# Data Engineer Bootcamp

# Capstone Project 1

Presented by Bador Alsharif and Sami Alfifi

# Agenda

## Capstone Project 1

01 Project Objectives

02 Data Overview

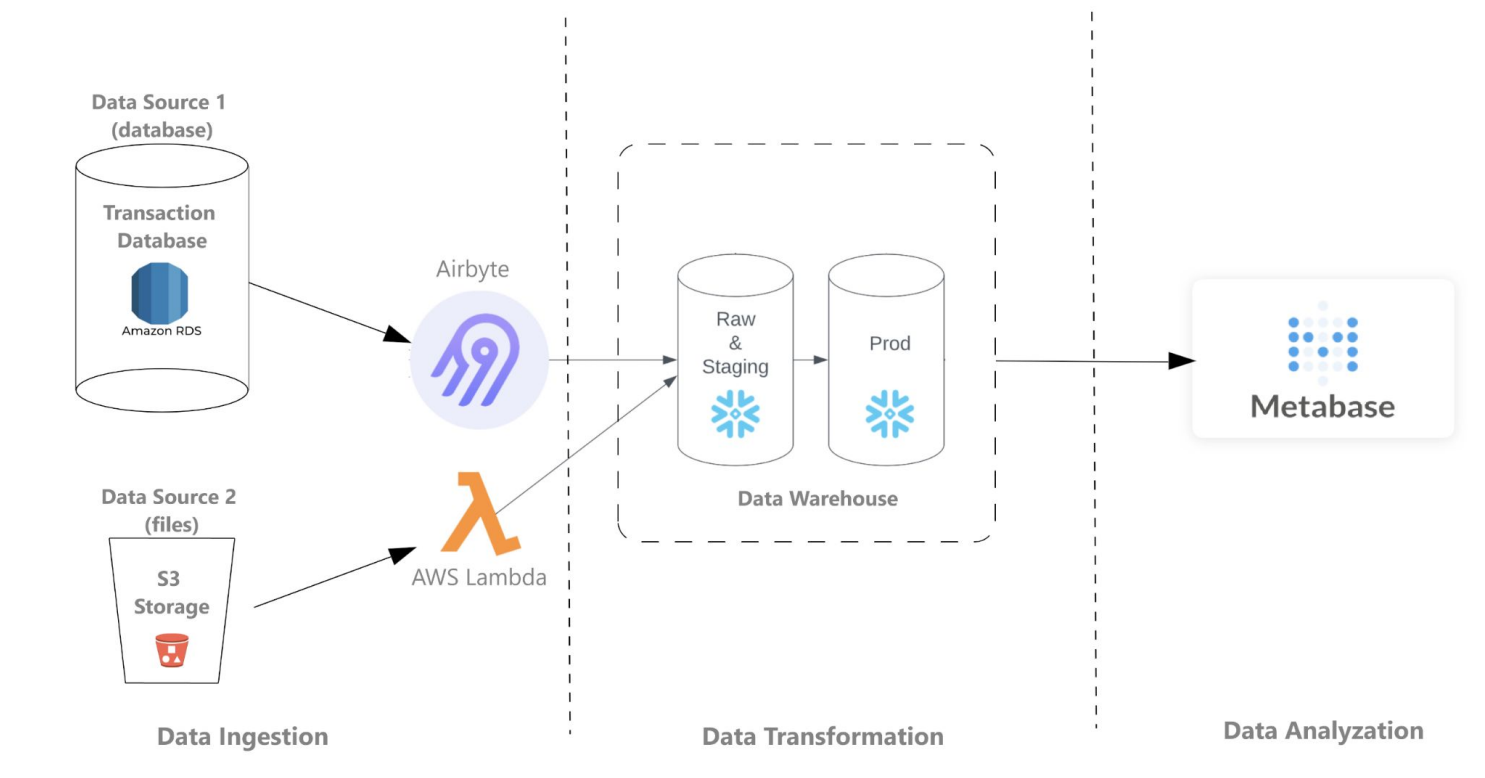
03 Project Steps

04 Project Achievement

05 Future

WeCloudData

# Project Objectives




WeCloudData

# Data Overview

- TPCDS a well-known dataset designed for database testing

- In TPC-DS, fact tables store quantitative data like sales, while dimension tables contain descriptive attributes like customer details.

| Fact tables | Dimension tables |
|---|---|
| Catalog_Sales | Date_Dim |
| Web_Sales | Customer |
| Inventory | Item |
| | Promotion |
| | Customer_Demographics |
| | Call_Center |
| | Customer_Address |
| | Catalog_Page |
| | Warehouse |
| | Time_Dim |
| | Ship_Mode |
| | Household_Demographics |
| | Income_Band |
| | Web_page |
| | Web_Site |
| | |

WeCloudData

# Project Steps

01 AWS Lambda Function

02 Airbyte Installation

03 Data Modeling

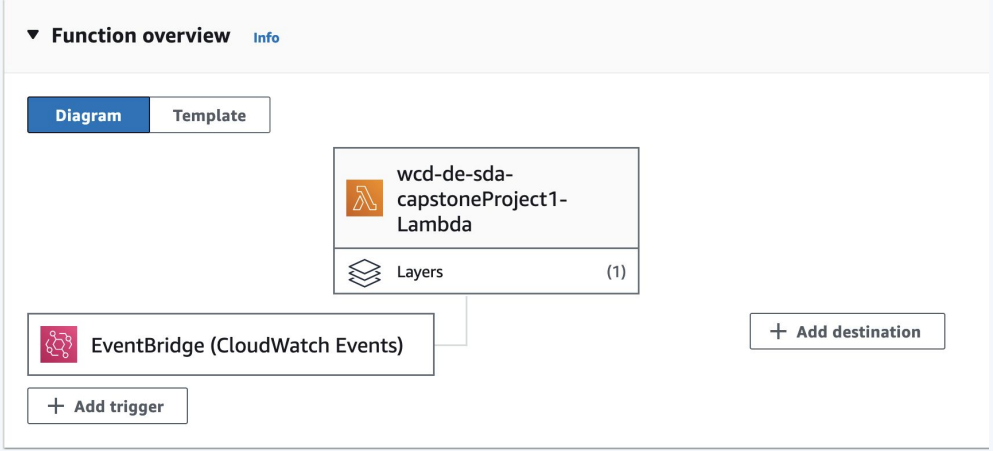
04 ETL and Data Loading

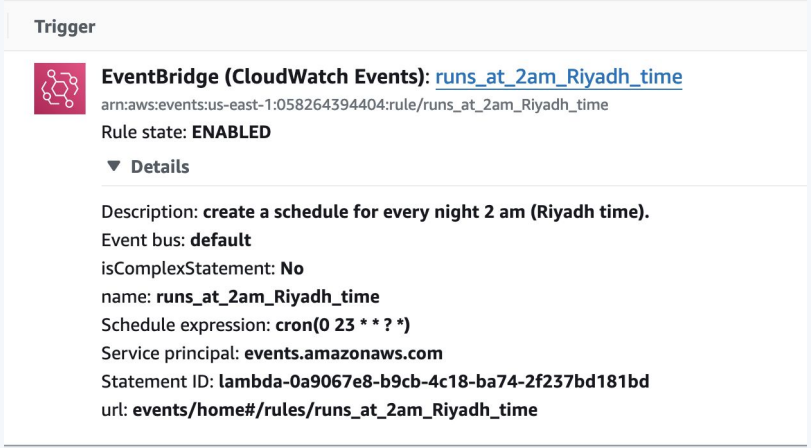
05 Metabase

WeCloudData

# AWS Lambda Function

1. download 'inventory.csv' from a public AWS S3 URL.
2. upload the 'inventory.csv' file to the Snowflake table.
3. create a schedule for every night 2 am (Riyadh time).

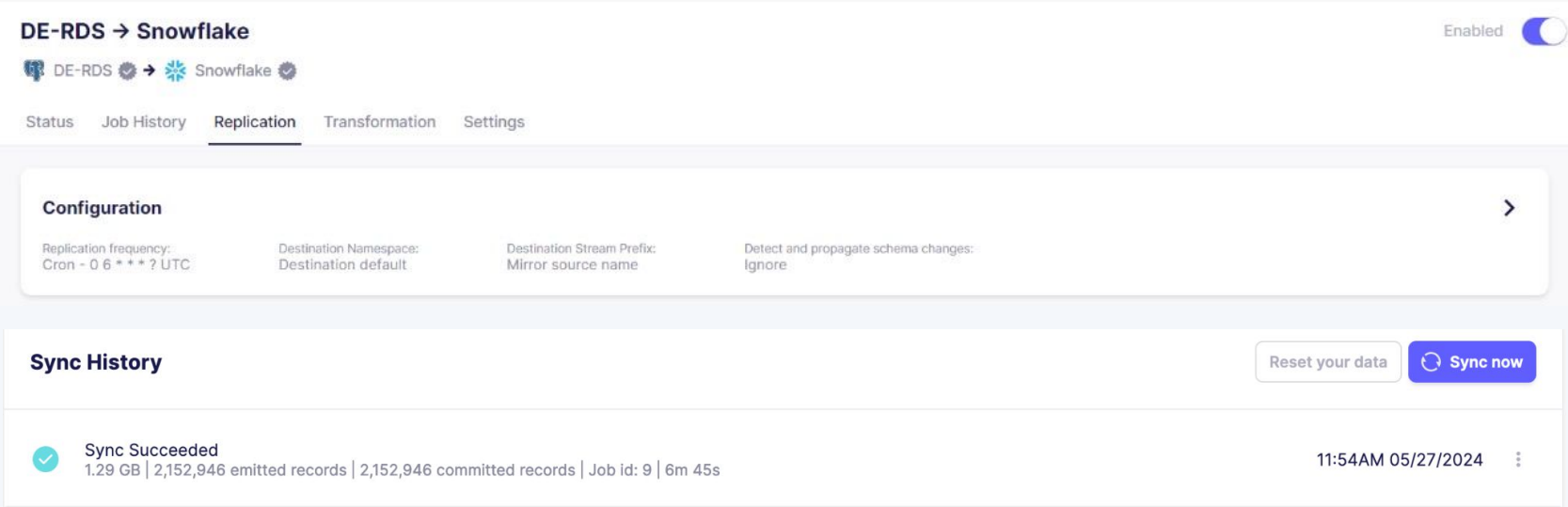



WeCloudData

# Airbyte Installation

1. Install Airbyte on EC2 instance.
2. Connect to the Postgres database in AWS RDS.
3. Create connections to data sources and set up replication pipelines.




WeCloudData

# Data Modeling

1. Create an analytics schema within the Snowflake database to hold the data model.
2. Develop an integrated customer dimension table from multiple related tables.
3. Design a weekly fact table containing sales and inventory metrics.

- TPCDS
  - ANALYTICS
    - Tables
      - CUSTOMER_DIM
      - DATE_DIM
      - ITEM_DIM
      - WAREHOUSE_DIM
      - WEEKLY_SALES_INVENTORY

**CUSTOMER_DIM** 100K Rows

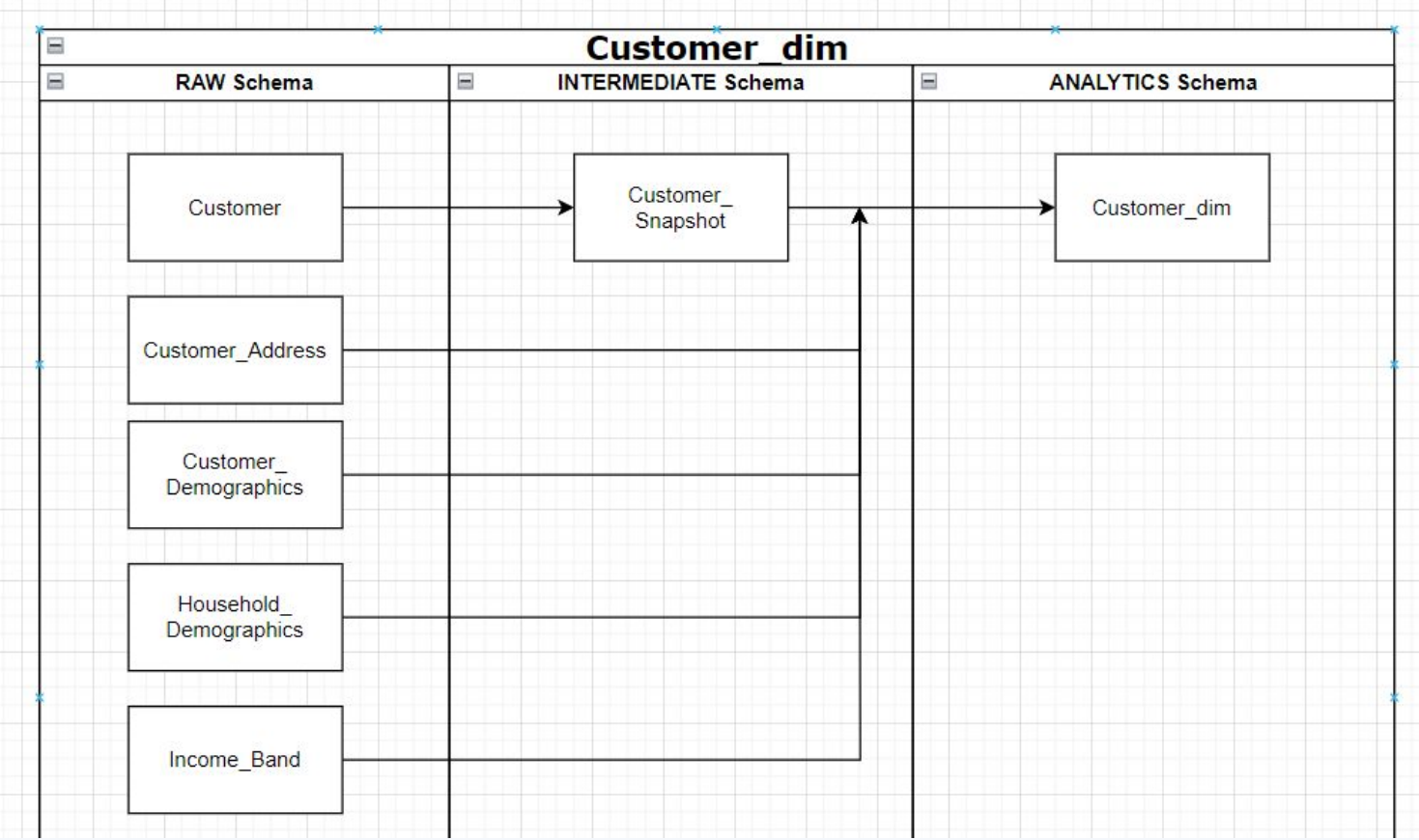
| | |
|---|---|
| C_SALUTATION | VARCHAR(16777216) |
| C_PREFERRED_CUST_F... | VARCHAR(16777216) |
| C_FIRST_SALES_DATE_SK | NUMBER(38,0) |
| C_CUSTOMER_SK | NUMBER(38,0) |
| C_LOGIN | VARCHAR(16777216) |
| C_CURRENT_CDEMO_SK | NUMBER(38,0) |
| C_FIRST_NAME | VARCHAR(16777216) |
| C_CURRENT_HDEMO_SK | NUMBER(38,0) |
| C_CURRENT_ADDR_SK | NUMBER(38,0) |
| C_LAST_NAME | VARCHAR(16777216) |
| C_CUSTOMER_ID | VARCHAR(16777216) |
| C_LAST_REVIEW_DATE_SK | NUMBER(38,0) |

**WEEKLY_SALES_INVENTORY** 1.2M Rows

| | |
|---|---|
| WAREHOUSE_SK | NUMBER(38,0) |
| ITEM_SK | NUMBER(38,0) |
| SOLD_WK_SK | NUMBER(38,0) |
| SOLD_WK_NUM | NUMBER(38,0) |
| SOLD_YR_NUM | NUMBER(38,0) |
| SUM_QTY_WK | NUMBER(38,0) |
| SUM_AMT_WK | FLOAT |
| SUM_PROFIT_WK | FLOAT |
| AVG_QTY_DY | NUMBER(38,6) |
| INV_QTY_WK | NUMBER(38,0) |
| WKS_SPLY | NUMBER(38,6) |
| LOW_STOCK_FLG_WK | BOOLEAN |

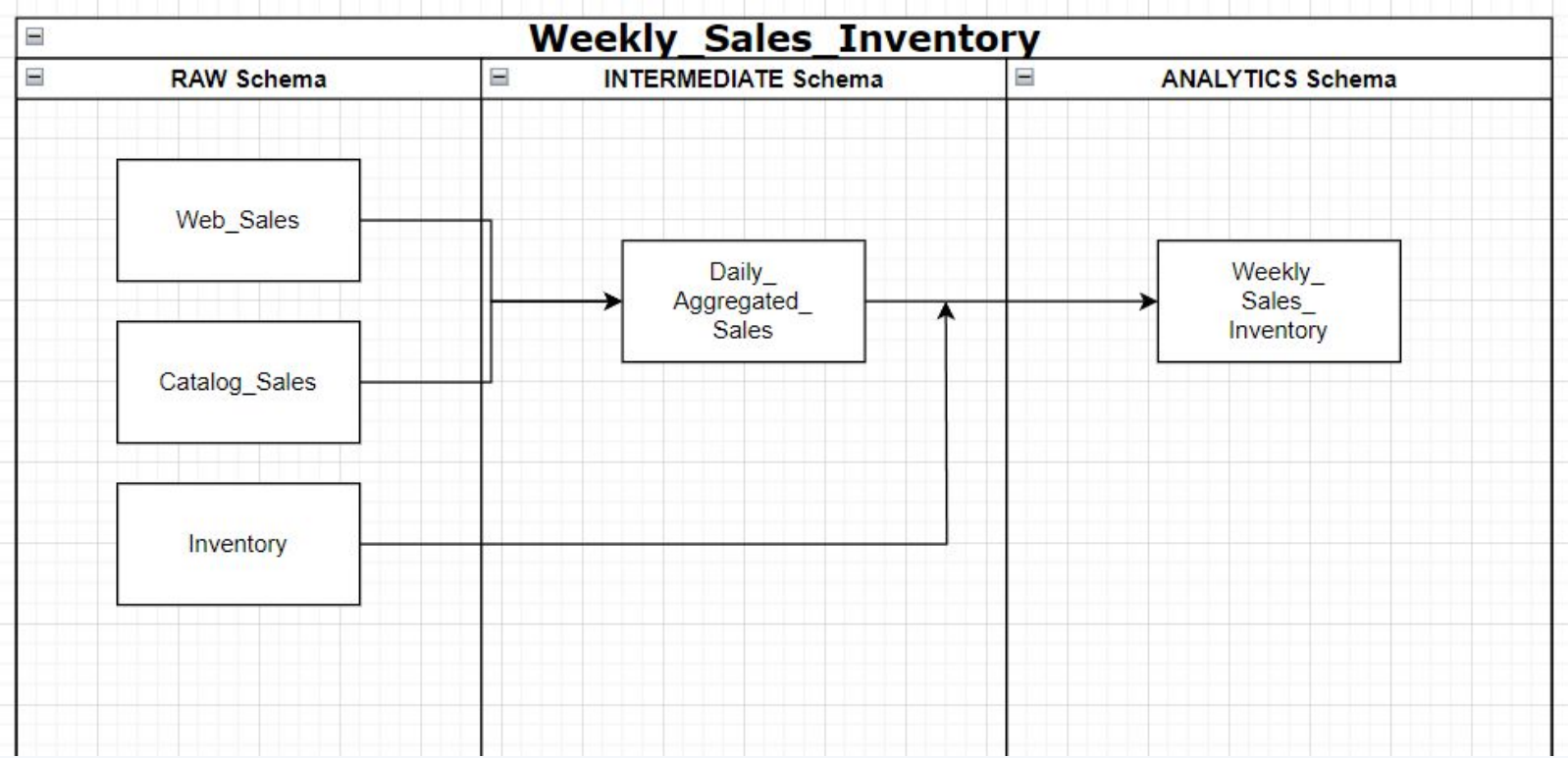
WeCloudData

# Data Modeling - (Customer Dimension)




WeCloudData

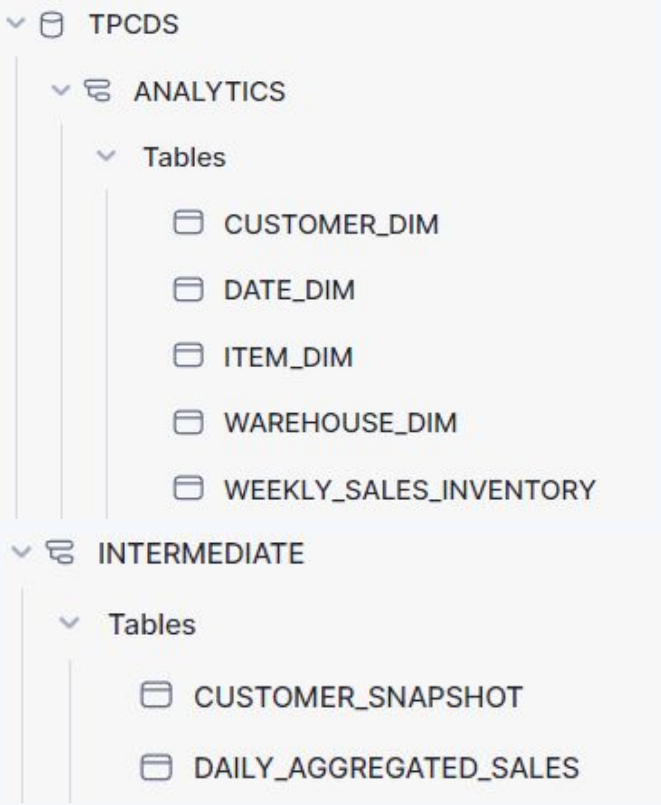
# Data Modeling - (weekly_sales_inventory)




WeCloudData

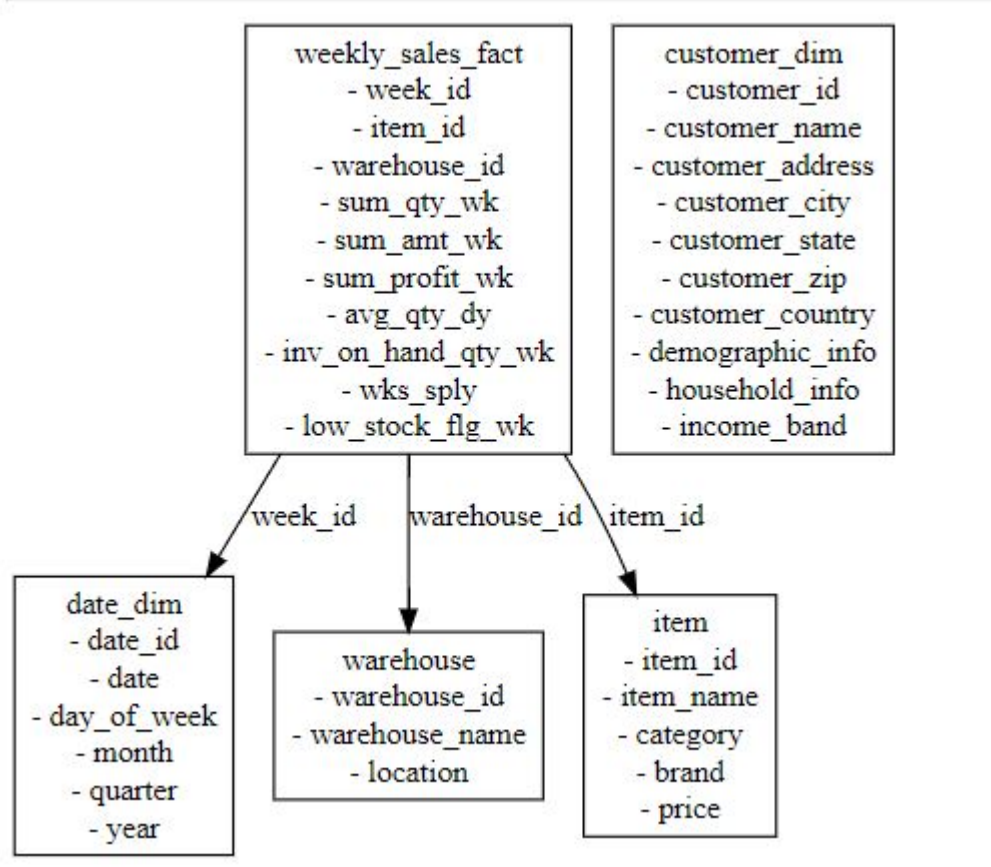
# ETL and Data Loading

1. Implement merge scripts for integrating customer dimension and daily sales records.
2. Create SQL queries for joining daily sales and updated inventory tables to generate the weekly sales and inventory fact table.
3. in order to realize the requirements from the Metabase BI purpose, we also need the Date_dim, Warehouse, and Item tables in the Prod schema.




WeCloudData

# ETL and Data Loading




WeCloudData

# Testing

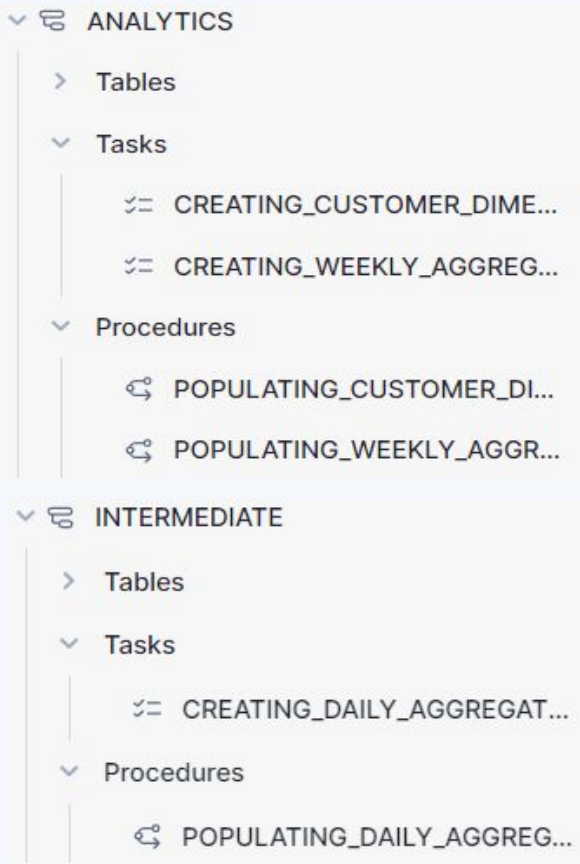
1. Not_null, Unique, Relationship, Accepted Value, Value Range, Length Tests.
2. Foreign Key Integrity Test to ensuring foreign key relationships are valid.

| | TEST_NAME | RESULT |
|---|---|---|
| 1 | Not Null Test | Pass |
| 2 | Unique Test | Pass |
| 3 | Relationship Test | Pass |
| 4 | Accepted Value Test | Pass |
| 5 | Foreign Key Integrity Test | Pass |
| 6 | Value Range Test | Pass |
| 7 | Length Test | Pass |

WeCloudData

# Automating

1. Set up automated processes for daily and weekly data aggregations.
2. Create SQL queries for joining daily sales and updated inventory tables to generate the weekly sales and inventory fact table.
3. Configure cron jobs using Snowflake Tasks to execute daily and weekly aggregation scripts at specified times.




WeCloudData

# Metabase (Business Requirements)

1. Install and configure Metabase for data visualization and report generation.
2. Connect Metabase to the Snowflake database.
3. Fix field names, and assign types to the columns to ensure proper visualization.
4. Generate reports based on business requirements using dashboards.

Weekly Inventory Shortages

| Warehouse Sk | Sold Wk Sk → Cal Dt: Week | Low Stock Flg Wk | Item Sk |
|---|---|---|---|
| 1 | December 27, 2020 – January 2, 2021 | true | 56 |
| 1 | December 27, 2020 – January 2, 2021 | true | 262 |
| 1 | December 27, 2020 – January 2, 2021 | true | 272 |
| 1 | December 27, 2020 – January 2, 2021 | true | 442 |
| 1 | December 27, 2020 – January 2, 2021 | true | 478 |
| 1 | December 27, 2020 – January 2, 2021 | true | 512 |

Rows 1-6 of first 2000

Total Weekly Shortages


Top Selling Items

| CAL_WK | ITEM_SK | SUM_AMT_WK | SUM_QTY_WK | RANKING |
|---|---|---|---|---|
| June 12, 2022 | 12,748 | 39,543.32 | 288 | 1 |
| June 12, 2022 | 15,976 | 27,140.85 | 244 | 2 |
| June 12, 2022 | 14,959 | 26,416.16 | 172 | 3 |
| June 12, 2022 | 13,754 | 26,333.56 | 97 | 4 |
| June 12, 2022 | 9,700 | 25,858.86 | 173 | 5 |
| June 12, 2022 | 2,257 | 25,597.49 | 221 | 6 |

Rows 1-6 of first 2000

Bottom Selling Items

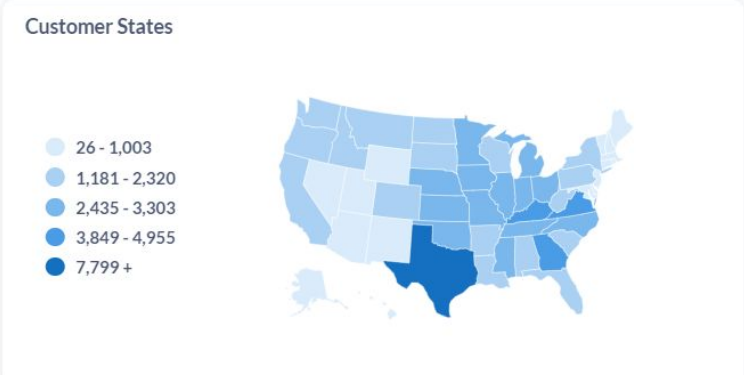
| CAL_WK | ITEM_SK | SUM_AMT_WK | SUM_QTY_WK | RANKING |
|---|---|---|---|---|
| September 11, 2022 | 14,467 | 0 | 1 | 1 |
| September 11, 2022 | 625 | 0 | 3 | 2 |
| September 11, 2022 | 9,038 | 0 | 5 | 3 |
| September 11, 2022 | 17,617 | 0 | 5 | 3 |
| September 11, 2022 | 4,477 | 0 | 6 | 5 |
| September 11, 2022 | 9,430 | 0 | 9 | 6 |
| September 11, 2022 | 7,231 | 0 | 13 | 7 |
| September 11, 2022 | 9,266 | 0 | 13 | 8 |

Rows 1-8 of first 2000

WeCloudData

# Metabase (Additional Requirements)






WeCloudData

# Challenges

1. Data Quality and Consistency:
   Implement data validation and cleansing routines to ensure data quality before aggregation.

2. Complex Transformations:
   Ensuring that the logic for these calculations is robust and handles edge cases properly.

3. Deployment and Migration:
   Migrating existing data to the new schema and ensuring that historical data is accurately transformed to include the new metrics.

# Project Achievement

1. Exploring the data and understanding business requirements.
2. Merging data in temporary tables to avoid affecting tables in use.
3. Testing tables to validate the results before automating the pipelines.

# Future

1. Enhanced Data Quality Measures:
   Implementing robust data cleansing and validation techniques to ensure high data quality and accuracy.

2. Advanced Analytics Capabilities:
   Incorporating advanced analytics techniques such as predictive modeling or machine learning to derive deeper insights from the data.

3. Security Enhancements:
   Implementing additional security measures to protect sensitive data throughout the entire data lifecycle.

# Thank you!

WeCloudData