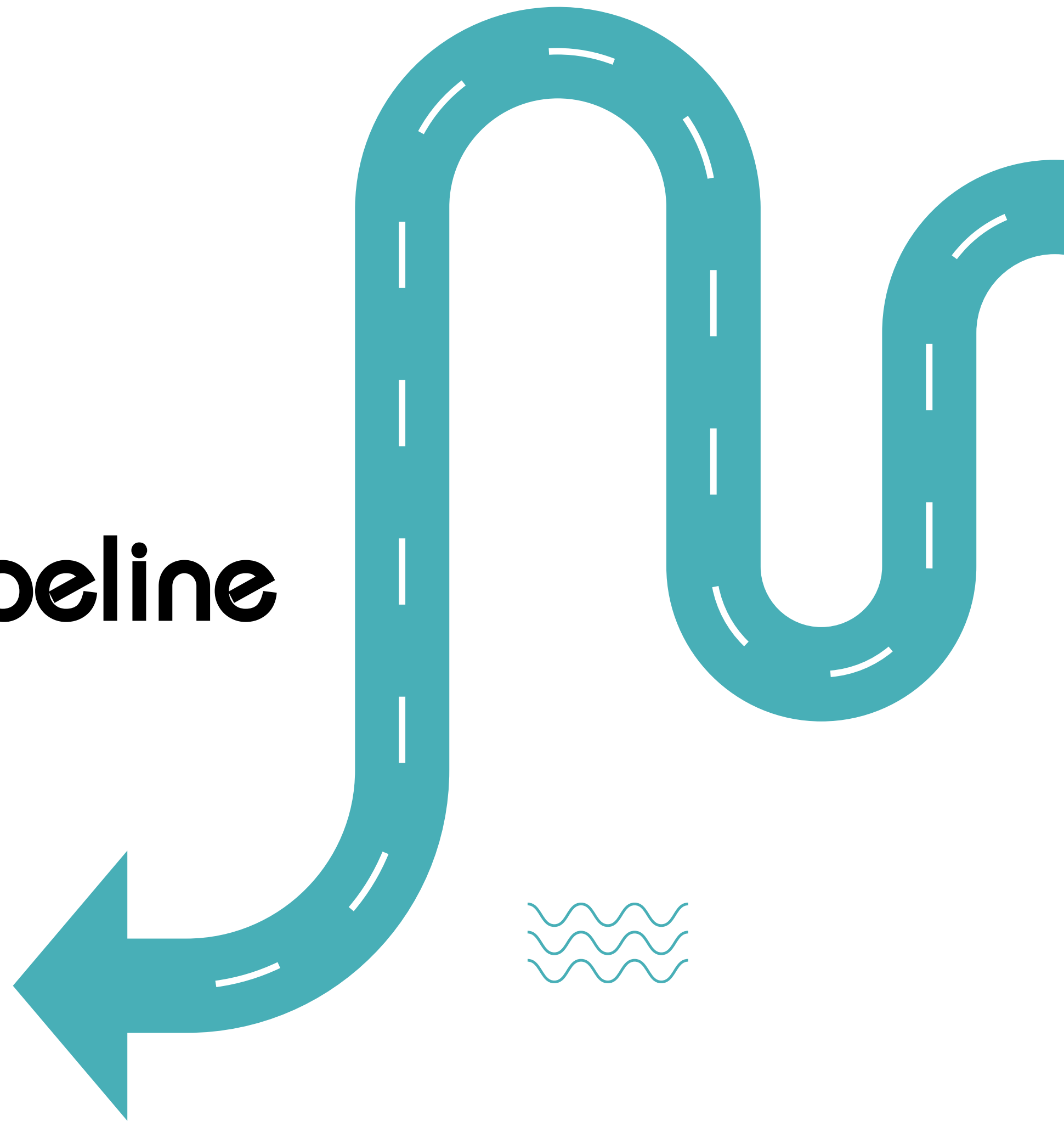
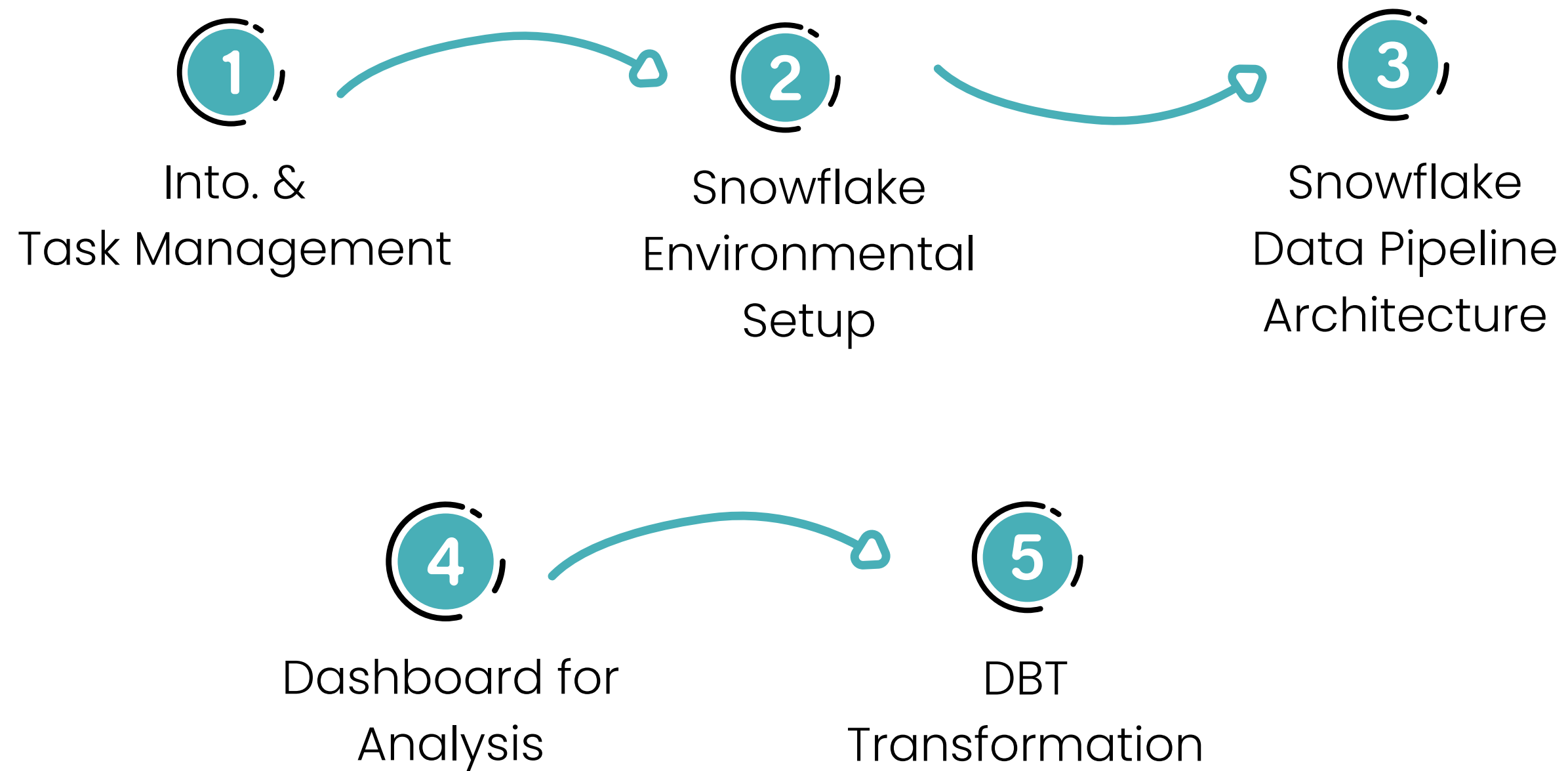


# Data Integration and Transformation Pipeline



# Content





# Brainstorming

Trello:

1. DB, Schema, WH

2. Grant privilege - Roles

Sug - 3

Tia - 4

Don

Sum

Select

3. 3 Tables - RAW

Housing Rent

(4) Refined - Table - Needed Col.

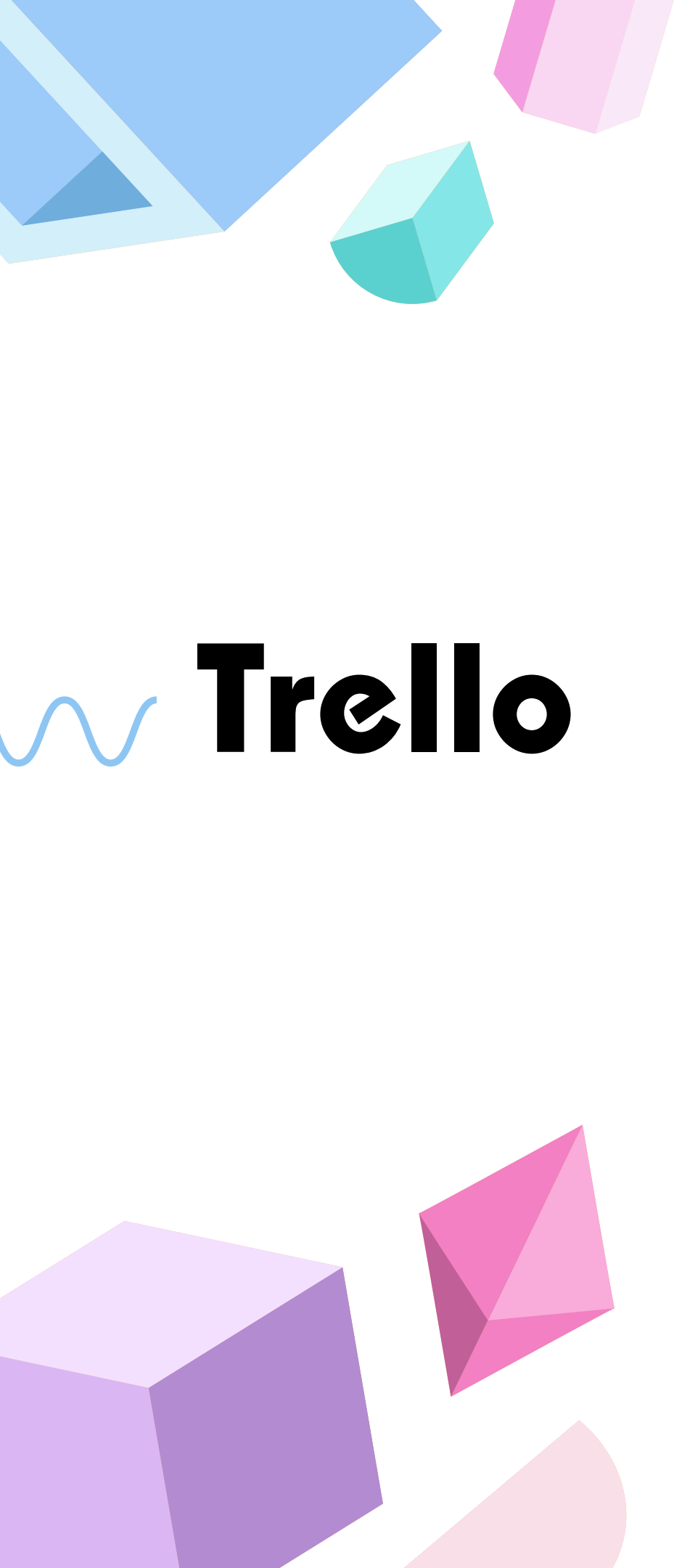
← S. cleaning & Analysing

Build Dbt modules

6. 4 analysis - 4 quest.

7. Dash - 4 analysis





# Trello

**Group 4 - Data Pipeline Task Management**

**Backlog**

- + Add a card

**In Progress**

- Final Deliverables
  - Jun 10 2/3
- Prepare the presentation slides
  - Jun 11
- Conduct sprint retrospective
  - 0/3
- + Add a card

**Review** 0

**Done**

- POD
  - May 28
    - SG SM TY DY
- Create: DB,Schema and Warehouse
  - May 30
    - SM
- Grant privilege to the Roles
  - May 31
    - TY
- Identify dataset with 3 different file formats
  - May 30
    - 3/3
      - SM DY TY SG
- Create RAW Schema: Upload 3 tables to Snowflake
  - May 31
    - 3/3
      - SM TY DY SG
- Create PREP Schema: Simple changes from raw schema
  - Jun 3
    - 3/3
      - SM DY SG TY
- Create REFINE Schema
  - May 31 0/1
    - SG
- Create Delivery Schema
  - May 31
    - DY

**Done**

- Set up stakeholders & 4 questions to lead the analysis
  - Jun 5 3/4
    - DY SM SG TY
- Study and understand the structure of all datasets
  - Jun 4
    - TY DY SM SG
- Data Cleaning
  - Jun 5 2/2
    - DY SM
- Refinement Schema Anlaysis
  - Jun 4
    - TY DY SM SG
- Check-in with IL(JAKOB) and group discussion about the doubts
  - SM DY SG TY
- Create a dashboard in Snowflake
  - 1 Jun 9 3/4
    - SM SG TY DY
- Build DBT models for the analysis and update DAG
  - 2 Jun 6
    - SM DY SG TY



# Snowflake Environmental Setup



Create Team Warehouse

Create Database and Schemas



- RAW
- PREP
- REFINED
- DELIVERY



Grant Privileges











# Grant Privileges



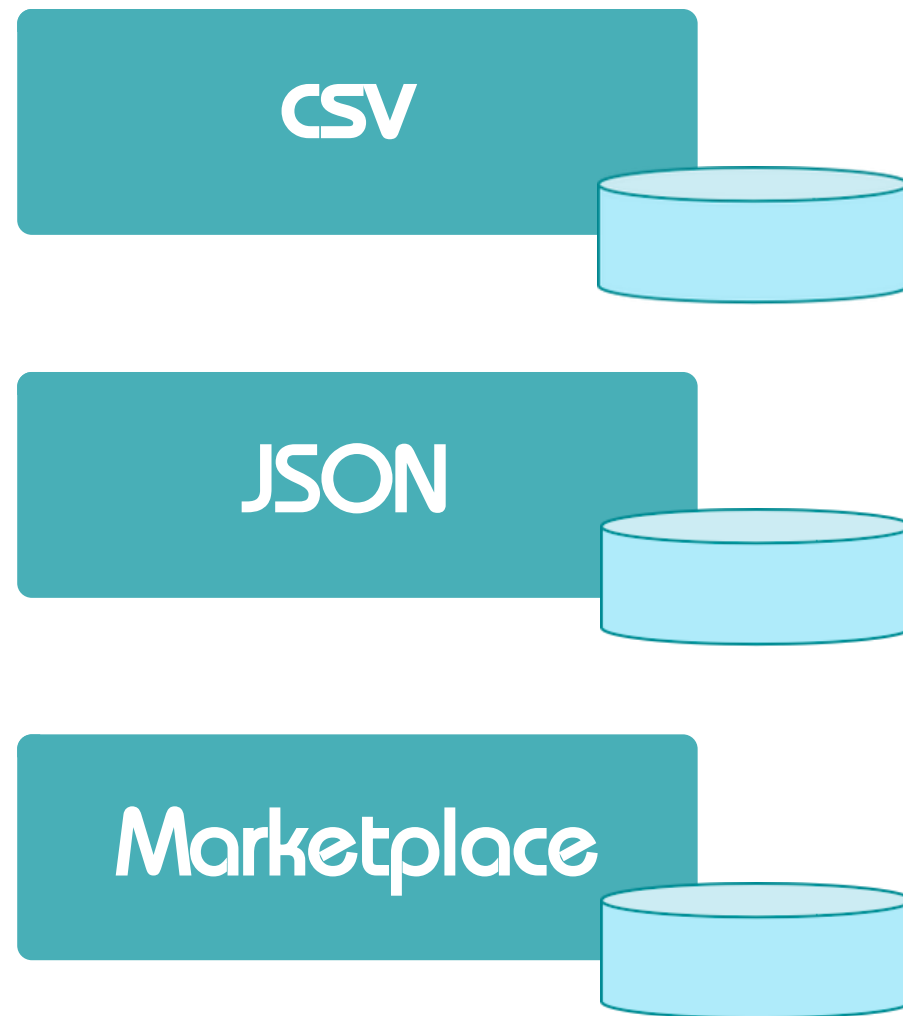
## Privileges

 CHIPMUNK_ROLE (Current Role)	USAGE
 KOALA_ROLE	USAGE
 LEMMING_ROLE	USAGE
 LEMUR_ROLE	USAGE
 TEAM_4_USER_ROLE	<div>🔍 OWNERSHIP</div> <div>DELETE - FUTURE TABLE</div> <div>INSERT - FUTURE TABLE</div> <div>SELECT - FUTURE TABLE</div> <div>SELECT - FUTURE VIEW</div> <div>UPDATE - FUTURE TABLE</div> <div>USAGE - FUTURE SCHEMA</div>
 TEAM_4_VIEWER_ROLE	<div>SELECT - FUTURE VIEW</div> <div>USAGE</div>

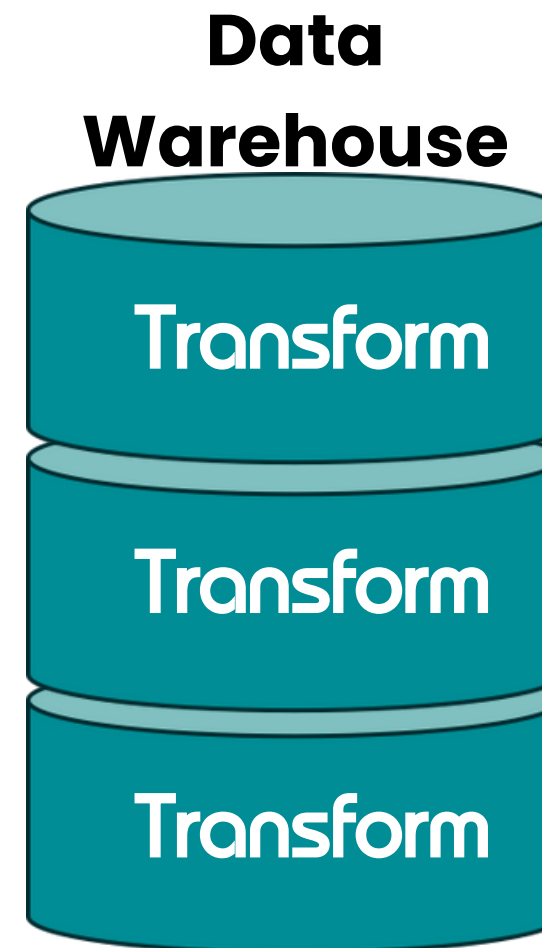


# Snowflake Data Pipeline Architecture

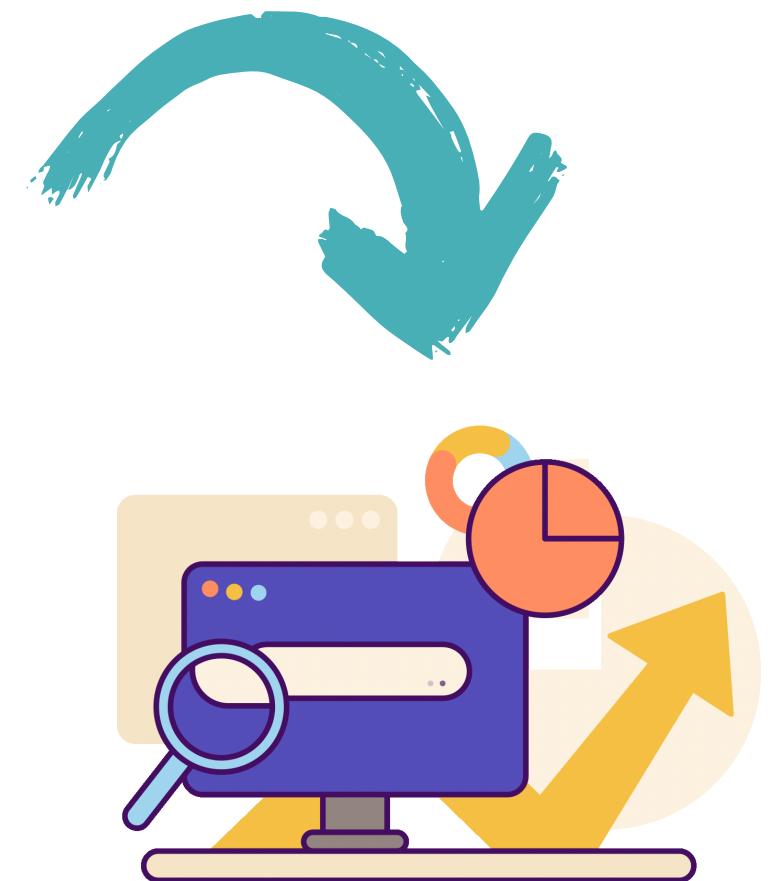
RAW → PREP → REFINED →  
DELIVERY



**RAW  
Data**



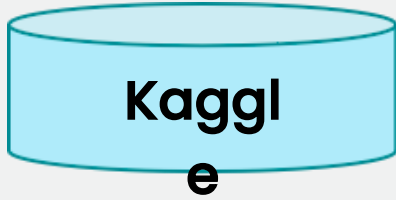
Analyze



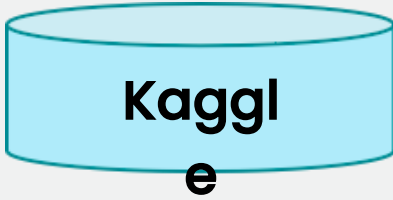
Extract



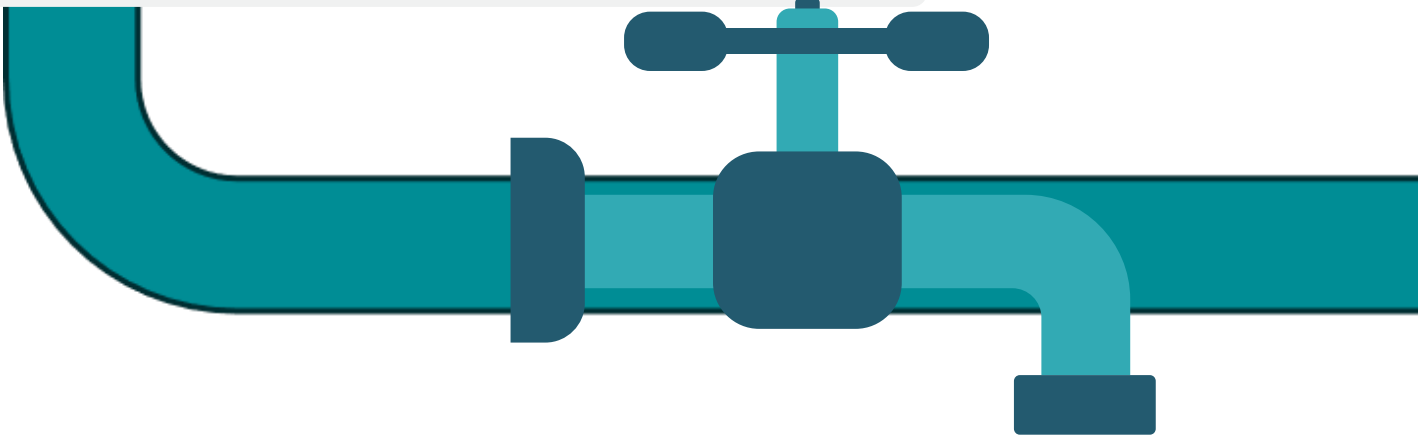
Housing rental information data



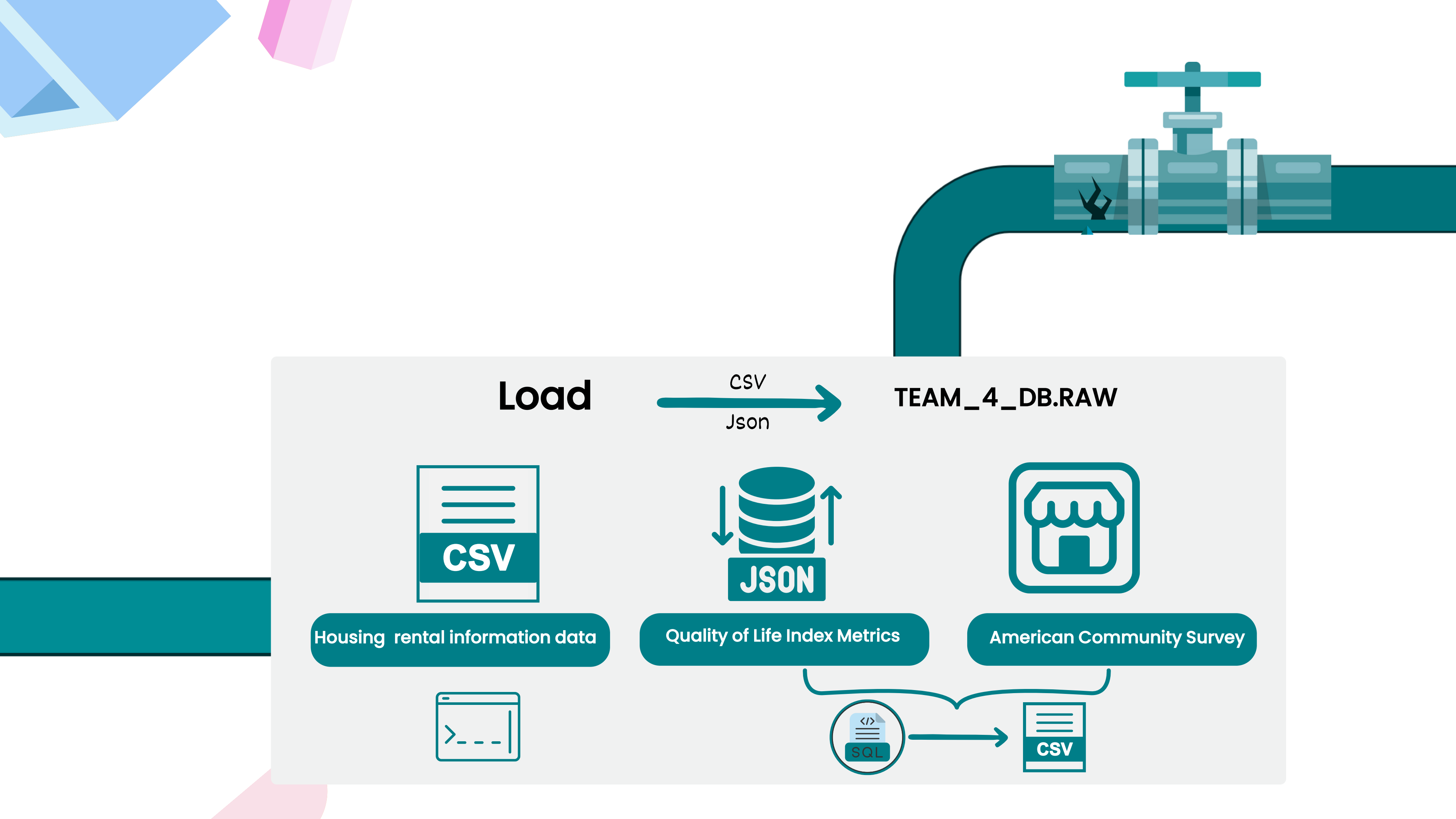
Quality of Life Index Metrics



American Community Survey









Load



```
C:\Users\sugan>snowsql -a nlb11398 -u KOALA -r KOALA_ROLE -d TEAM_4_DB -s RAW
* SnowSQL * v1.4.1
Type SQL statements or !help
KOALA#LEARNER_WH@TEAM_4_DB.RAW>USE WAREHOUSE TEAM_4_WAREHOUSE;
+-----+
| status |
+-----+
| Statement executed successfully. |
+-----+
1 Row(s) produced. Time Elapsed: 0.246s
KOALA#TEAM_4_WAREHOUSE@TEAM_4_DB.RAW>SELECT CURRENT_WAREHOUSE();
+-----+
| CURRENT_WAREHOUSE() |
+-----+
| TEAM_4_WAREHOUSE    |
+-----+
1 Row(s) produced. Time Elapsed: 0.233s
KOALA#TEAM_4_WAREHOUSE@TEAM_4_DB.RAW>PUT
file://C:/Users/sugan/OneDrive/Desktop/Hyper_Island/Course_8_Data_Engineering/
```

#Raw



Transform



- Standardized column names
- Limited to the needed columns
- Type conversion
- Parsed JSON to structured columns

#Raw → Prep

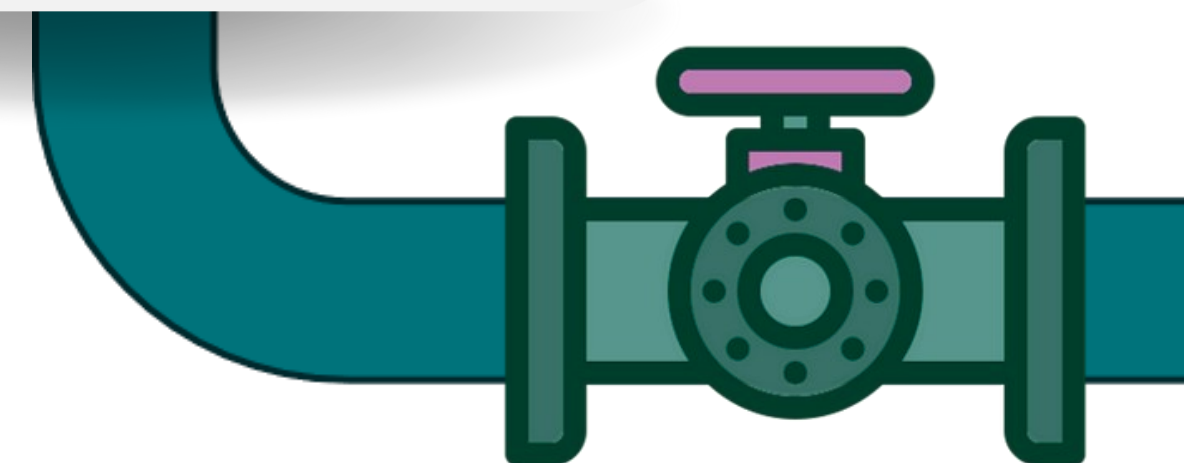


Transform



- Data Cleaning:
- Standardizing State Codes
- Removing outliers

#Raw → Prep → Refined



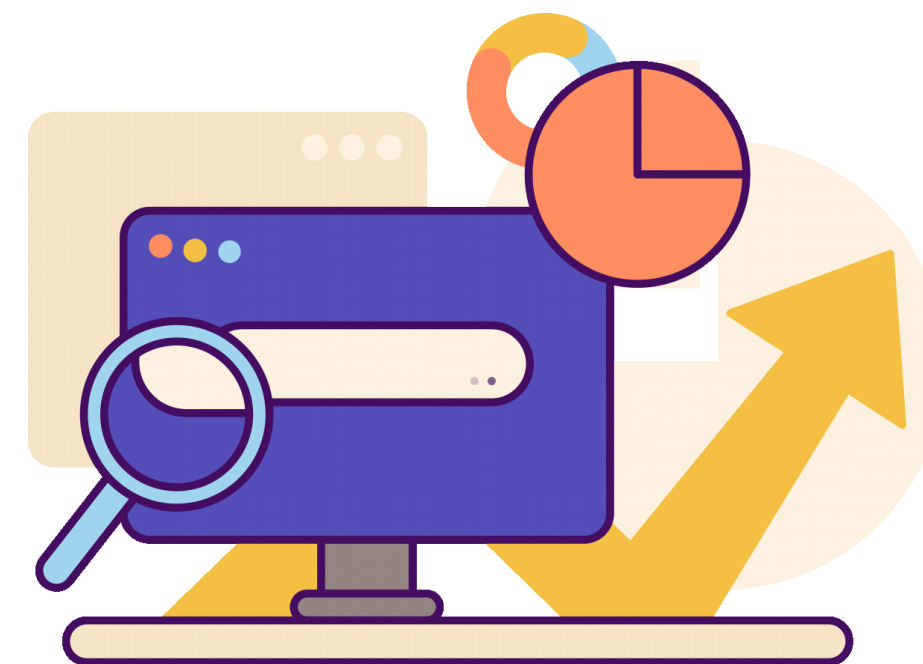


Dashboard

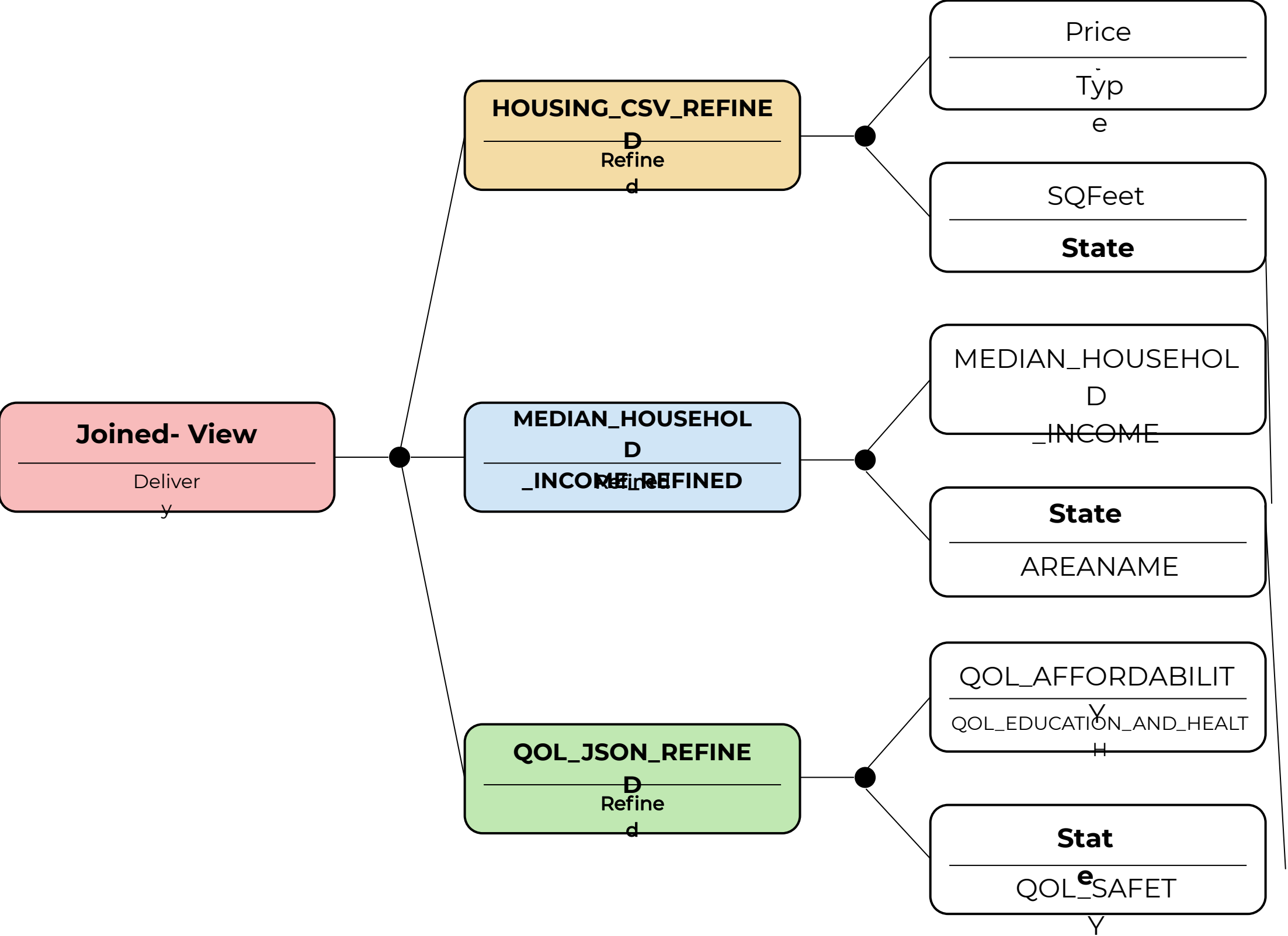
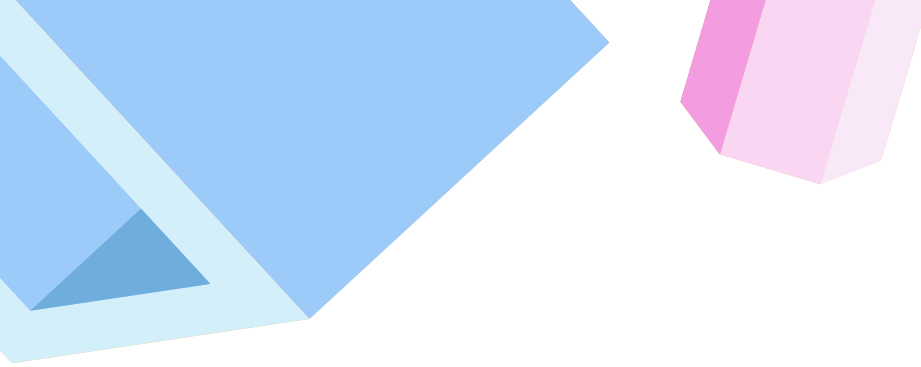


- Using Team\_4\_Viewer\_role to create dashboard
- From refined schema joined 3 tables together for analysis.

#Raw → Prep → Refined → Delivery

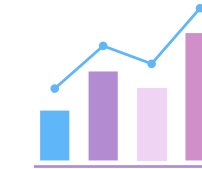








# Snowflake Dashboard



< Dashboards Housing\_Rent\_Analysis ▾

TEAM\_4\_VIEWER\_ROLE TEAM\_4\_WAREHOUSE (X-Small)

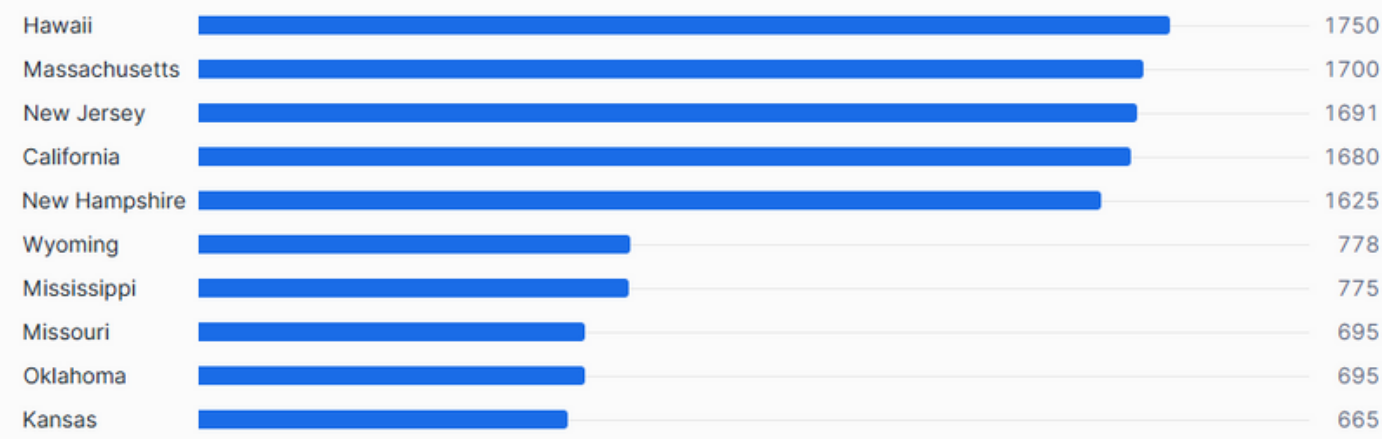
Share

▶ Run

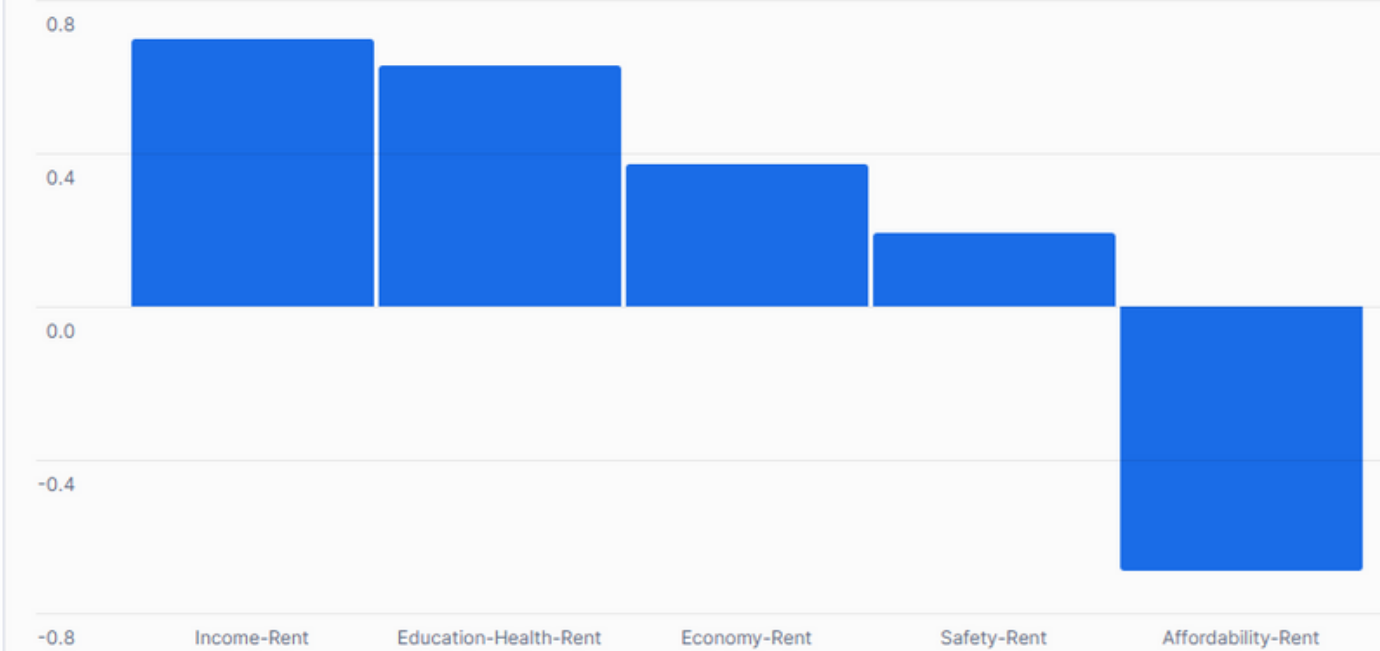


Updated 11h ago

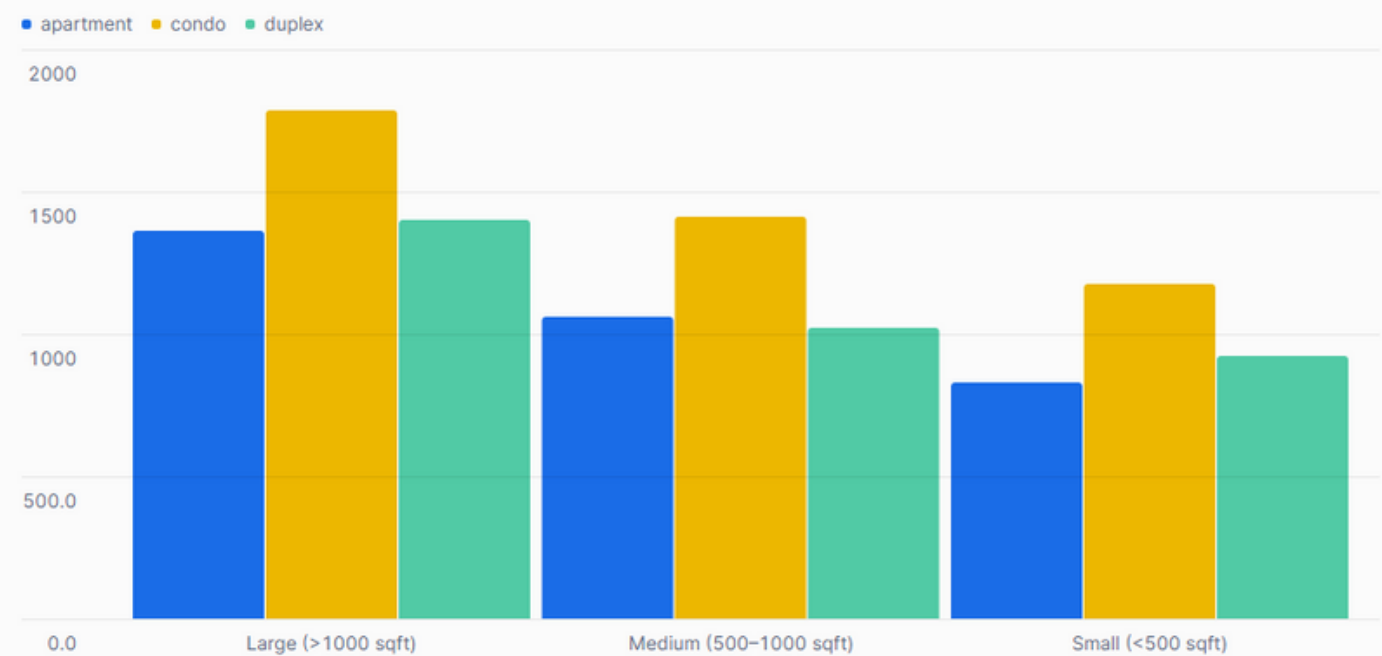
Top 5 & Bottom 5 States By Monthly Rent Price



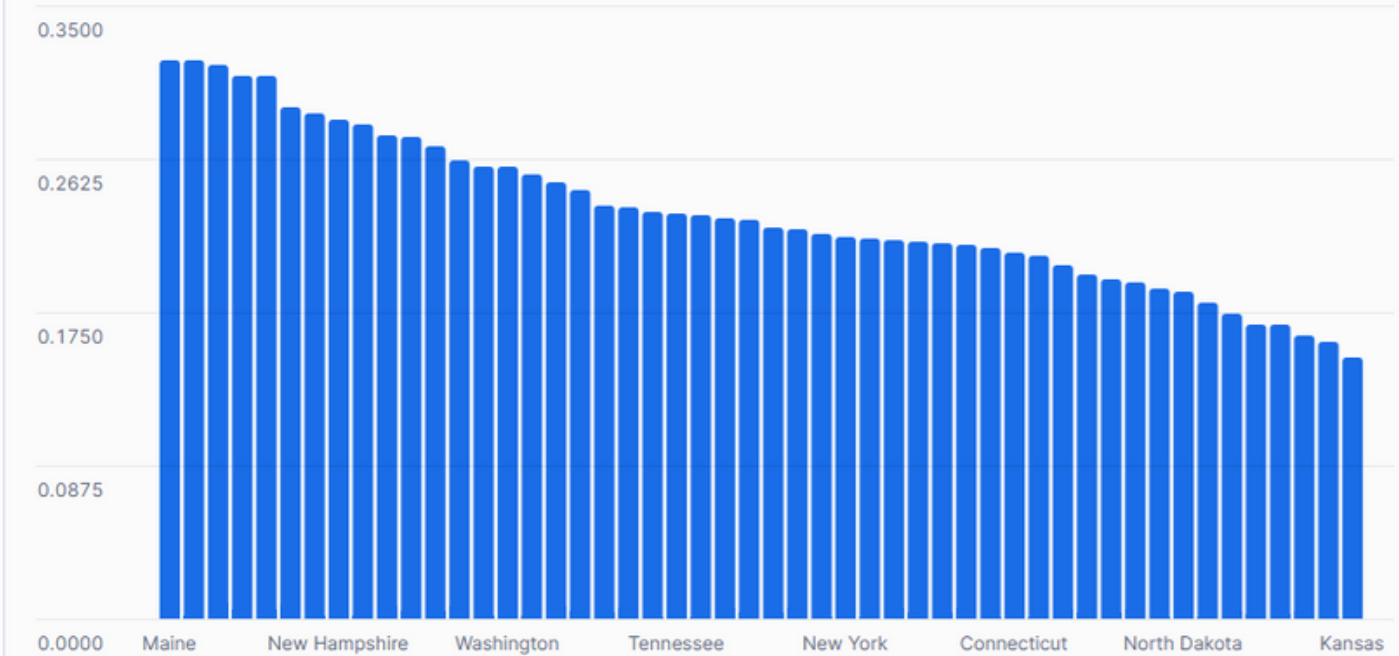
Rent\_Price\_Influencer



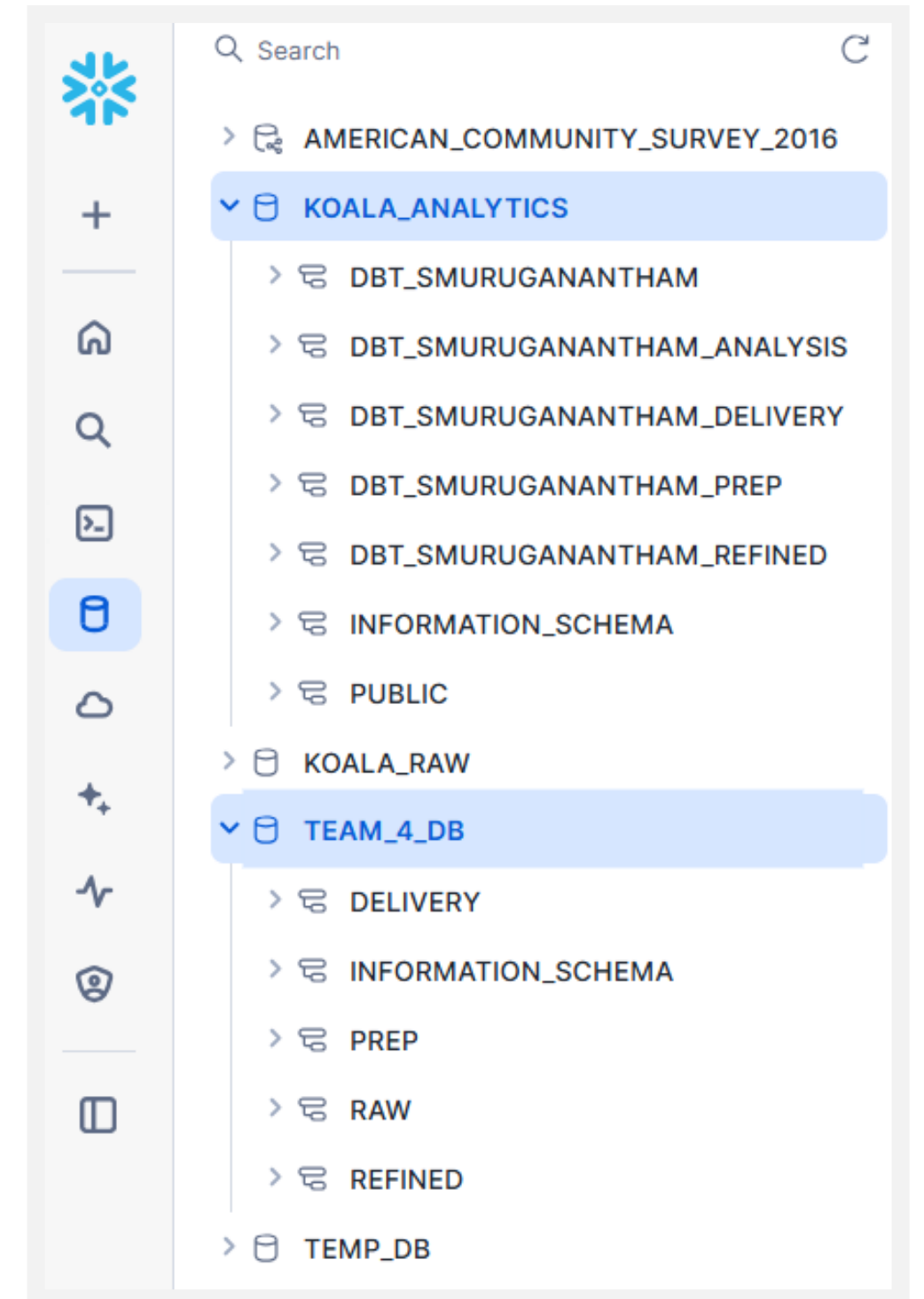
Average Rent by Size Group & Type



Rend Burden



# Integrated pipeline





# dbt Snowflake Pipeline Architecture

- Transformation
- Data Tests



PREP  
REFINED



RAW



DELIVERY



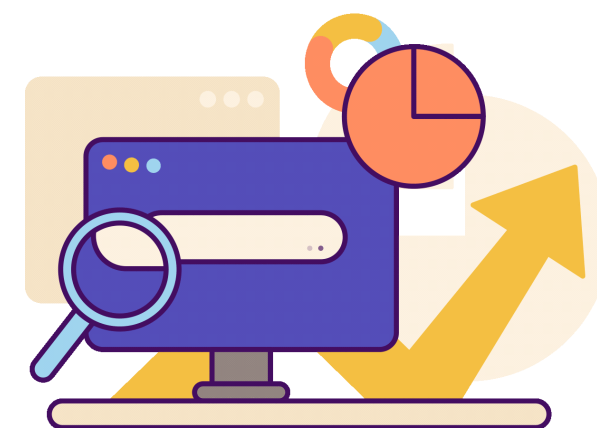
Raw Data

Transformed Data

CSV

JSON

Marketplace





# GitHub collaboration

## Change Branch

Git Branch

main — updated Mon Jun 09 2025

main — updated Mon Jun 09 2025

suganyam2001-housing — updated Mon Jun 09 2025 (current branch)

## Users

🔍 Search users by name or email

Name ↑

License

Danqing Yao

danqing.yao@hyperisland.se

Developer

Suganya Muruganantham

suganya.m2001@gmail.com

Developer

Sunny Gustavsson

sunny.zhang.qing@gmail.com

Developer

Tanglan Yang

volingla@gmail.com

Developer

# DBT Setup Data Pipeline

RAW → PREP → REFINED → DELIVERY



## Sources



raw

- HOUSING\_CSV\_RAW
- MARKETPLACE\_INCOME\_RAW
- QOL\_JSON\_RAW

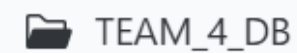
## Projects



my\_new\_project



models



TEAM\_4\_DB

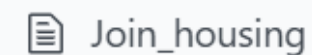


Analysis

- Danq\_rent\_burden
- Sunny\_states\_rent\_price
- sug\_correlation\_metrics
- tiana\_avg\_rent\_size\_and\_type



DELIVERY

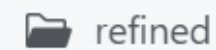


Join\_housing



PREP

- prep\_housing\_data
- prep\_median\_income
- prep\_qol\_data



refined

- refined\_housing\_data
- refined\_median\_income
- refined\_qol\_data



tests



Joined\_Analysis



Databases Worksheets

Search objects

- > AMERICAN\_COMMUNITY\_SURVEY\_2016
- > KOALA\_ANALYTICS
  - > DBT\_SMURUGANANTHAM
  - > DBT\_SMURUGANANTHAM\_ANALYSIS
  - > DBT\_SMURUGANANTHAM\_DELIVERY
  - > DBT\_SMURUGANANTHAM\_PREP
  - > DBT\_SMURUGANANTHAM\_REFINED
  - > INFORMATION\_SCHEMA
  - > PUBLIC

# DAG

