view. This approach replaces nulls in host names, registration dates, superhost status, and neighborhoods with modal values, ensuring data completeness. The transformation includes lowercase standardization for textual data and preserves critical identifiers like HOST_ID, listing_id, and SCRAPED_DATE, culminating in a comprehensive view tailored for subsequent stages of analysis.

The property_review_stg view in the staging layer addresses missing review scores within airbnb_listing_all_view. It calculates median values for various review metrics, such as rating, accuracy, and cleanliness, using the PERCENTILE_CONT function. These medians are then applied to fill any nulls, ensuring each property has complete review data. This enriched view maintains essential identifiers

like listing_id and host_id, and includes the scraped_date to provide temporal context for each review record, creating a robust dataset for analyzing guest feedback trends.

In the nsw_lga_suburb_stg and nsw_lga_code_stg views transform lga_name and suburb_name to lowercase, enhancing consistency and facilitating more accurate joins and analyses across our data models. Whereas The census_g01_nsw_lga_2016_stg and census_g02_nsw_lga_2016_stg views transform lga_code_2016, stripping the 'LGA' prefix and casting the remaining numeric code to integers, thereby standardizing LGA identifiers to facilitate accurate and efficient joins within our data warehouse architecture.

**Warehouse layer**

Within the warehouse layer, after ensuring the integrity of our data through rigorous validation of staging views, we established a robust structure of dimension tables: dim_property, dim_host, dim_nsw_lga_code, dim_nsw_lga_suburbs, dim_census_g01_nsw_lga_2016, and dim_census_g02_nsw_lga_2016. These tables encapsulate various data facets, from property details to geographic and demographic information. Central to this architecture is the fact_listings table, which serves as the single fact table, integrating key metrics and dimensions. The inclusion of the dm_geographic table, a product of joined LGA and suburb data, completes the framework, offering a holistic geographic perspective for in-depth analytical endeavors.

**Datamart layer**

In the data mart layer, we consolidate and transform data into insightful views, addressing business questions on listings and demographics, to empower stakeholders with actionable insights into market trends and guest-host dynamics within Airbnb's ecosystem. We have created the 3 views in it as follows.

The `dm_listing_neighbourhood` view consolidates Airbnb data into a monthly digest per neighbourhood, calculating active listings rates, pricing trends, and host metrics. It details minimum, maximum, median, and average active listing prices, distinct host counts, superhost rates, and average review scores. Additionally, it computes the total number of stays and average estimated revenue for active listings, offering a comprehensive monthly financial and operational overview. This organized view, grouped by neighbourhood and date, is invaluable for market trend analysis and strategic planning.

The view dm_property_type meticulously curates an analytical snapshot of Airbnb's diverse accommodations. It joins property details, host information, and listing reviews to construct a multi-dimensional perspective. Active listings are isolated using availability status, with a subsequent breakdown of pricing stats—minimum, maximum, median, and average. The view also calculates the rate of superhosts, the average review scores for active listings, and synthesizes an average estimated revenue figure. Grouped by neighbourhood and time, it provides stakeholders with a detailed monthly analysis of the listings, enabling data-driven decisions tailored to the nuances of property types and their performance over time.

The view dm_host_neighbourhood is a dbt materialized view that provides a summarized financial outlook per host and neighbourhood. It aggregates data on a yearly and monthly basis, combining host details from dim_host with property financial estimates from dim_property. The view computes the total and average estimated revenue per host, offering insights into the economic impact of Airbnb hosts in their respective localities. It also counts distinct hosts, presenting a clear picture of host activity and financial contribution across neighbourhoods. Ordered by neighbourhood and date, this view is instrumental for evaluating host performance and revenue trends.

## AD-HOC ANALYSIS

**Analysis One**

Mosman stands out as the best-performing neighbourhood with a high average revenue per listing of $9,326.52, compared to Fairfield, the worst-performing neighbourhood, with an average revenue of just $1,298.32. The output shows Mosman's residents have a median age of 42, a median monthly mortgage repayment of $3,000, and a higher weekly personal income of $1,295, indicative of its affluent status. Fairfield, in contrast, has a younger median age of 36, lower median mortgage repayments of $1,800, and a weekly personal income of $439, reflecting a more budget-conscious community. These economic contrasts underscore why Mosman achieves higher Airbnb revenues compared to Fairfield.

| lga_code | lga_name | median_age_ persons | median_mortgage_repay _monthly | median_tot_prsnl_inc _weekly | median_rent_ weekly | median_tot_fam _inc_weekly | average_num_psns _per_bedroom | median_tot_hhd_inc_ weekly | average_house hold_size |
|---|---|---|---|---|---|---|---|---|---|
| 12850 | fairfield | 36 | 1800 | 439 | 350 | 1263 | 1.1 | 1222 | 3.3 |
| 15350 | mosman | 42 | 3000 | 1295 | 560 | 3671 | 0.9 | 2522 | 2.4 |

**Analysis Two**

The data shows that private rooms in campers/RVs, entire bungalows, lofts, and villas are among the types of properties that have a full booking capacity, with an average stay of 30 days. This suggests that listings which offer a unique and possibly more intimate or specialized experience are highly sought after and could be considered the best type of listing in terms of occupancy and stay duration within the top-performing neighbourhoods.

| property_type | room_type | accommodates | avg_stays |
|---|---|---|---|
| private room in camper/rv | private room | 3 | 30 |
| camper/rv | entire home/apt | 3 | 30 |
| shared room in condominium | shared room | 3 | 30 |
| bungalow | private room | 1 | 30 |
| loft | entire home/apt | 5 | 30 |
| private room in villa | private room | 4 | 30 |
| villa | private room | 5 | 30 |
| entire floor | entire home/apt | 4 | 30 |
| hostel | shared room | 12 | 30 |
| apartment | shared room | 5 | 30 |
| entire home/apt | entire home/apt | 2 | 30 |
| entire chalet | entire home/apt | 2 | 30 |
| private room in villa | private room | 5 | 30 |
| bed and breakfast | shared room | 12 | 30 |
| hostel | shared room | 10 | 30 |
| barn | entire home/apt | 4 | 30 |
| shared room in bed and breakfast | shared room | 12 | 30 |
| villa | entire home/apt | 5 | 30 |
| entire house | entire home/apt | 15 | 30 |
| casa particular | entire home/apt | 2 | 30 |

**Analysis Three**

The data indicates that the majority of hosts with multiple listings tend to have their properties in the same LGA where they reside. Specifically, 24,327 hosts, representing the larger proportion, have over 50% of their listings in the same LGA, suggesting a preference or convenience for managing properties within the same geographic area. Conversely, a smaller segment of 5,393 hosts has less than half of their listings in their residential LGA, this suggests that the hosts with multiple listings are more inclined to have their listings in the same LGA as where they live.

| host_id | same_lga_count | total_listings | same_lga_percentage | | | | | |
|---|---|---|---|---|---|---|---|---|
| 14093 | 4 | 4 | 100 | | | | | |
| 16474 | 9 | 9 | 100 | | | | | |
| 17061 | 529 | 529 | 100 | | | | | |
| 17331 | 576 | 576 | 100 | | | | | |
| 18459 | 144 | 144 | 100 | | | | | |
| 20258 | 49 | 49 | 100 | | | | | |
| 27184 | 36 | 36 | 100 | | | COUNTIF(C:C, ">50") | 24327 | #number of host having more than 50% of the property in same lga |
| 34305 | 144 | 144 | 100 | | | COUNTIF(C:C, "<50") | 5393 | #number of host having less than 50% of the property in same lga |
| 40855 | 144 | 144 | 100 | | | | | |
| 41302 | 100 | 100 | 100 | | | | | |
| 42647 | 144 | 144 | 100 | | | | | |
| 44298 | 121 | 121 | 100 | | | | | |
| 44490 | 144 | 144 | 100 | | | | | |
| 46116 | 121 | 121 | 100 | | | | | |
| 47896 | 144 | 144 | 100 | | | | | |
| 52279 | 1296 | 1296 | 100 | | | | | |
| 55948 | 576 | 576 | 100 | | | | | |
| 57949 | 4 | 4 | 100 | | | | | |

**Analysis Four**

Analysis of annual estimated revenues against median mortgage repayments reveals a favorable trend for hosts with a unique listing. Out of the evaluated hosts, 127,719 can cover their annualized median mortgage repayments through their listing's revenue, surpassing the 126,264 who cannot. This indicates a majority of single-listing hosts generate sufficient income from their Airbnb to meet or exceed their yearly mortgage obligations, reflecting a positive economic impact of hosting on individual finances.

| host_id | annual_revenue | annual_median_mortgage_repayment | can_cover_mortgage | | | |
|---|---|---|---|---|---|---|
| 40855 | 2663 | 29988 | FALSE | | | |
| 59850 | 7332 | 29988 | FALSE | | | |
| 112237 | 42930 | 29988 | TRUE | | | |
| 279955 | 37638 | 36000 | TRUE | COUNTIF(D:D, TRUE) | 127719 | #can cover |
| 318390 | 61200 | 36000 | TRUE | COUNTIF(D:D, FALSE) | 126264 | cannot cover |
| 322887 | 30270 | 38400 | FALSE | | | |
| 323771 | 4087 | 29988 | FALSE | | | |
| 16474 | 7230 | 29988 | FALSE | | | |
| 160705 | 1877 | 31200 | FALSE | | | |
| 333581 | 4782 | 28800 | FALSE | | | |
| 345292 | 22024 | 33600 | FALSE | | | |
| 352614 | 26913 | 28800 | FALSE | | | |
| 431242 | 0 | 36000 | FALSE | | | |
| 434914 | 12400 | 31200 | FALSE | | | |
| 476047 | 23369 | 36000 | FALSE | | | |
| 453466 | 534 | 29988 | FALSE | | | |
| 480315 | 22051 | 33600 | FALSE | | | |
| 407155 | 95600 | 31200 | TRUE | | | |
| 172878 | 88641 | 36000 | TRUE | | | |

## CONCLUSION

Our analysis effectively merges Airbnb listing data with Australian Census information to tackle important business questions. We reveal the profitability of distinct property types in Sydney's prime areas and the prevalence of hosts managing properties in their own local government areas. Significantly, many single-listing hosts can cover their annual mortgage payments through Airbnb income, emphasizing the platform's financial potential for individuals. Our project showcases the importance of organized data pipelines and warehouses in extracting valuable insights from intricate datasets.

Word Count- 2996