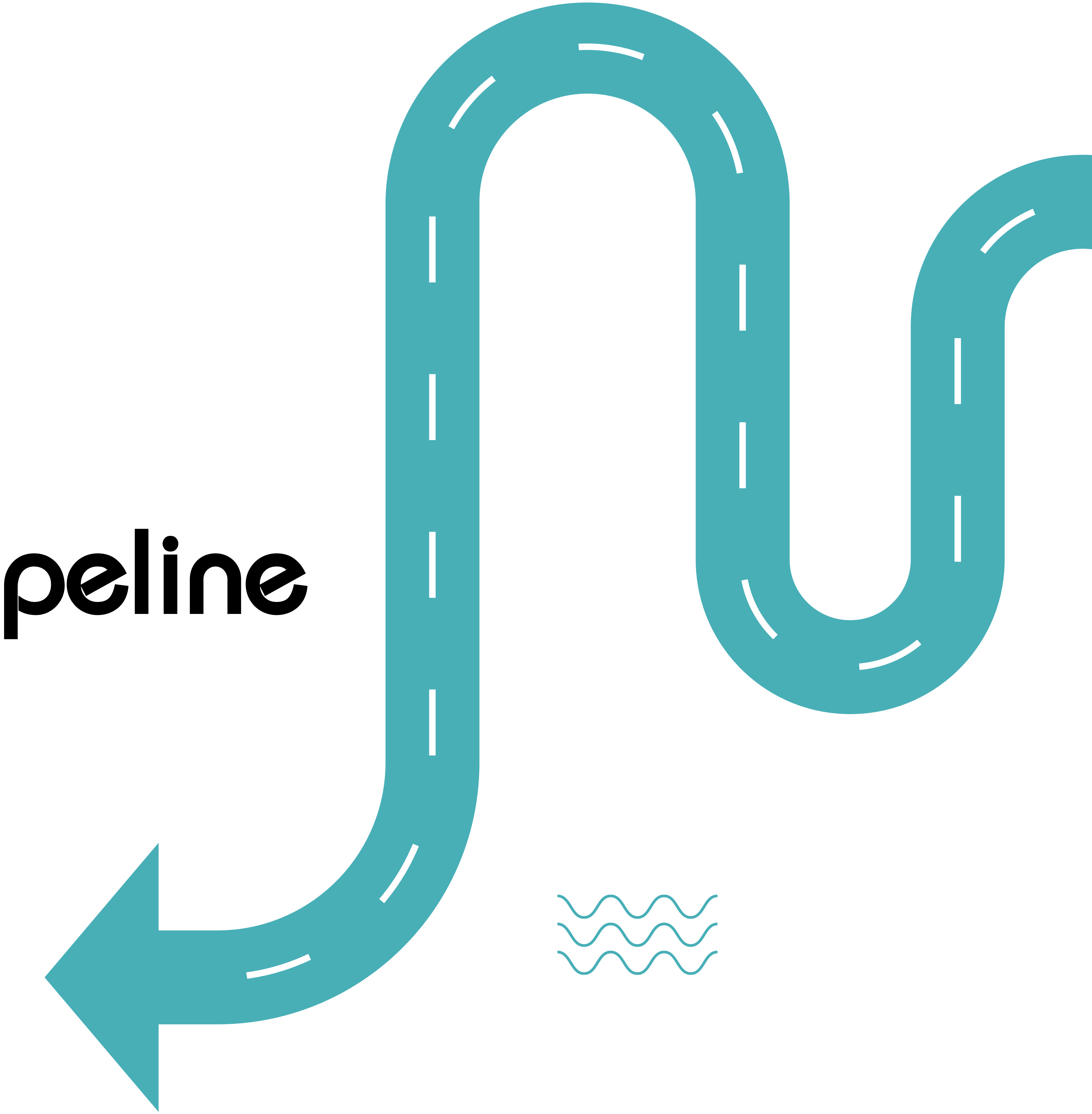
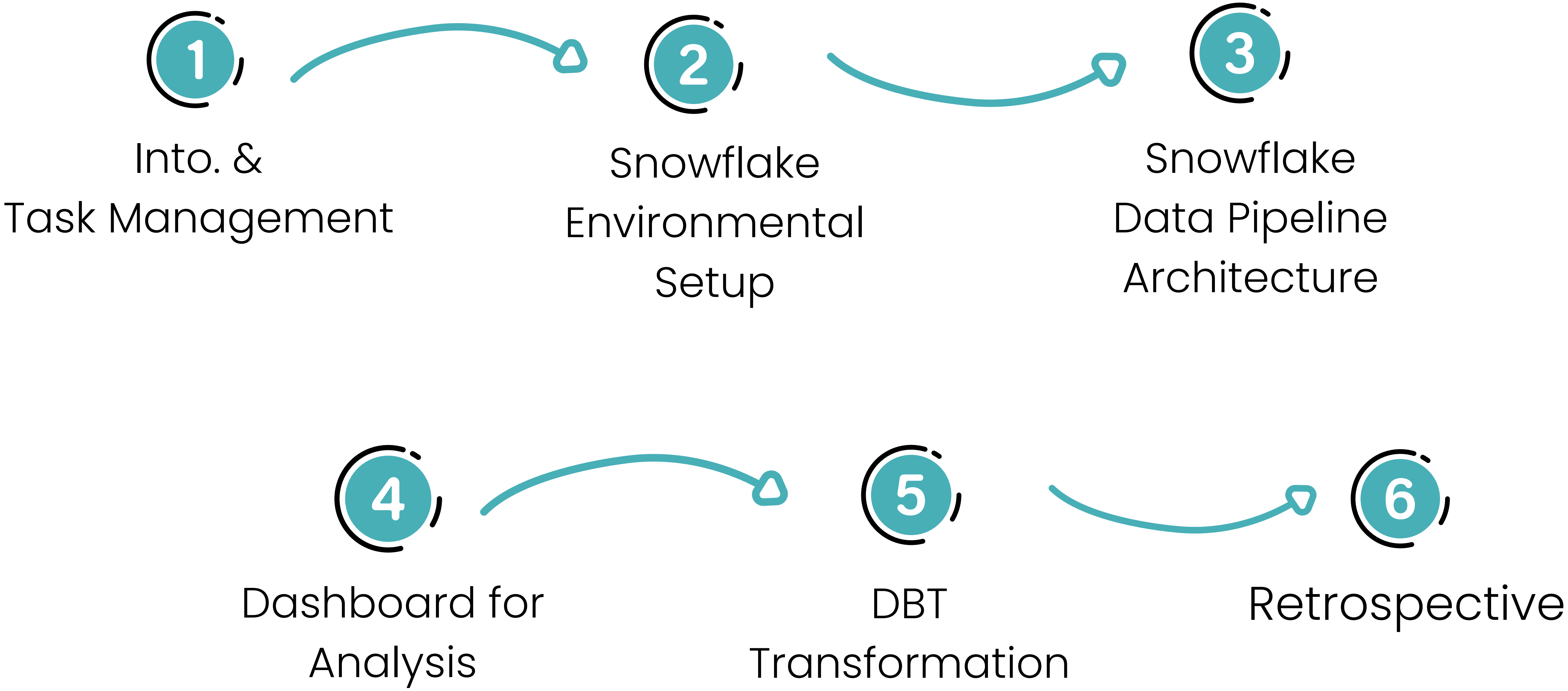
Data Integration and Transformation Pipeline



Content



1 Into. & Task Management

2 Snowflake Environmental Setup

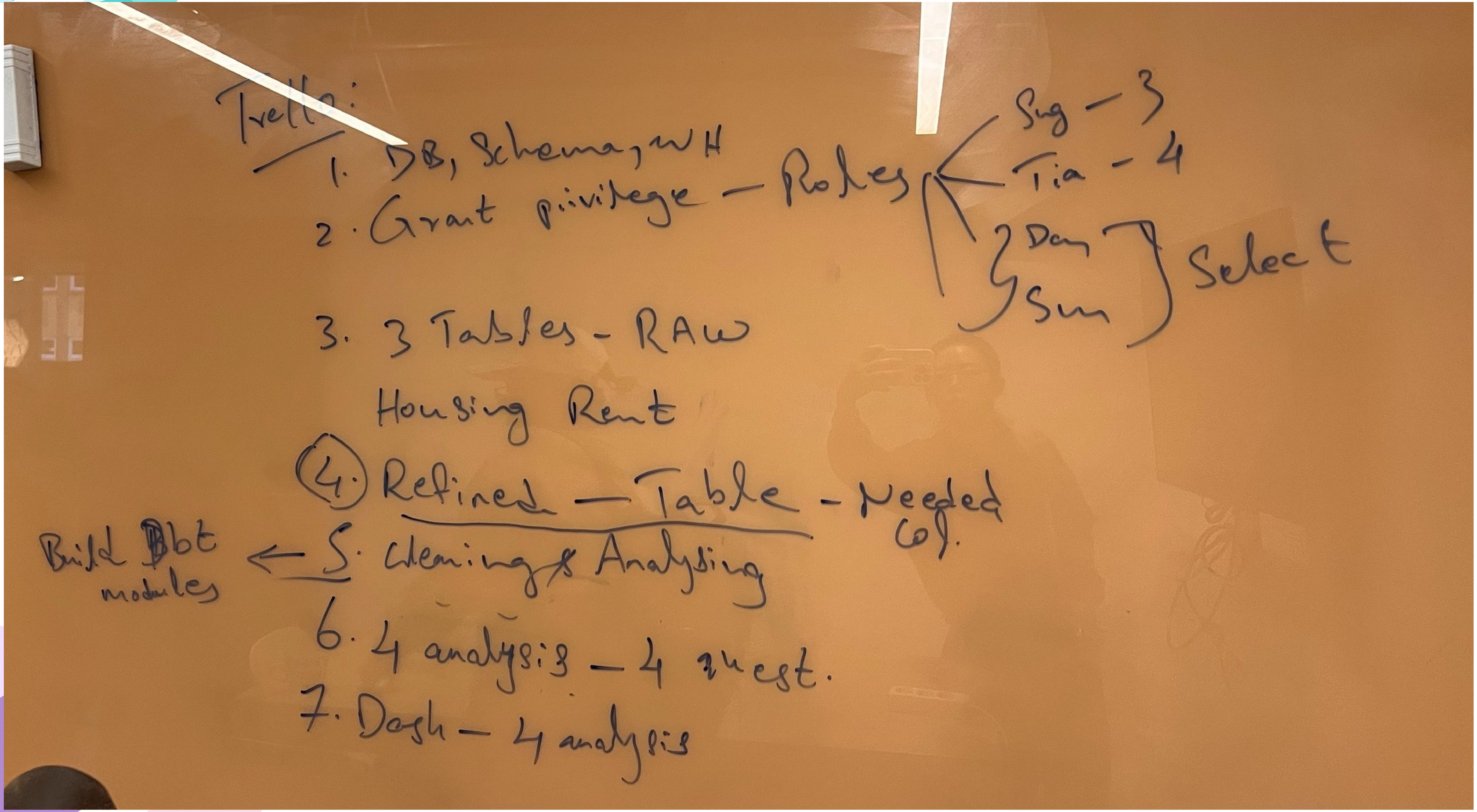
3 Snowflake Data Pipeline Architecture

4 Dashboard for Analysis

5 DBT Transformation

6 Retrospective

Brainstorming



Trello:

1. DB, Schema, WH
2. Grant privilege — Roles

Sug — 3

Tia — 4

Dan

Sun

Select

3. 3 Tables - RAW

Housing Rent

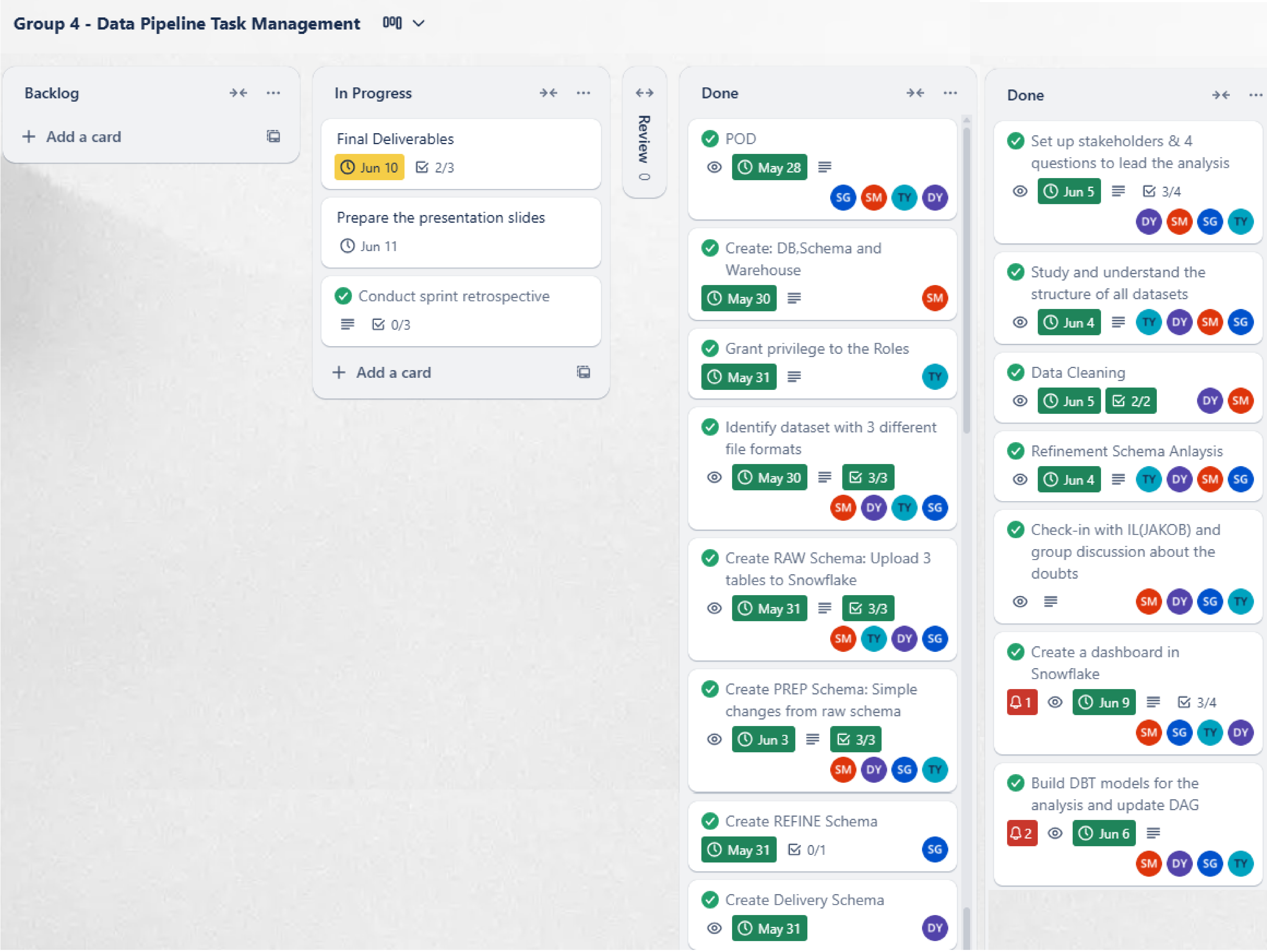
4. Refined — Table - Needed Col.

Build Dbt modules ← 5. Cleaning & Analysing

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

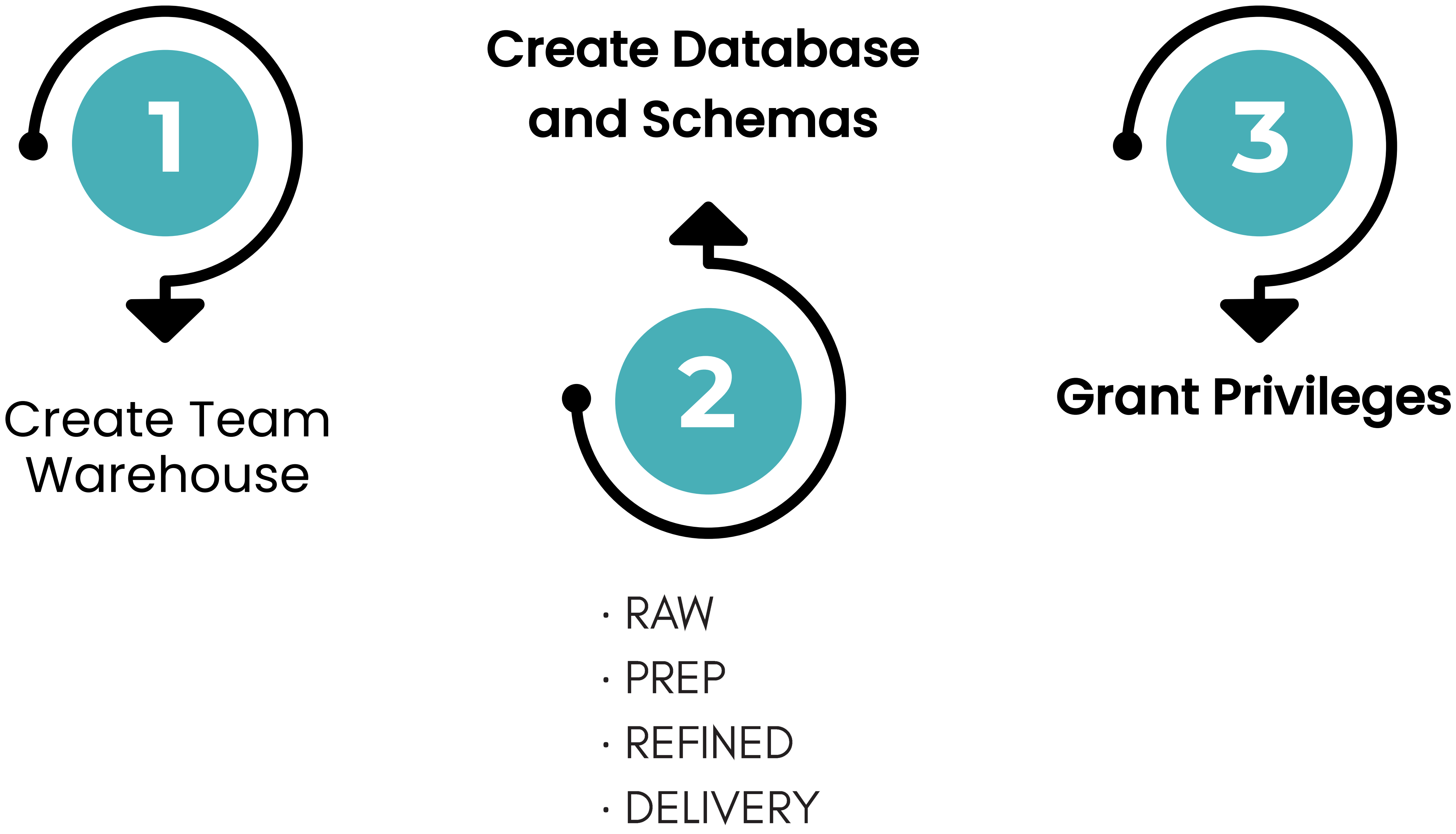
✓ Refinement Schema Anlaysis
Jun 4
TY DY SM SG

✓ 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



1

Create Team Warehouse

2

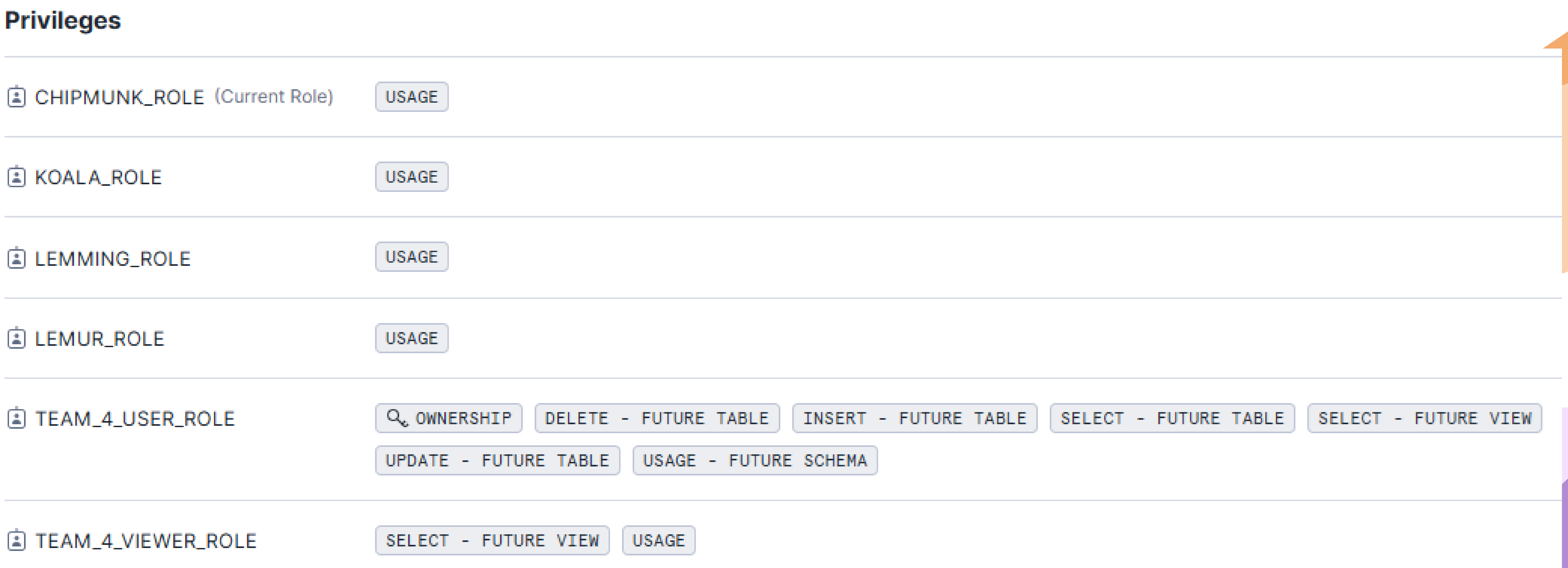
Create Database and Schemas

- RAW
- PREP
- REFINED
- DELIVERY

3

Grant Privileges

Grant Privileges



Privileges

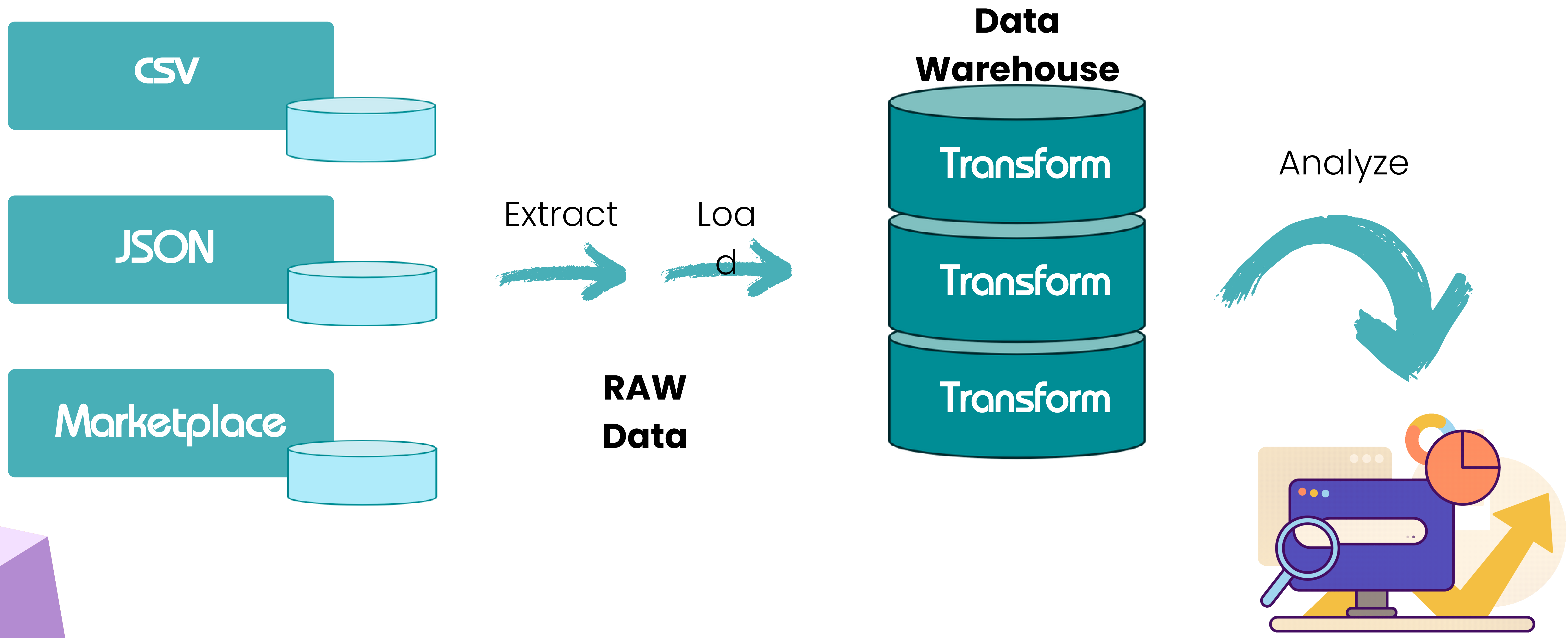
CHIPMUNK_ROLE (Current Role)	USAGE
KOALA_ROLE	USAGE
LEMMING_ROLE	USAGE
LEMUR_ROLE	USAGE

TEAM_4_USER_ROLE	OWNERSHIP
	DELETE - FUTURE TABLE
	INSERT - FUTURE TABLE
	SELECT - FUTURE TABLE
	SELECT - FUTURE VIEW
	UPDATE - FUTURE TABLE
	USAGE - FUTURE SCHEMA

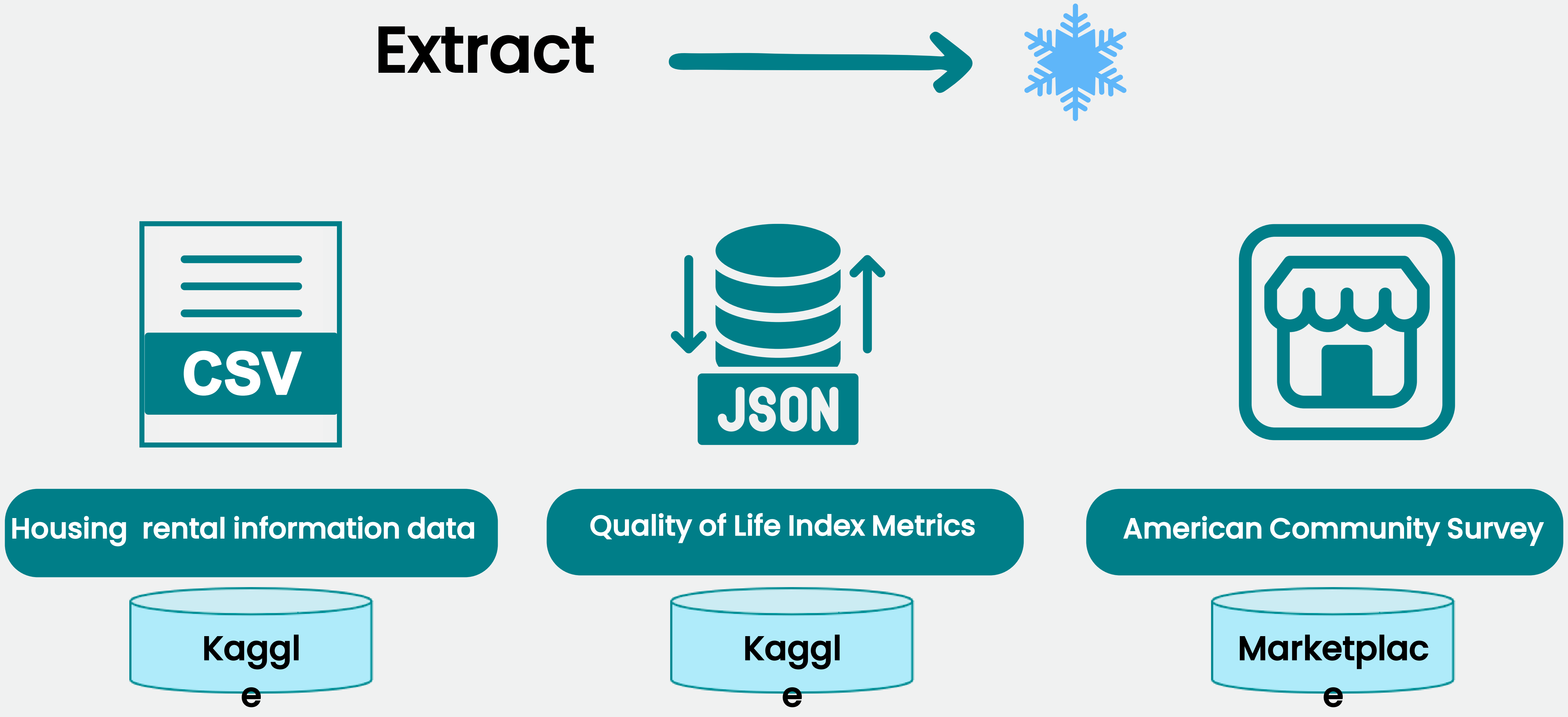
TEAM_4_VIEWER_ROLE	SELECT - FUTURE VIEW USAGE

Snowflake Data Pipeline Architecture

RAW ➔ PREP ➔ REFINED ➔
DELIVERY



Extract



CSV

JSON

Housing rental information data

Quality of Life Index Metrics

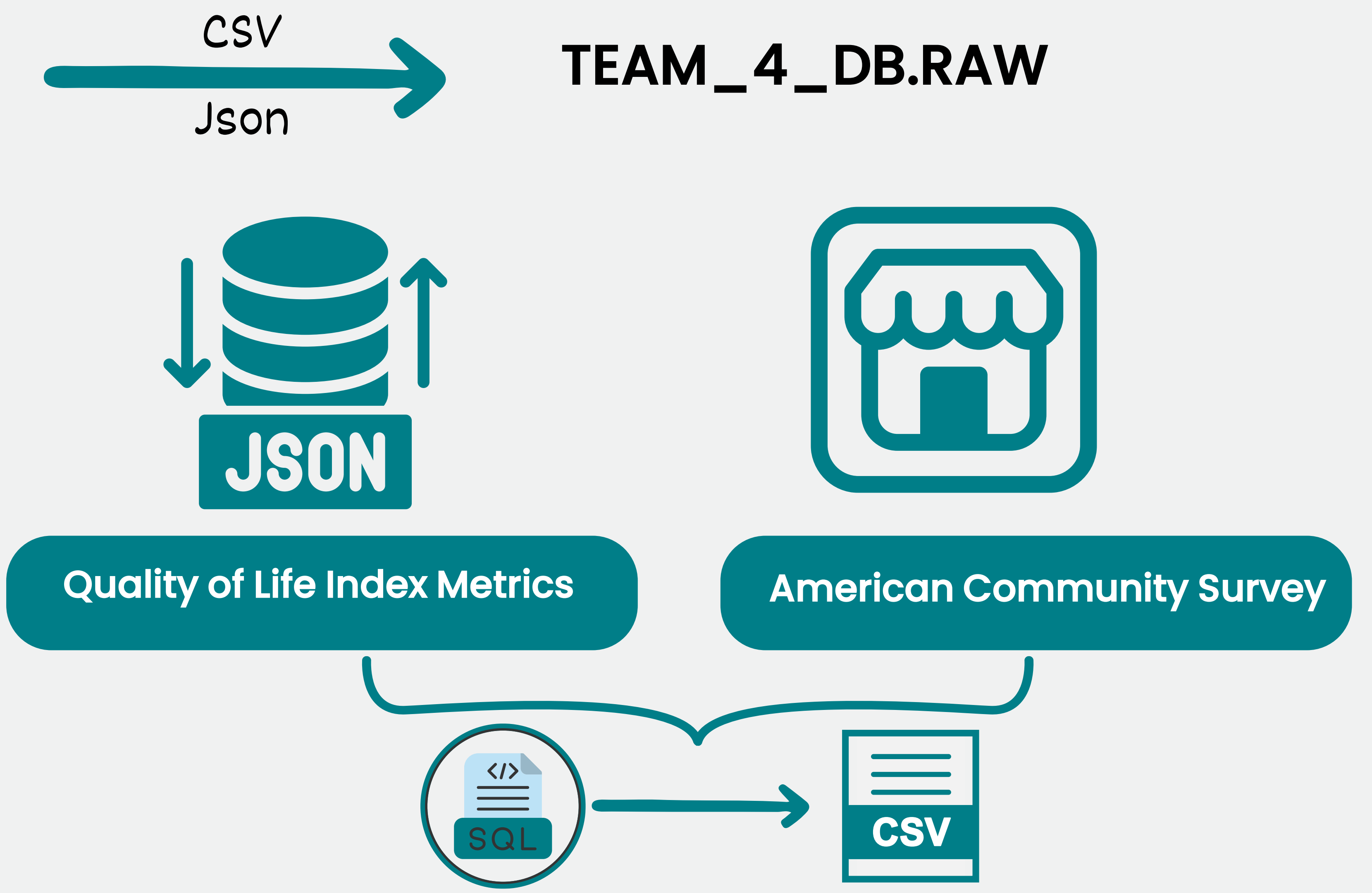
American Community Survey

Kaggl
e

Kaggl
e

Marketplac
e

Load

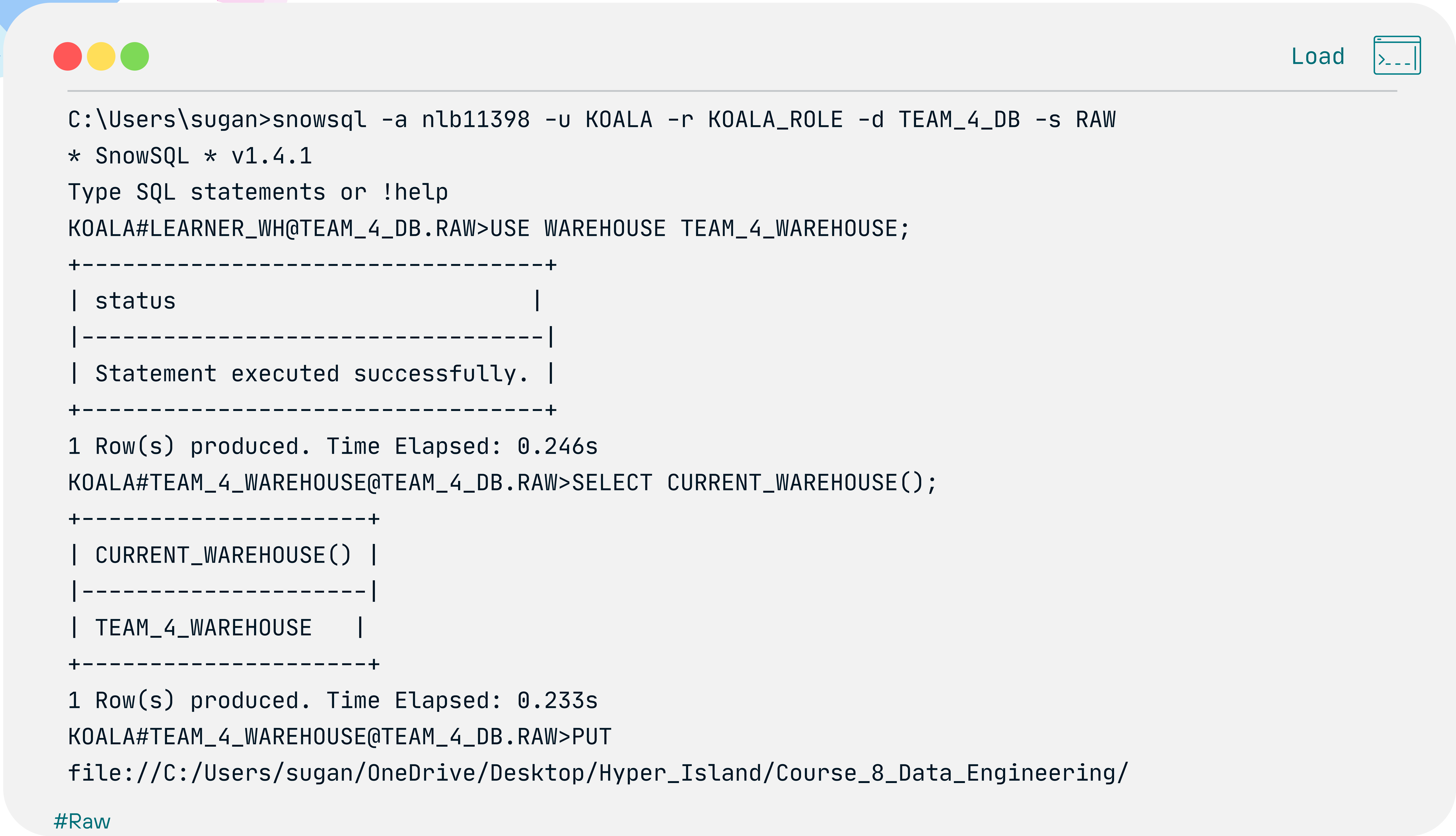


CSV → TEAM_4_DB.RAW
Json

Housing rental information data

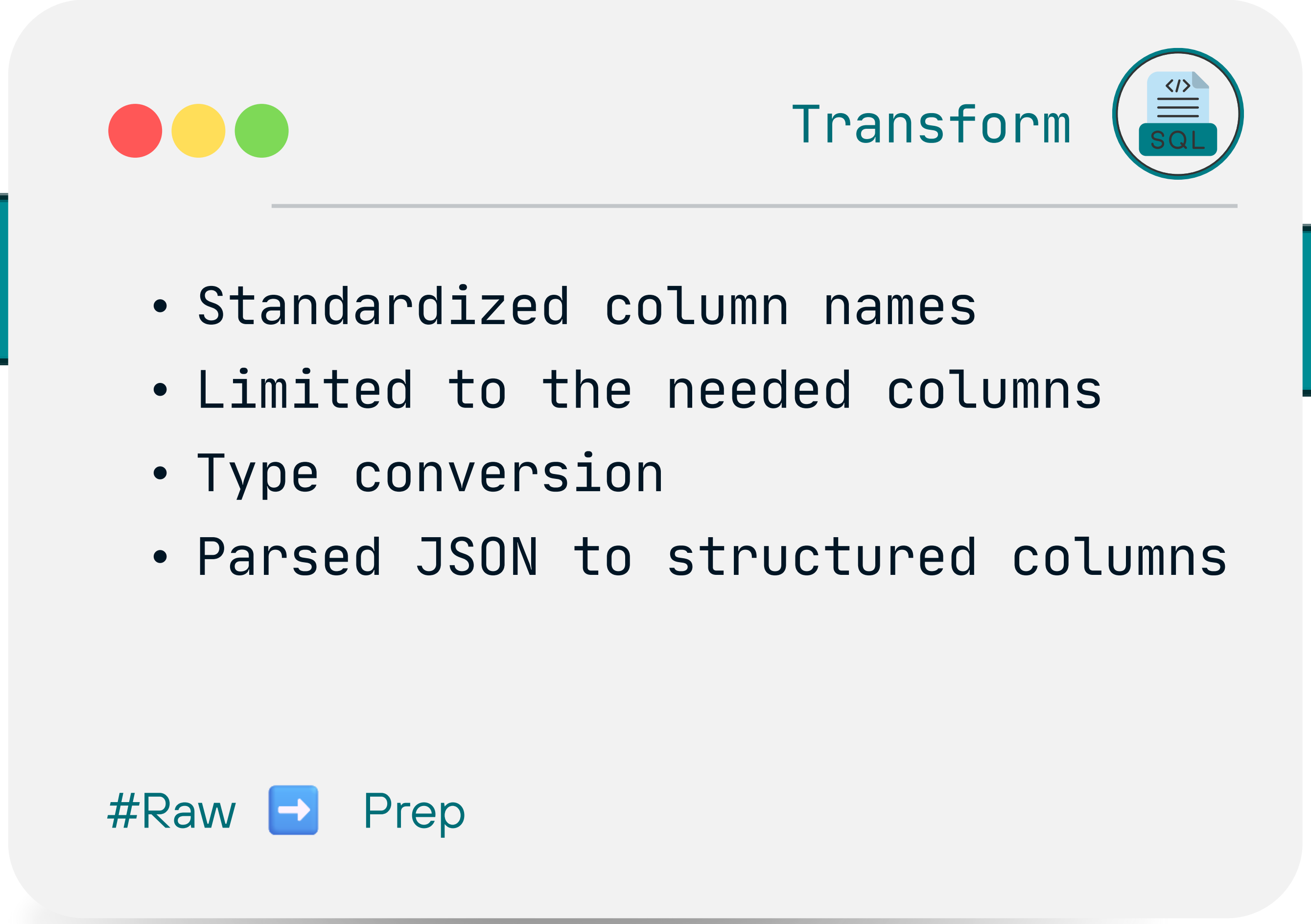
Quality of Life Index Metrics

American Community Survey



```
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/
```

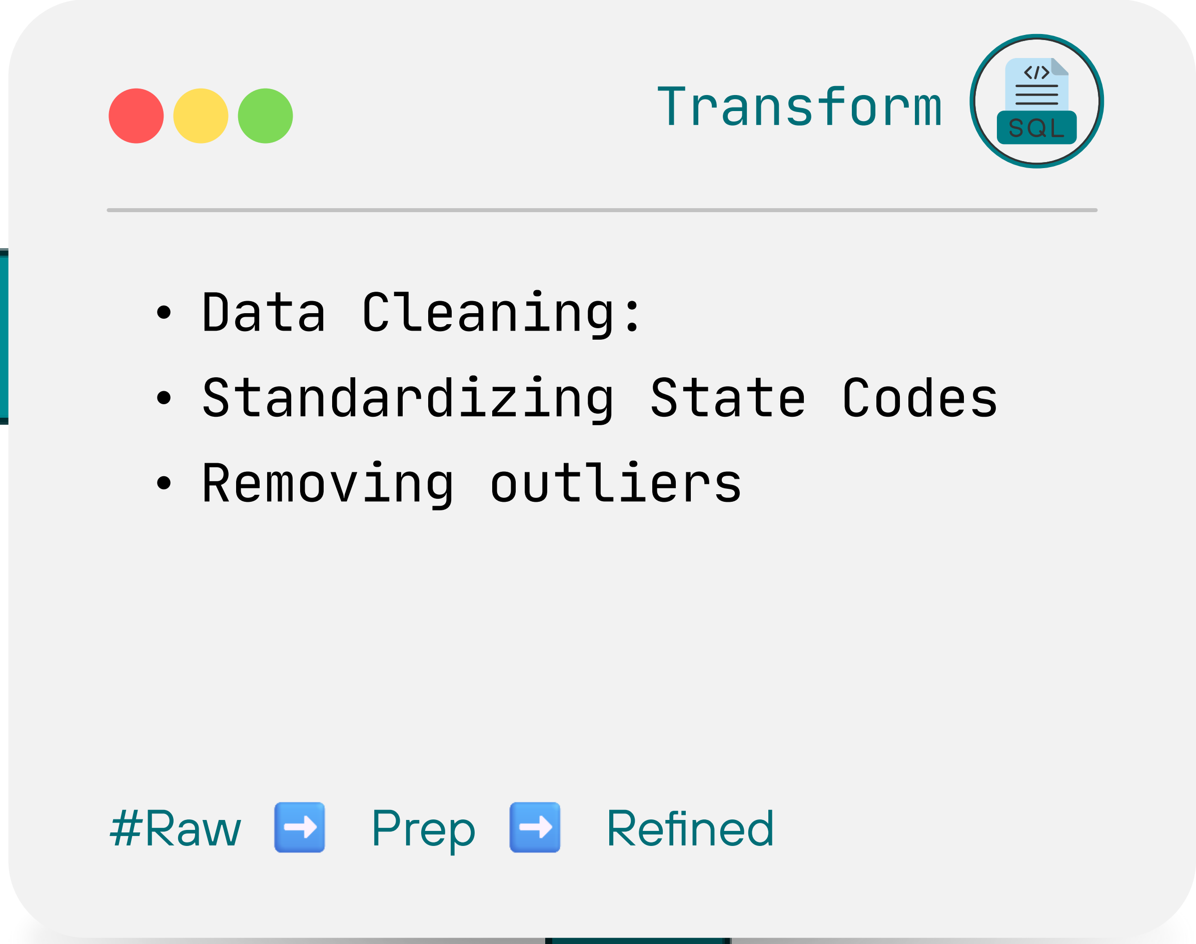
#Raw



Transform

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

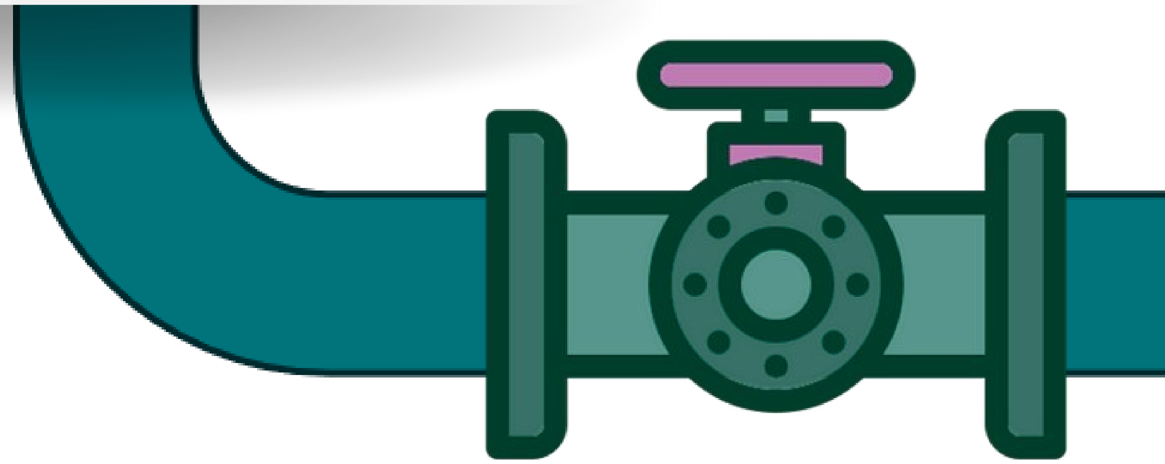
#Raw ➡ Prep



Transform

- Data Cleaning:
- Standardizing State Codes
- Removing outliers

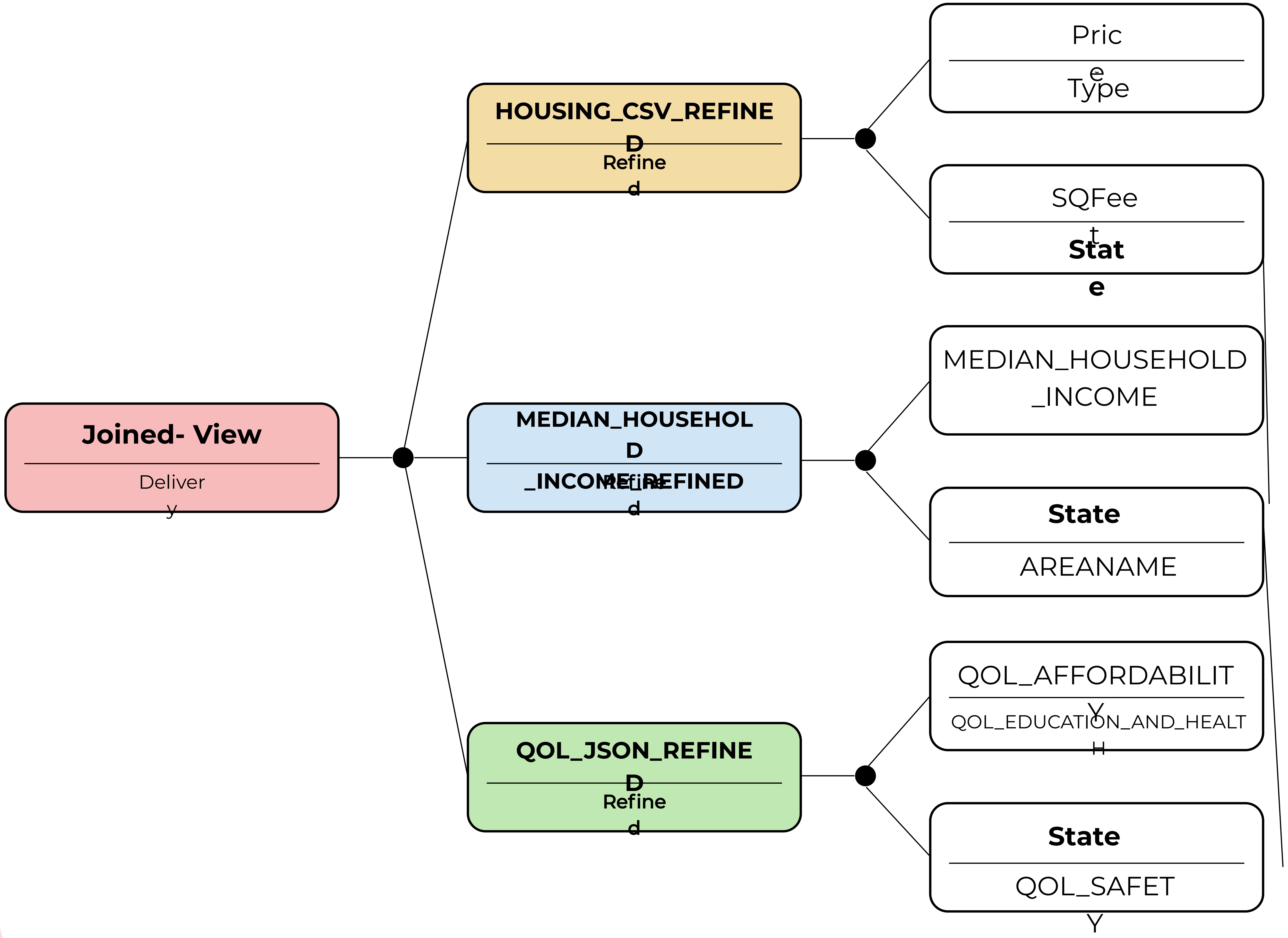
#Raw ➡ Prep ➡ Refined



Dashboard ❄

- Using Team_4_Viewer_role to create dashboard
- From refined schema joined 3 tables together for analysis.

#Raw ➡ Prep ➡ Refined ➡ Delivery



Joined- View

Delivery

HOUSING_CSV_REFINED

Refined

MEDIAN_HOUSEHOLD_INCOME_REFINED

Refined

QOL_JSON_REFINED

Refined

Price

Type

SQFeet

State

MEDIAN_HOUSEHOLD_INCOME

State

AREANAME

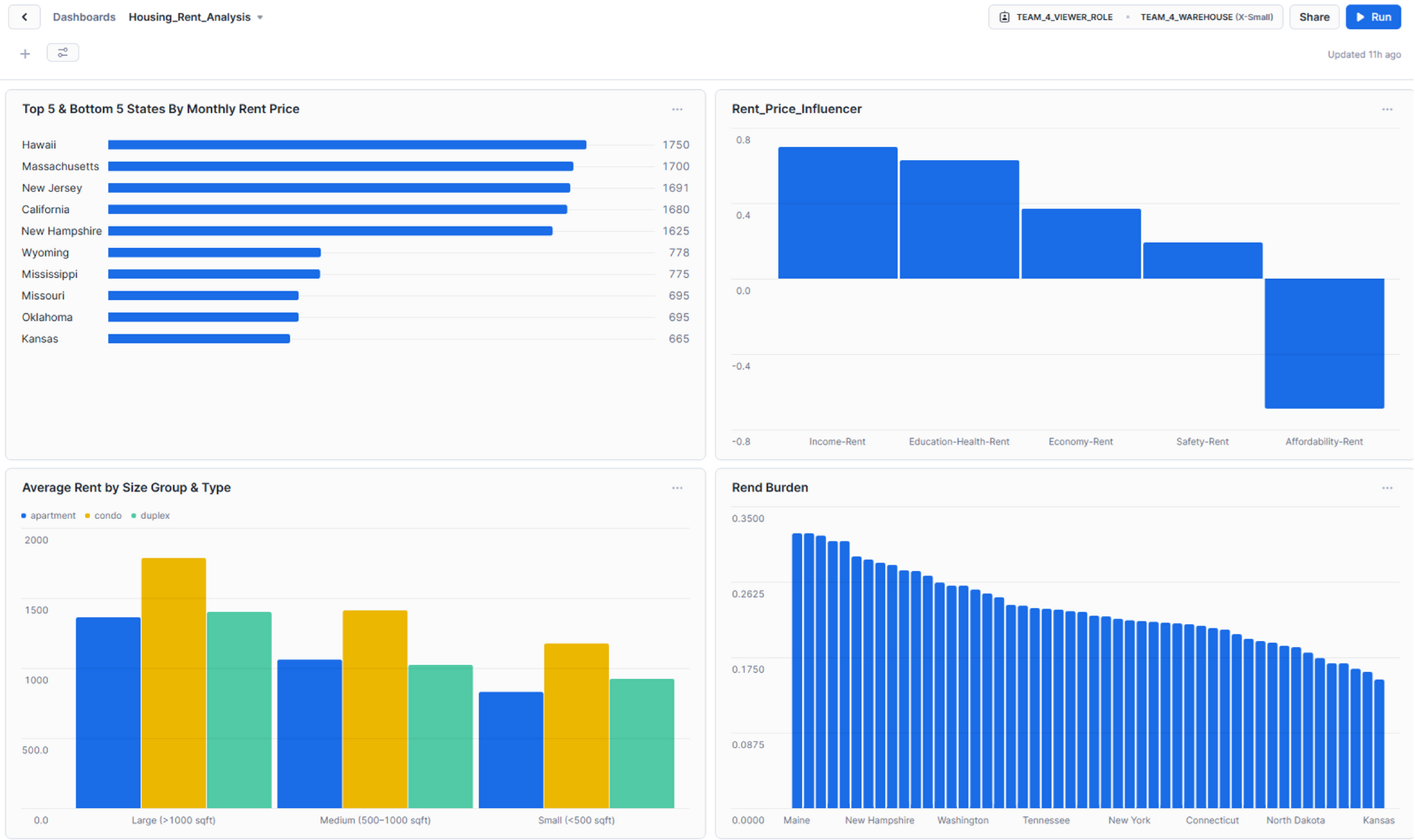
QOL_AFFORDABILITY

QOL_EDUCATION_AND_HEALTH

State

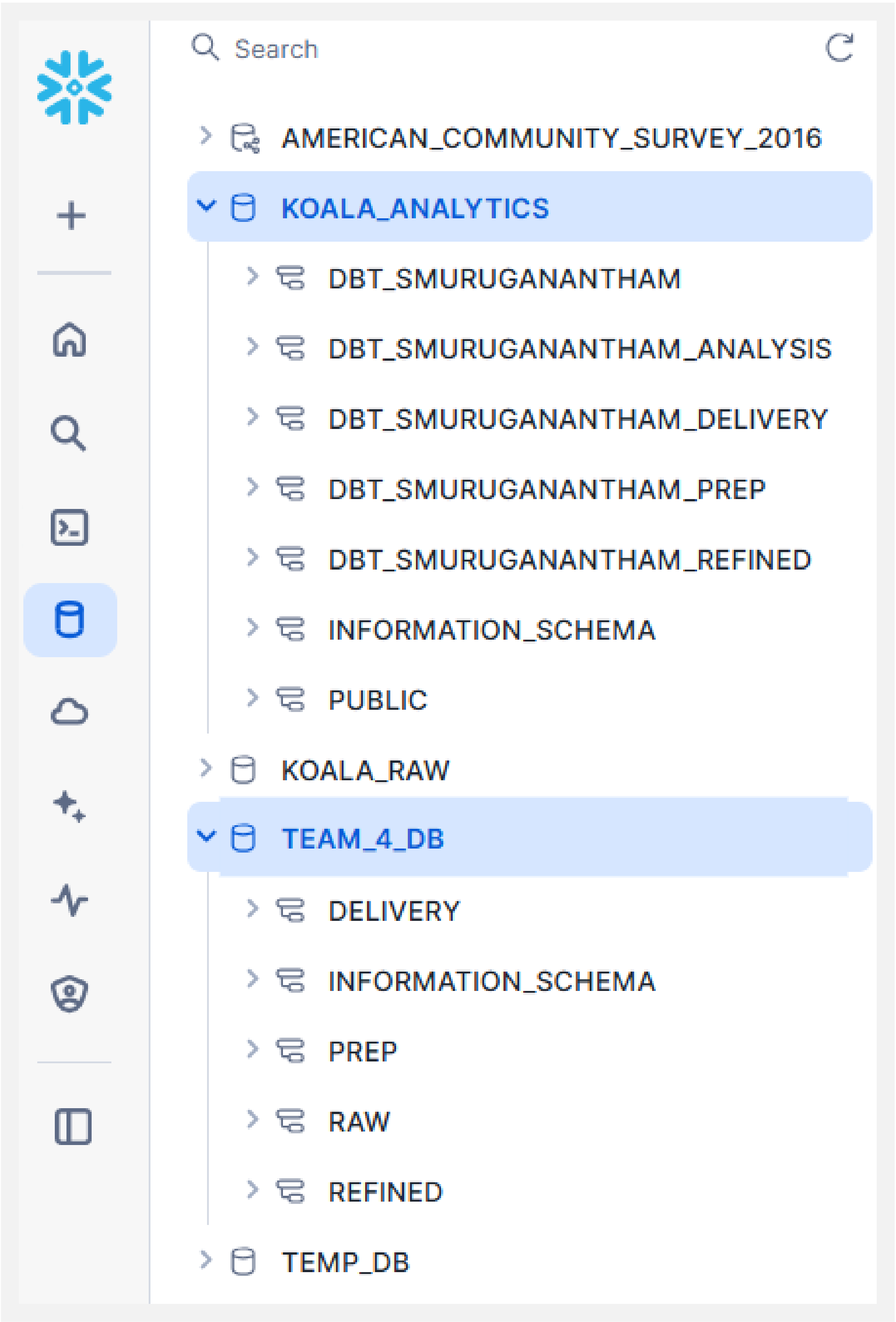
QOL_SAFETY

Snowflake Dashboard



Integrated pipeline

dbt
snowflake



Search

- AMERICAN_COMMUNITY_SURVEY_2016
- KOALA_ANALYTICS
  - DBT_SMURUGANANTHAM
  - DBT_SMURUGANANTHAM_ANALYSIS
  - DBT_SMURUGANANTHAM_DELIVERY
  - DBT_SMURUGANANTHAM_PREP
  - DBT_SMURUGANANTHAM_REFINED
  - INFORMATION_SCHEMA
  - PUBLIC
- KOALA_RAW
- TEAM_4_DB
  - DELIVERY
  - INFORMATION_SCHEMA
  - PREP
  - RAW
  - REFINED
- TEMP_DB

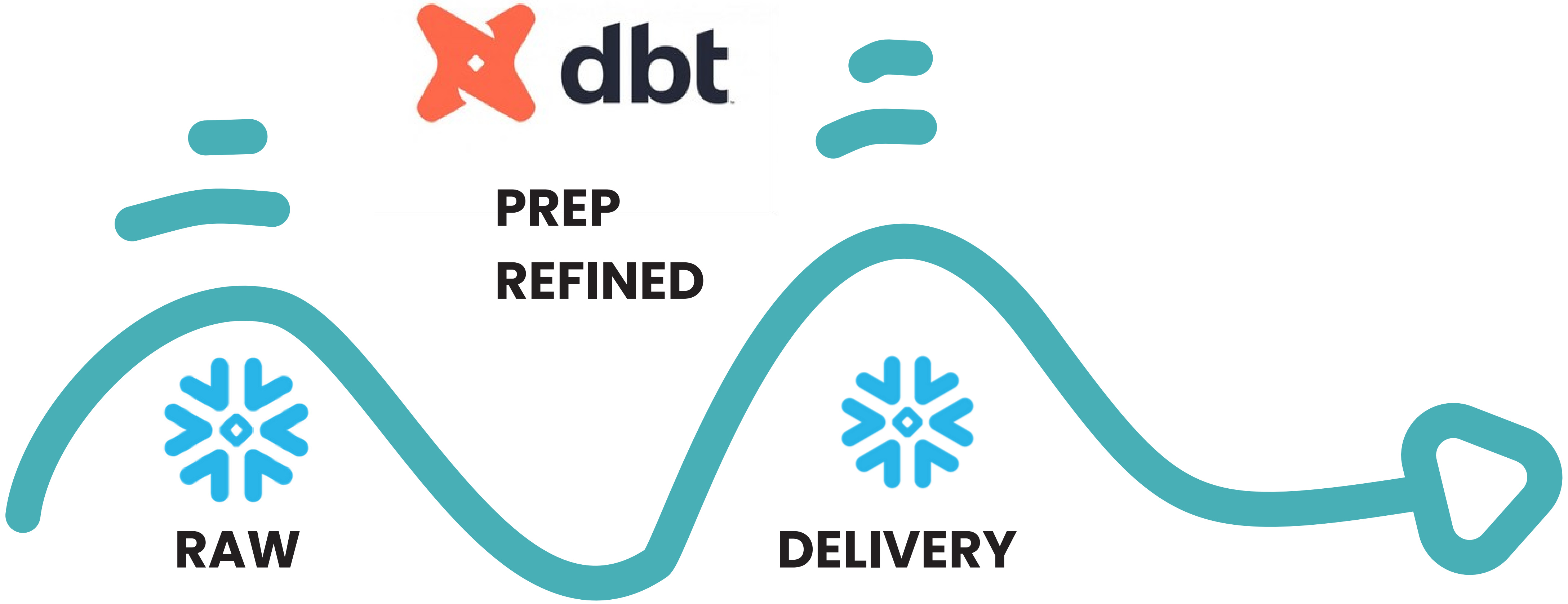
dbt Snowflake Pipeline Architecture

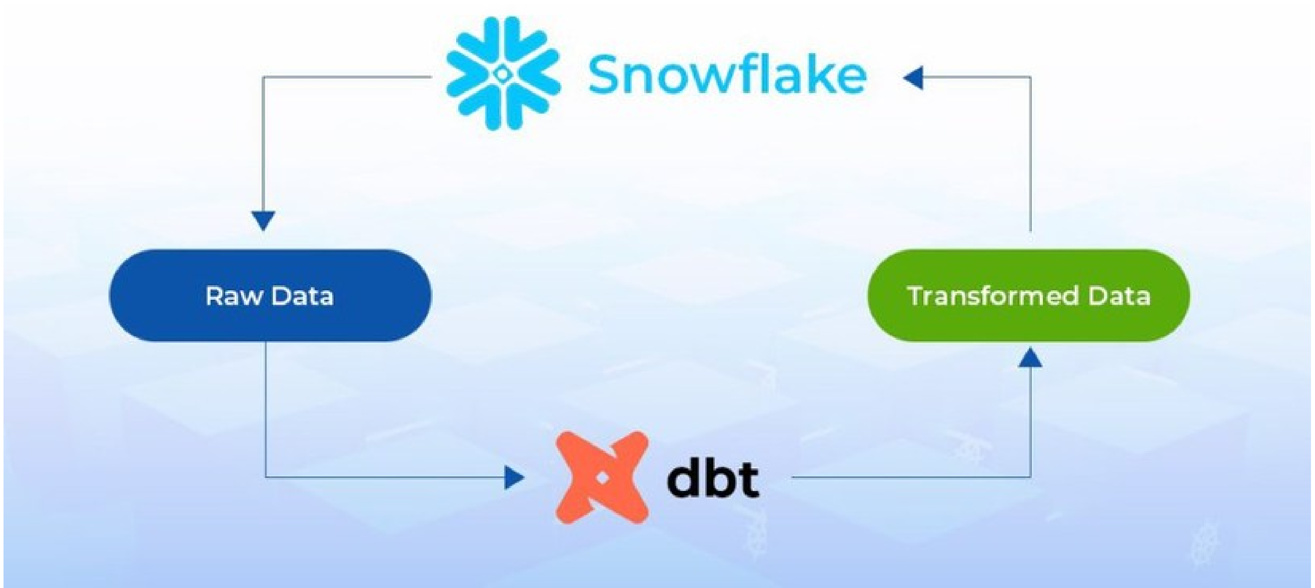
- Transformation
- Data Tests

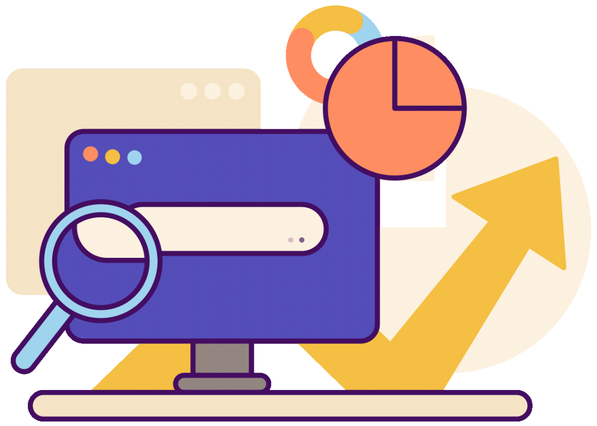
CSV

JSON

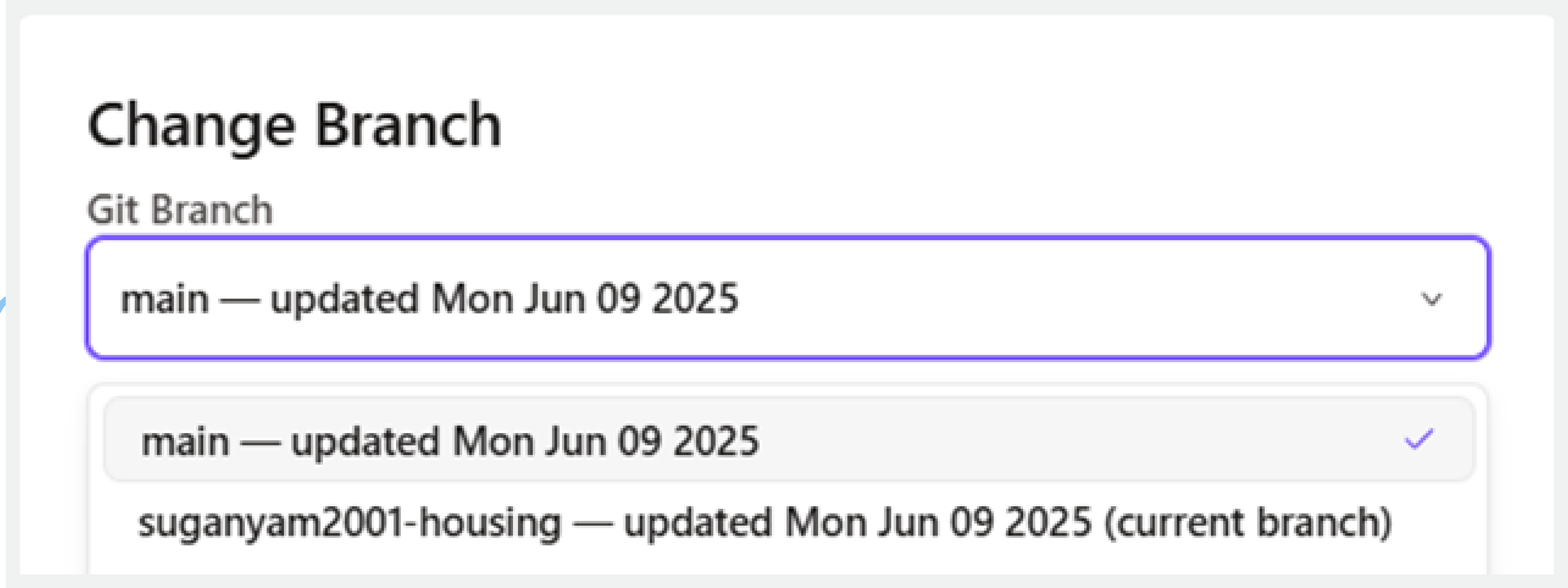
Marketplace







GitHub collaboration

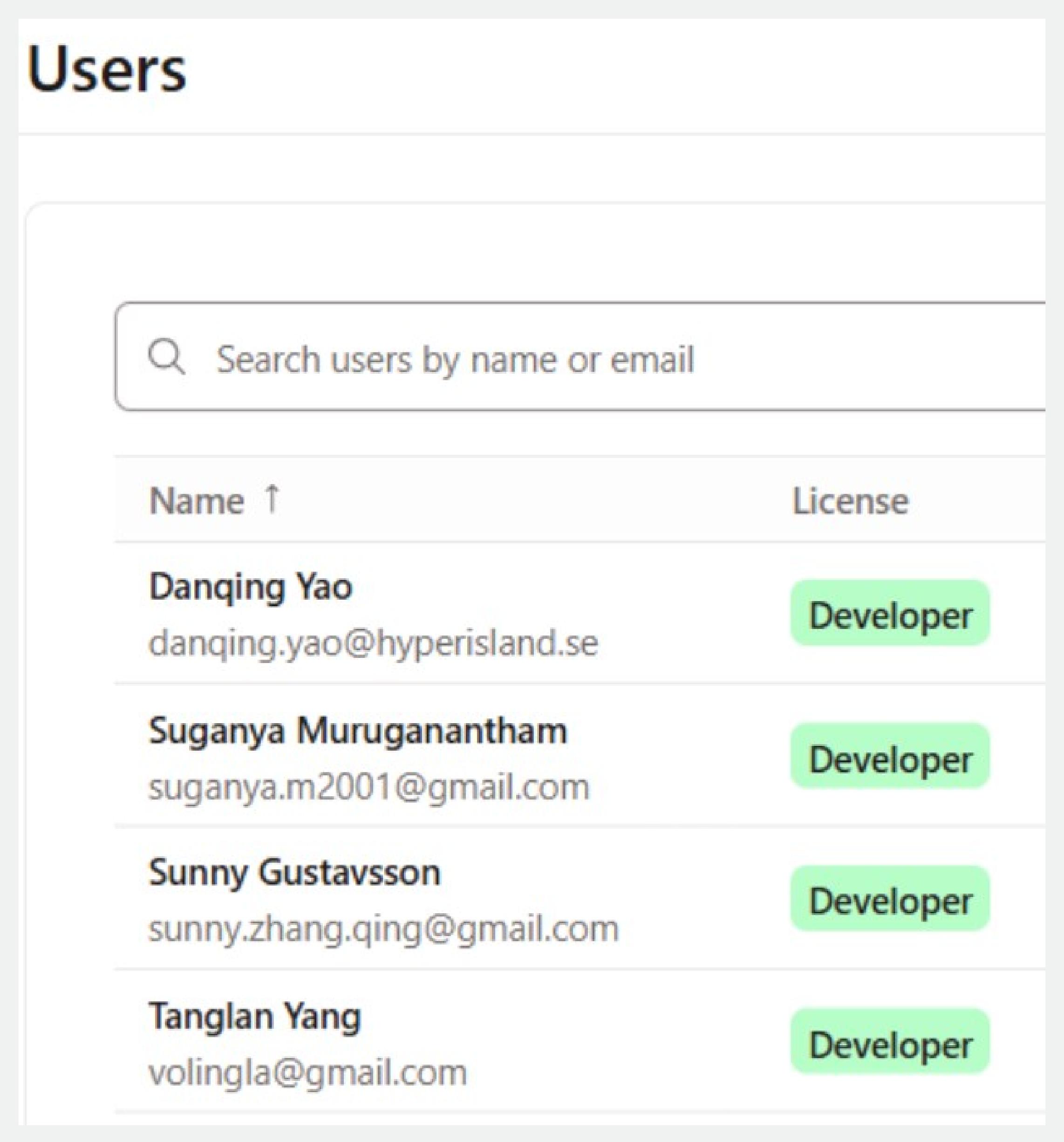


Change Branch

Git Branch

main — updated Mon Jun 09 2025

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



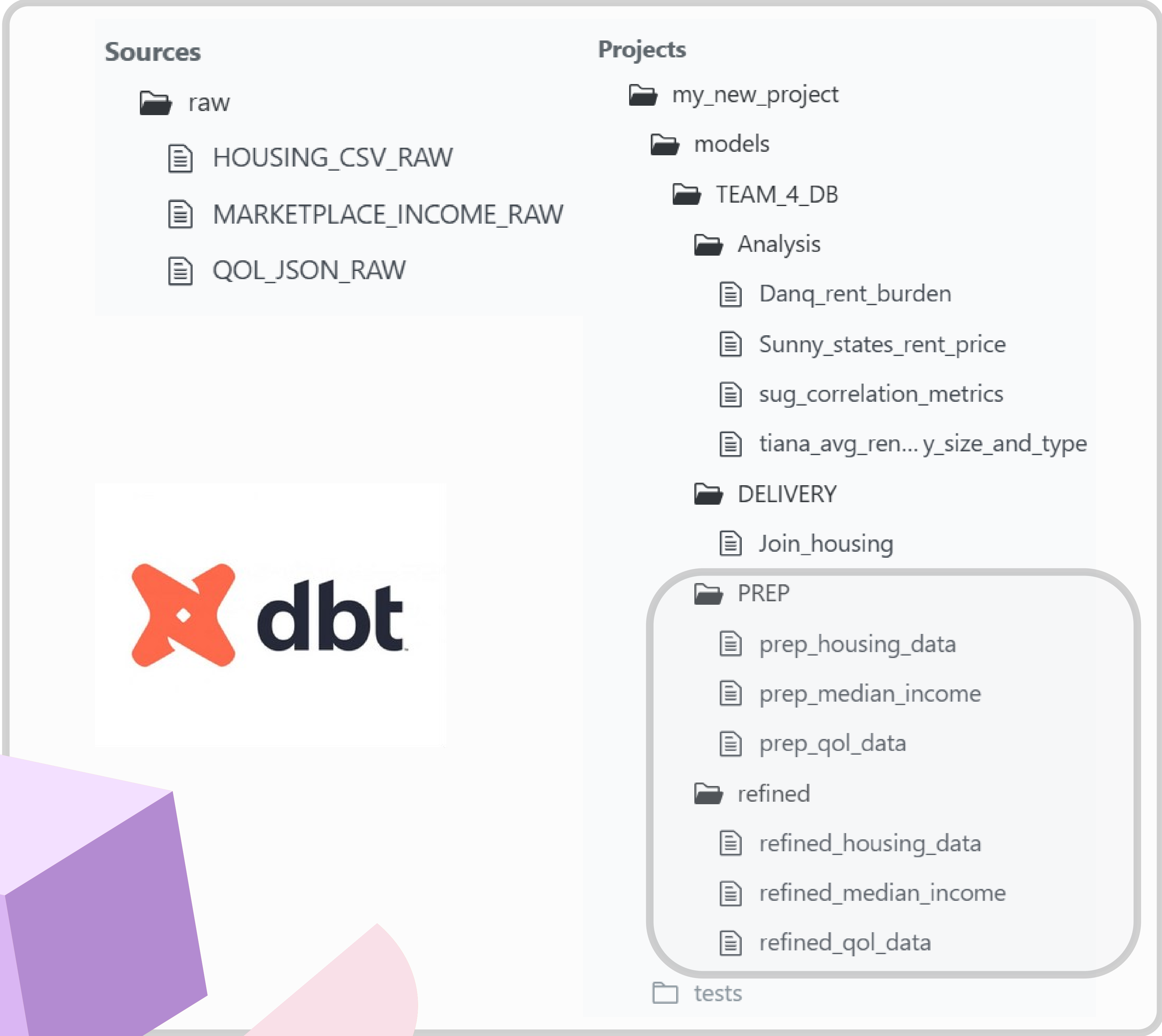
Users

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

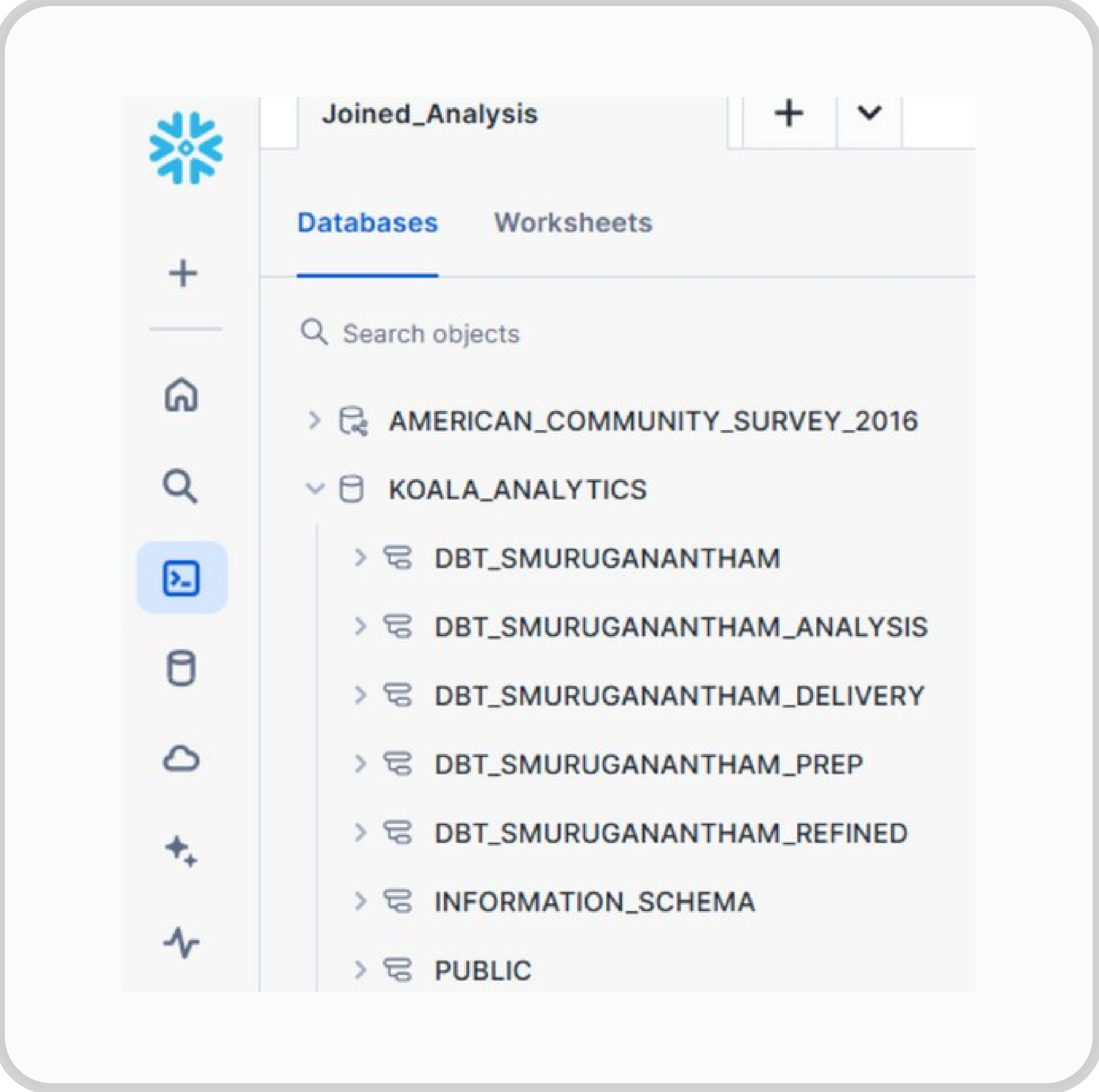


Sources

- 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_ren... y_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



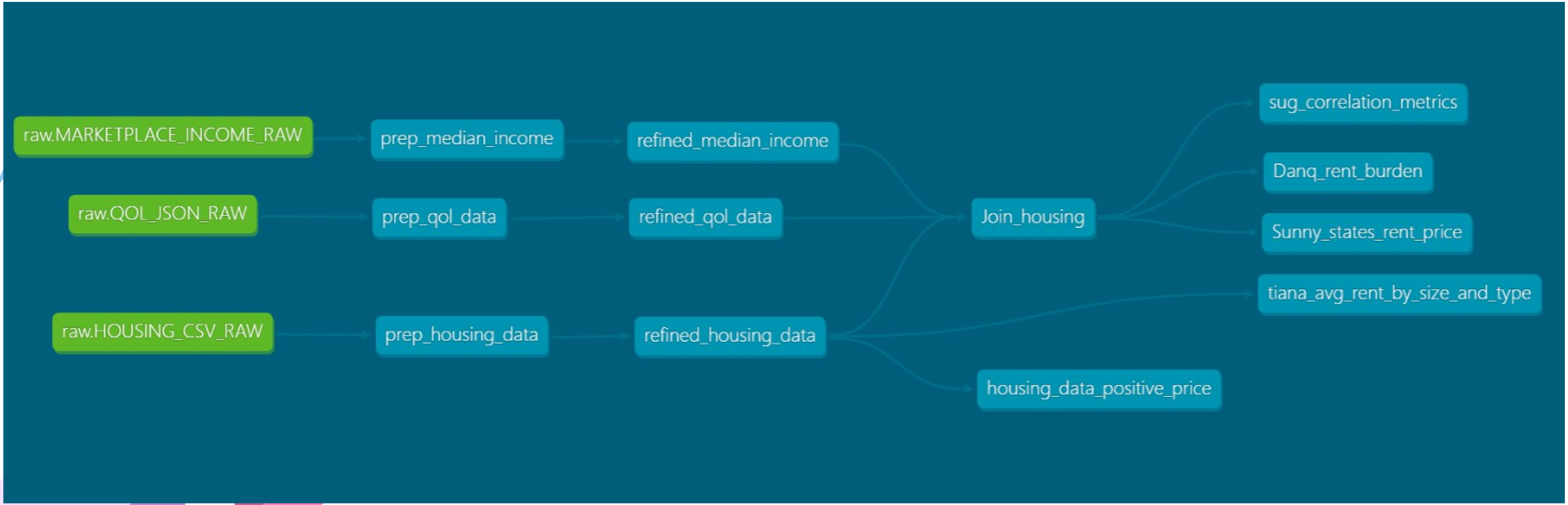
Joined_Analysis

Databases Worksheets

Search objects

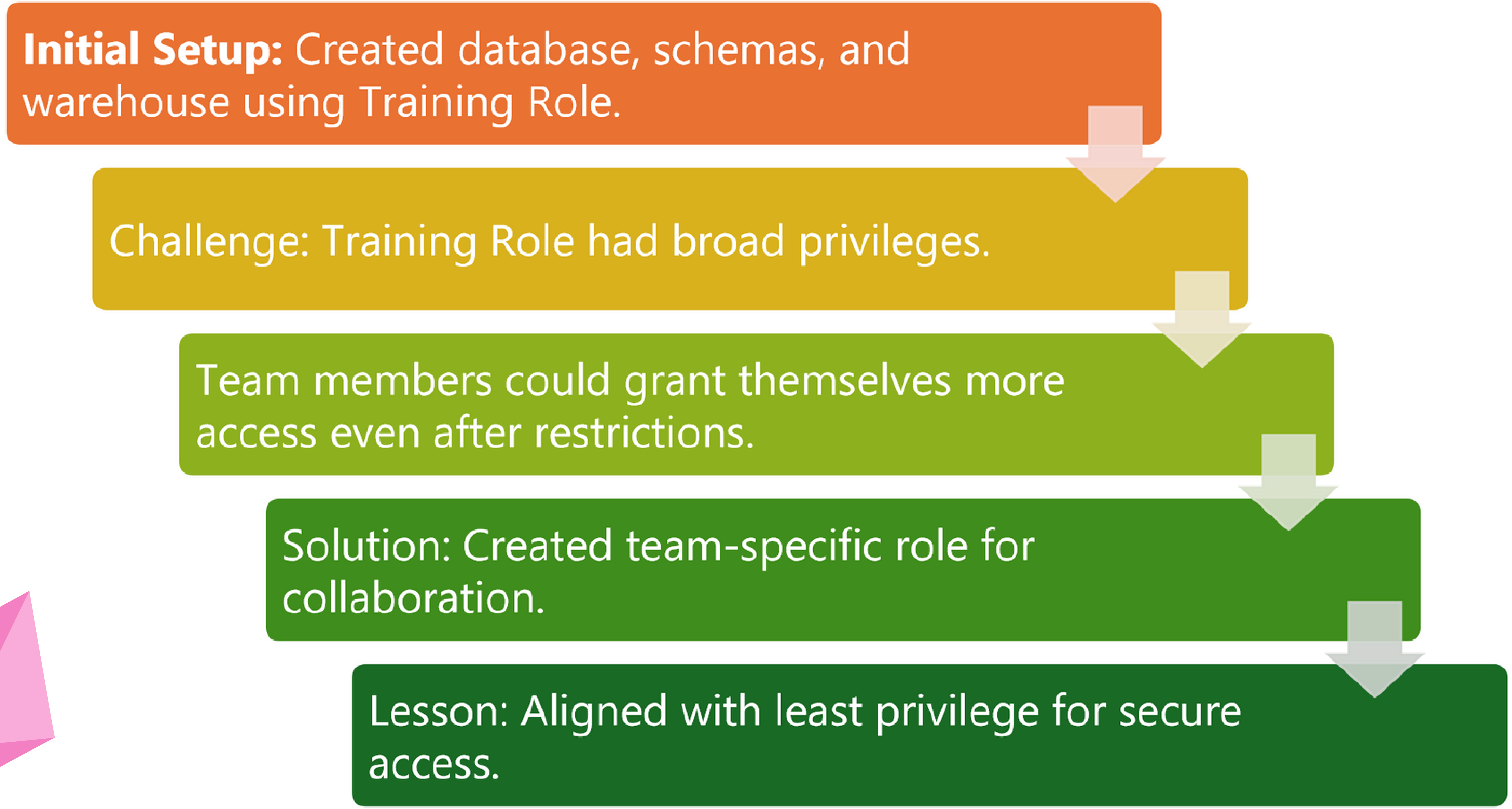
- AMERICAN_COMMUNITY_SURVEY_2016
- KOALA_ANALYTICS
  - DBT_SMURUGANANTHAM
  - DBT_SMURUGANANTHAM_ANALYSIS
  - DBT_SMURUGANANTHAM_DELIVERY
  - DBT_SMURUGANANTHAM_PREP
  - DBT_SMURUGANANTHAM_REFINED
  - INFORMATION_SCHEMA
  - PUBLIC

DAG


raw.MARKETPLACE_INCOME_RAW
raw.QOL_JSON_RAW
raw.HOUSING_CSV_RAW
prep_median_income
prep_qol_data
prep_housing_data
refined_median_income
refined_qol_data
refined_housing_data
Join_housing
housing_data_positive_price
sug_correlation_metrics
Danq_rent_burden
Sunny_states_rent_price
tiana_avg_rent_by_size_and_type

Retrospective

RBAC (Role-Based Access Control) Implementation



Initial Setup: Created database, schemas, and warehouse using Training Role.

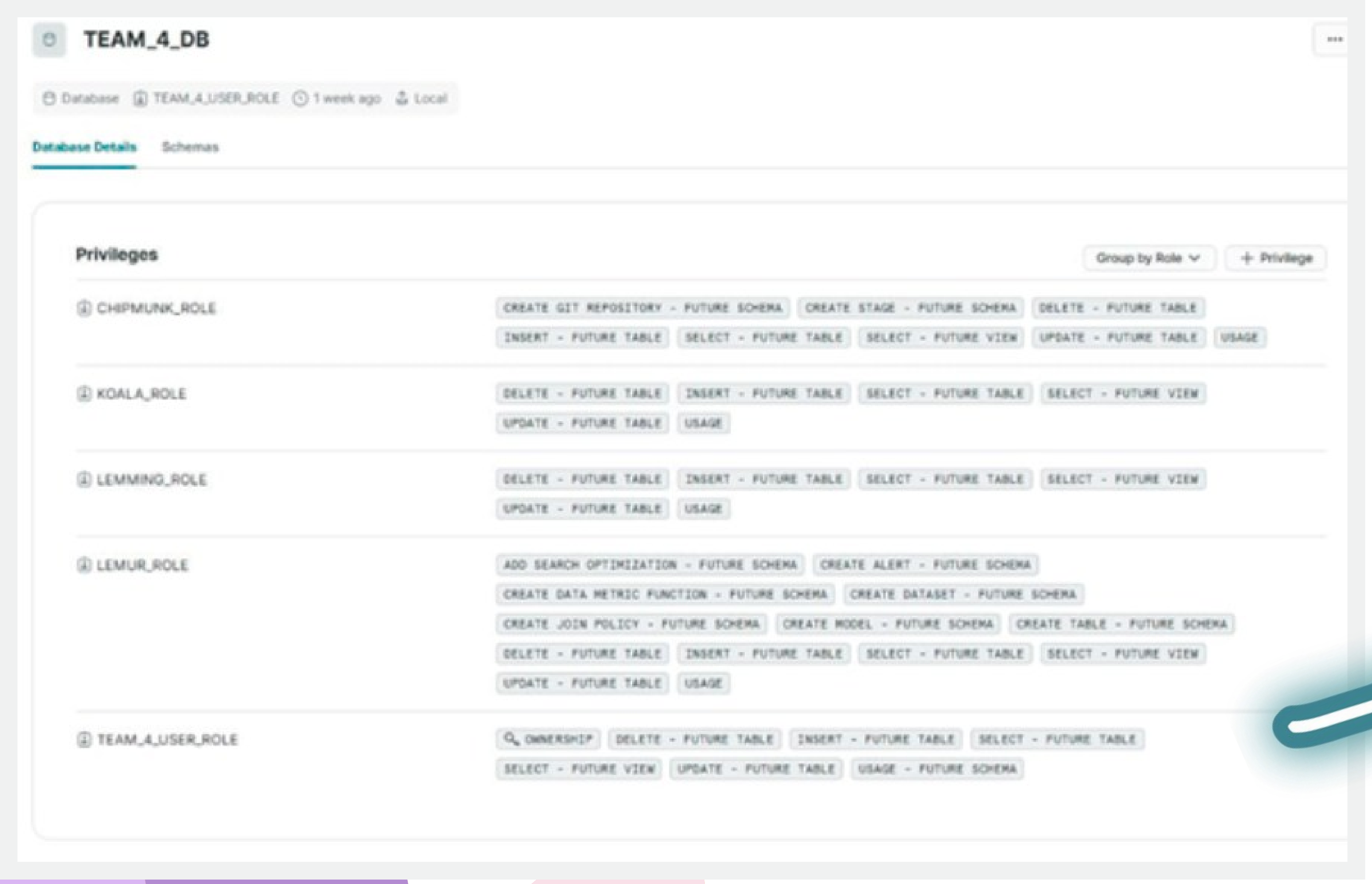
Challenge: Training Role had broad privileges.

Team members could grant themselves more access even after restrictions.

Solution: Created team-specific role for collaboration.

Lesson: Aligned with least privilege for secure access.

RBAC (Role-Based Access Control) Implementation



TEAM_4_DB

Database · TEAM_4_USER_ROLE · 1 week ago · Local

Database Details · Schemas

Privileges

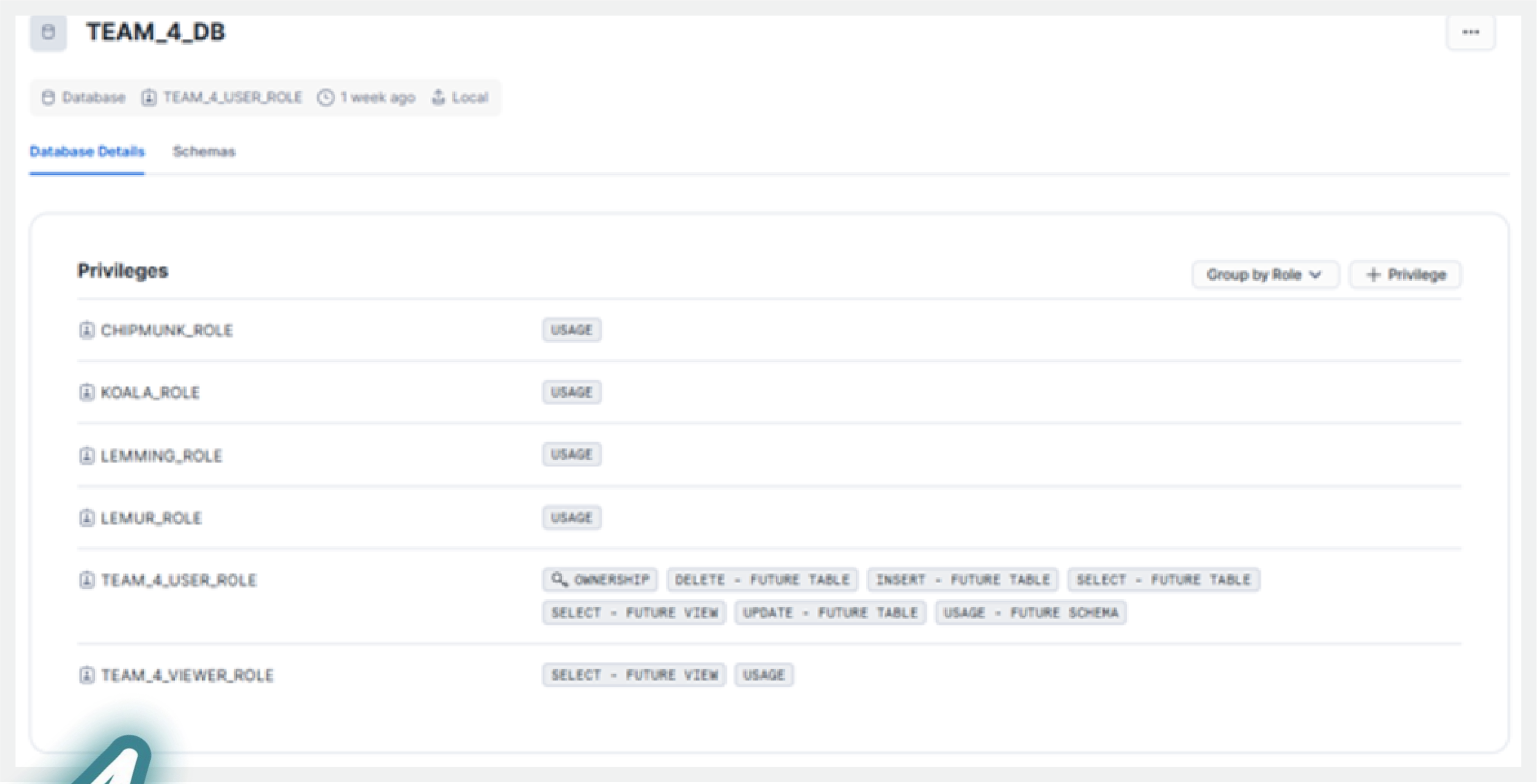
CHIPMUNK_ROLE	CREATE GIT REPOSITORY - FUTURE SCHEMA
	CREATE STAGE - FUTURE SCHEMA
	DELETE - FUTURE TABLE
	INSERT - FUTURE TABLE
	SELECT - FUTURE TABLE
	SELECT - FUTURE VIEW
	UPDATE - FUTURE TABLE
	USAGE

KOALA_ROLE	DELETE - FUTURE TABLE
	INSERT - FUTURE TABLE
	SELECT - FUTURE TABLE
	SELECT - FUTURE VIEW
	UPDATE - FUTURE TABLE
	USAGE

LEMMING_ROLE	DELETE - FUTURE TABLE
	INSERT - FUTURE TABLE
	SELECT - FUTURE TABLE
	SELECT - FUTURE VIEW
	UPDATE - FUTURE TABLE
	USAGE

LEMUR_ROLE	ADD SEARCH OPTIMIZATION - FUTURE SCHEMA
	CREATE ALERT - FUTURE SCHEMA
	CREATE DATA METRIC FUNCTION - FUTURE SCHEMA
	CREATE DATASET - FUTURE SCHEMA
	CREATE JOIN POLICY - FUTURE SCHEMA
	CREATE MODEL - FUTURE SCHEMA
	CREATE TABLE - FUTURE SCHEMA
	DELETE - FUTURE TABLE
	INSERT - FUTURE TABLE
	SELECT - FUTURE TABLE
	SELECT - FUTURE VIEW
	UPDATE - FUTURE TABLE
	USAGE

TEAM_4_USER_ROLE	OWNERSHIP
	DELETE - FUTURE TABLE
	INSERT - FUTURE TABLE
	SELECT - FUTURE TABLE
	SELECT - FUTURE VIEW
	UPDATE - FUTURE TABLE
	USAGE - FUTURE SCHEMA



TEAM_4_DB

Database · TEAM_4_USER_ROLE · 1 week ago · Local

Database Details · Schemas

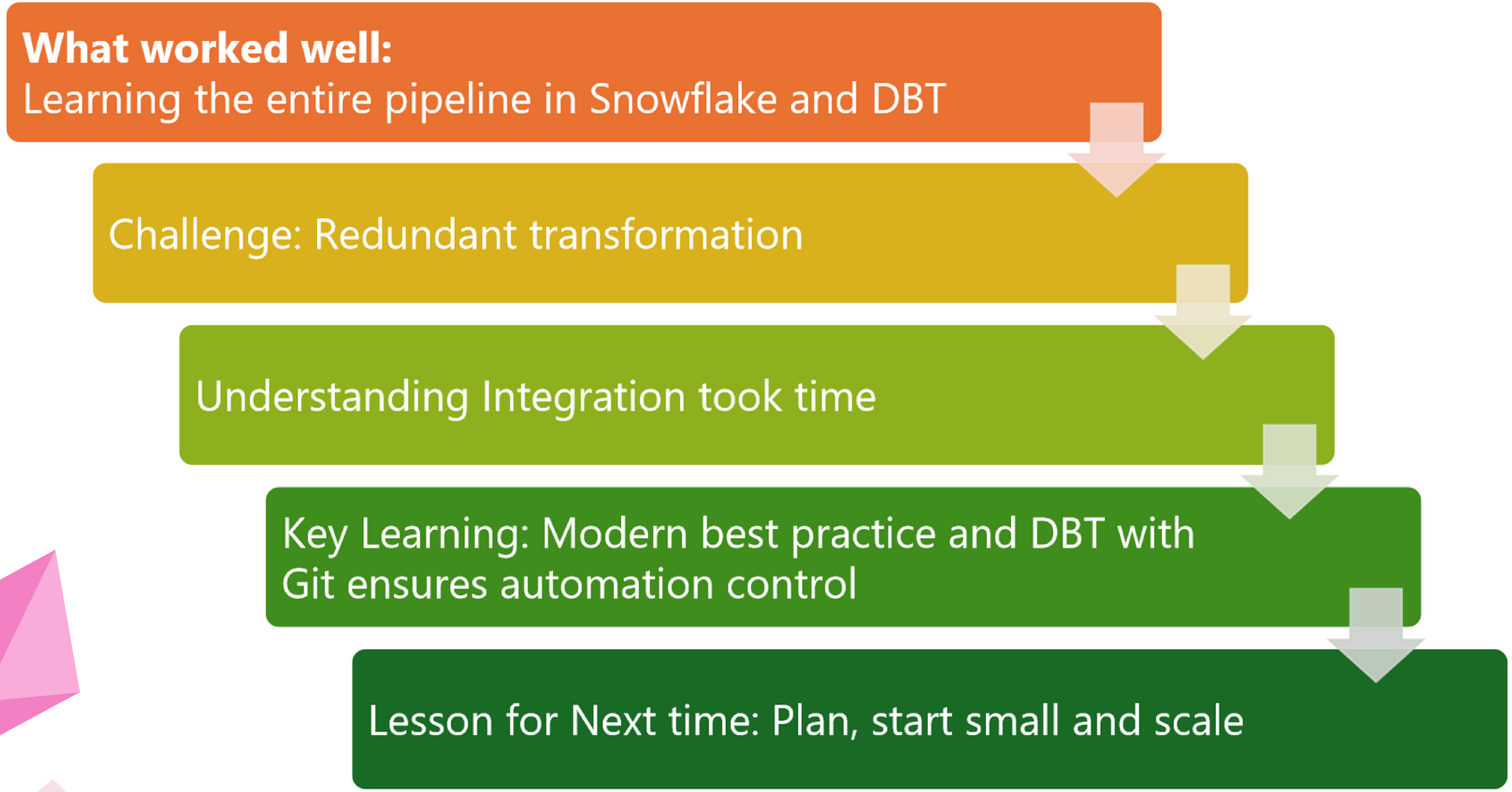
Privileges

CHIPMUNK_ROLE	USAGE
KOALA_ROLE	USAGE
LEMMING_ROLE	USAGE
LEMUR_ROLE	USAGE

TEAM_4_USER_ROLE	OWNERSHIP
	DELETE - FUTURE TABLE
	INSERT - FUTURE TABLE
	SELECT - FUTURE TABLE
	SELECT - FUTURE VIEW
	UPDATE - FUTURE TABLE
	USAGE - FUTURE SCHEMA

TEAM_4_VIEWER_ROLE	SELECT - FUTURE VIEW USAGE

Building an End-to-End Data Pipeline in Snowflake and DBT Integrated



What worked well:
Learning the entire pipeline in Snowflake and DBT

Challenge: Redundant transformation

Understanding Integration took time

Key Learning: Modern best practice and DBT with Git ensures automation control

Lesson for Next time: Plan, start small and scale