Data Engineer Bootcamp

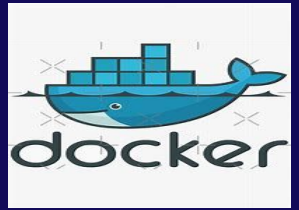
Kholoud Ali Ahmad Bin Shwayah,
Shatha albader,
Sultan Mohammed Almalki

Capstone Project

WeCloudData

Project Objectives

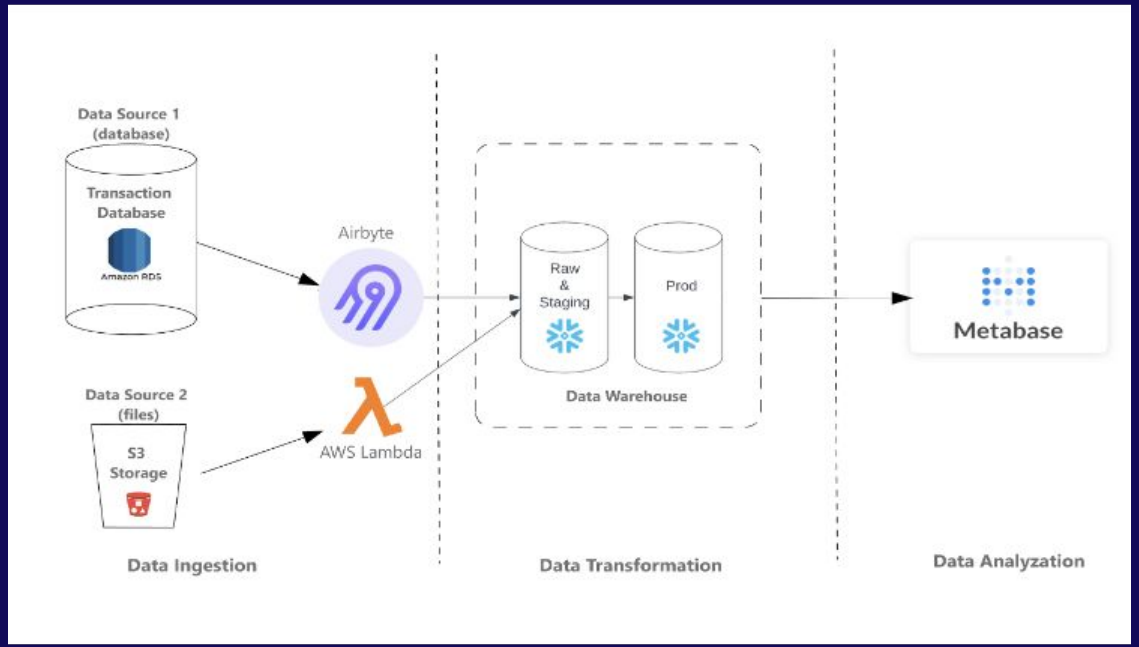
- The project requirements involve building a data warehouse to support business intelligence (BI) reporting
- The main tools we used in this project:



WeCloudData

Project Objectives

- The project process



WeCloudData

Data Overview

- The weekly sales fact table will be linked to the customer dimension table through a foreign key referencing customer ID.
- This allows us to analyze sales performance by customer segments.

Fact tables	Dimention tables
Catalog_Sales	Date_Dim
Web_Sales	Customer
Inventory	Item
FROM S3	Promotion
	Customer_Gemographics
	Call_Center
	Customer_Address
	Catalog_Page
	Warehouse
	Time_Dim
	Ship_Mode
	Household_Demographics
	Icome_Band
	Web_page
	Web_Site

WeCloudData

Data Overview

- The weekly sales fact table will be continuously loaded with new sales data each day to keep the weekly sales figures up-to-date.

- dimension tables as containing relatively static data about entities like customers or products.

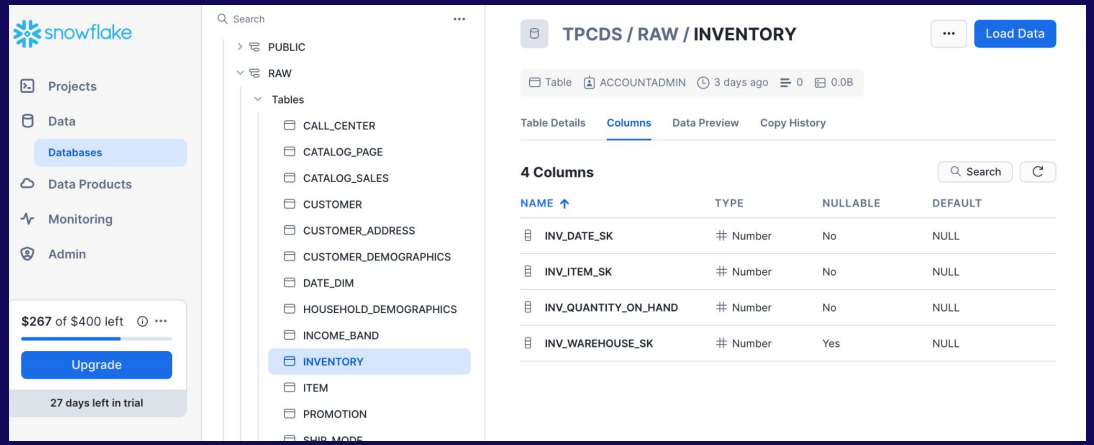
WeCloudData

Project Steps (cont)..

We will walk through the process of building:

Part 1: Data Ingestion

-Create a Database on Snowflake to store the data.



WeCloudData

Project Steps (cont)..

-Launching 2 Ubuntu EC2 Instances in AWS Console,

one for **Airbyte** and one for **Metabase**.

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone
Airbyte	i-06d391f927c505642	Stopped	t2.large	–	View alarms +	us-east-1e
Metabase	i-0f0a7301cb2879878	Stopped	t2.small	–	View alarms +	us-east-1c

-Install Docker on both EC2 instances.

WeCloudData

Project Steps, (cont)..

-Create an AWS Lambda Function

File Edit Find View Go Tools Window Test Deploy

lambda_functior × Environment Var ×

Environment

```python
import os
import boto3
import requests
import snowflake.connector as sf


def lambda_handler(event, context):

    url = 'https://de-materials-tpcds.s3.ca-central-1.amazonaws.com/inventory.csv'
    destination_folder = '/tmp'
    file_name = 'inventory.csv'
    local_file_path = '/tmp/inventory.csv'

    # Snowflake connection parameters
    account ='OPSJOQP-CI49455'
    warehouse ='COMPUTE_WH'
    database = 'TPCDS'
    schema = 'RAW'
    table = 'inventory'
    user = 'shatha1'
    password ='AbS@123456789'
    role='accountadmin'
    stage_name ='inv_Stage'

    # Download the data from the API endpoint
    response = requests.get(url)
    response.raise_for_status()



    # Save the data to the destination file in /tmp directory
    file_path = os.path.join(destination_folder, file_name)
```

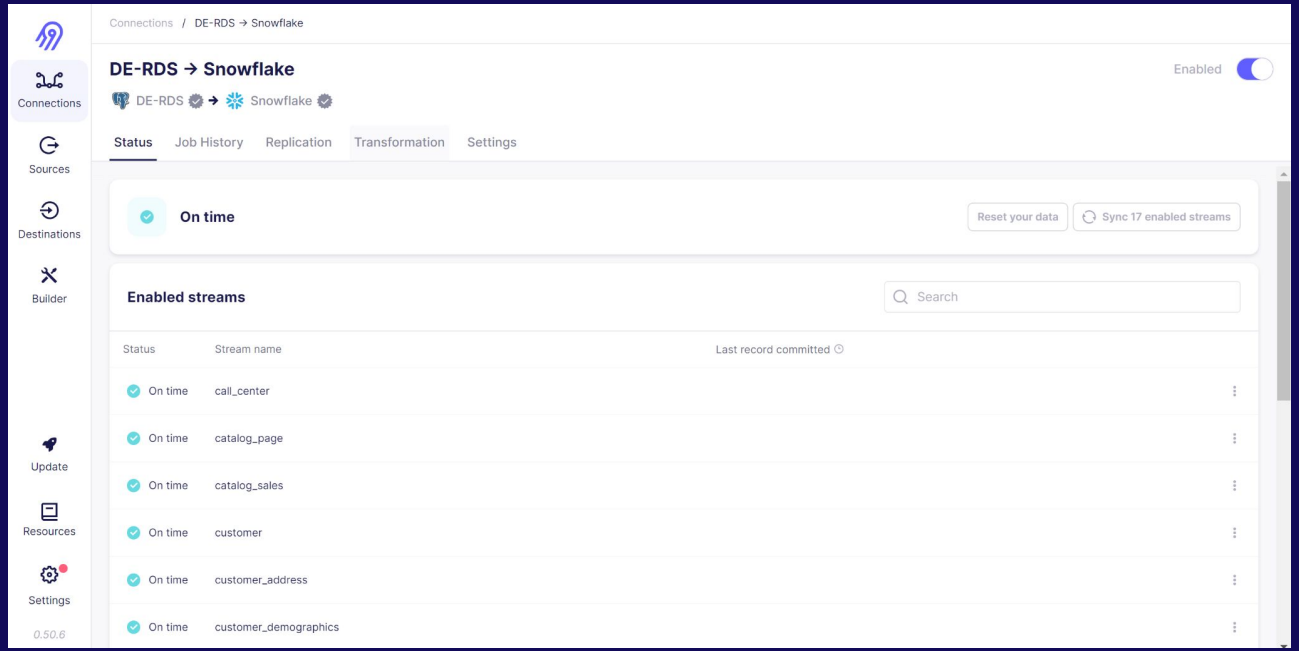
lambda_function × Environment Var ×

```python

    # Establish Snowflake connection
    conn = sf.connect(user = user, password = password, \
                      account = account, warehouse=warehouse, \
                       database=database,  schema=schema,  role=role)


    cursor = conn.cursor()

    # use schema
    use_schema = f"use schema {schema};"
    cursor.execute(use_schema)

    # create CSV format
    create_csv_format = f"CREATE or REPLACE FILE FORMAT COMMA_CSV TYPE ='CSV' FIELD_DELIMITER = ',';"
    cursor.execute(create_csv_format)


    create_stage_query = f"CREATE OR REPLACE STAGE {stage_name} FILE_FORMAT =COMMA_CSV "
    cursor.execute(create_stage_query)

    # Copy the file from local to the stage
    copy_into_stage_query = f"PUT 'file://{local_file_path}' @{stage_name}"
    cursor.execute(copy_into_stage_query)

    # List the stage
    list_stage_query = f"LIST @{stage_name}"
    cursor.execute(list_stage_query)

    # truncate table
    truncate_table = f"truncate table {schema}.{table};"
    cursor.execute(truncate_table)


    # Load the data from the stage into a table (example)
    copy_into_query = f"COPY INTO {schema}.{table} FROM @{stage_name}/{file_name} FILE_FORMAT =COMMA_CSV
    cursor.execute(copy_into_query)
```

WeCloudData

Project Steps, (cont)..

-Install and configure Airbyte on one of the EC2 instances. (port: 8000)



WeCloudData

Project Steps, (cont)..

Part 2: Data Modeling

1. Business Requirements and Data Modeling

   *Gather and understand the business requirements.

   *Design the data model to meet the business requirements.

2. ETL and Data Loading

   *Extract, transform, and load the data into the Snowflake database.

3. Scheduling with Snowflake

   *to automate the data loading process.

4. Creating Snowflake Stored Procedure
   - Create a Snowflake stored procedure to handle data processing or transformation.

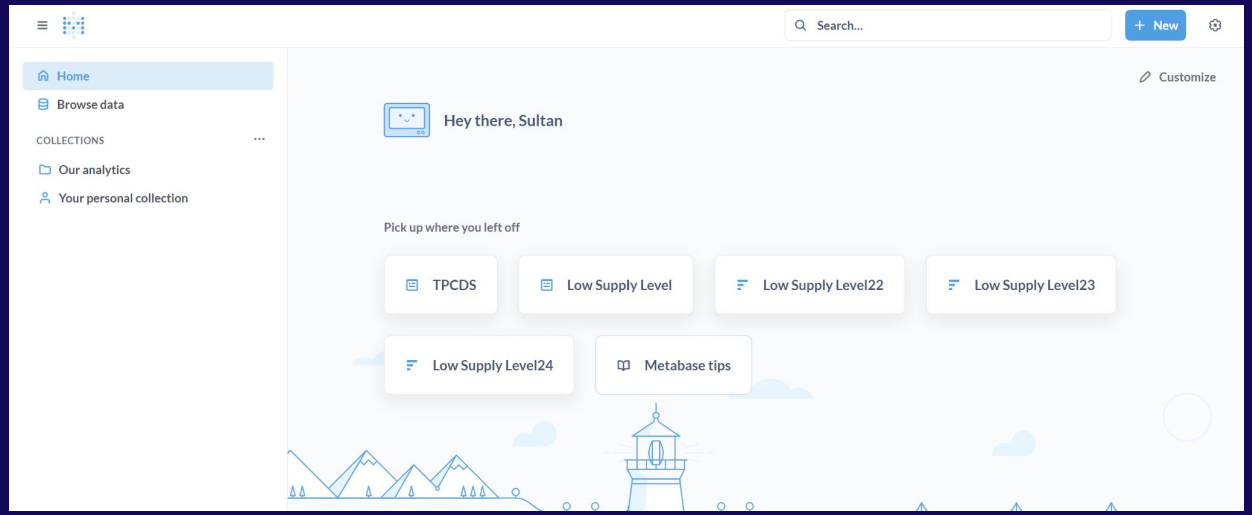
5. Creating Snowflake Tasks
   - Create Snowflake tasks to automate the execution of the stored procedure.

WeCloudData

Project Achievement:

Final Part: Data Visualization:

-Install and configure Metabase on the second EC2 instance. (port: 3000)



WeCloudData

Project Achievement

1. Analyze sales amounts to identify the highest performing items of the week.(2022,2023)

WeCloudData

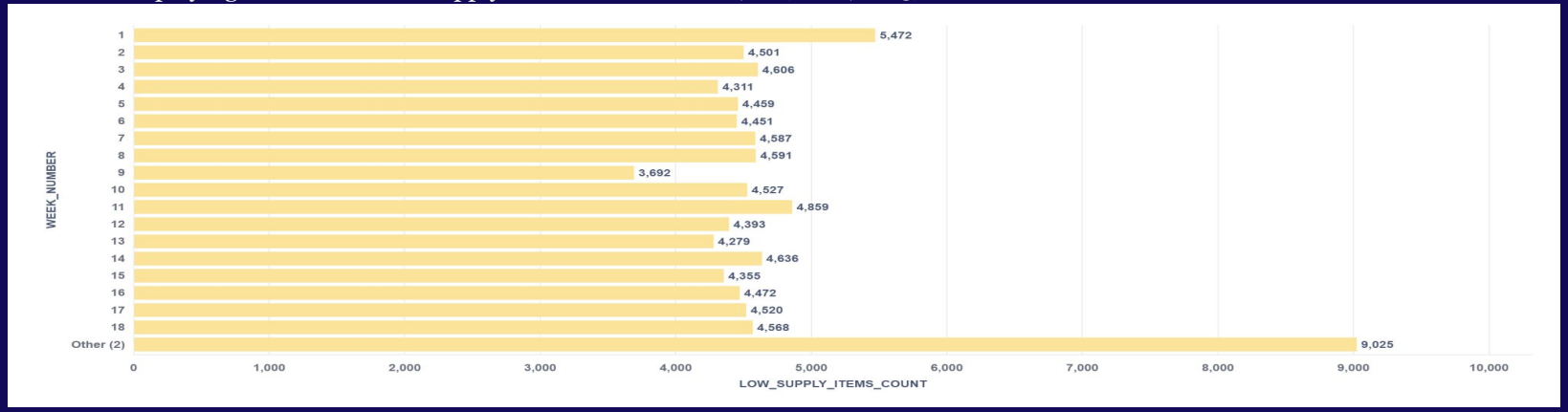
Project Achievement

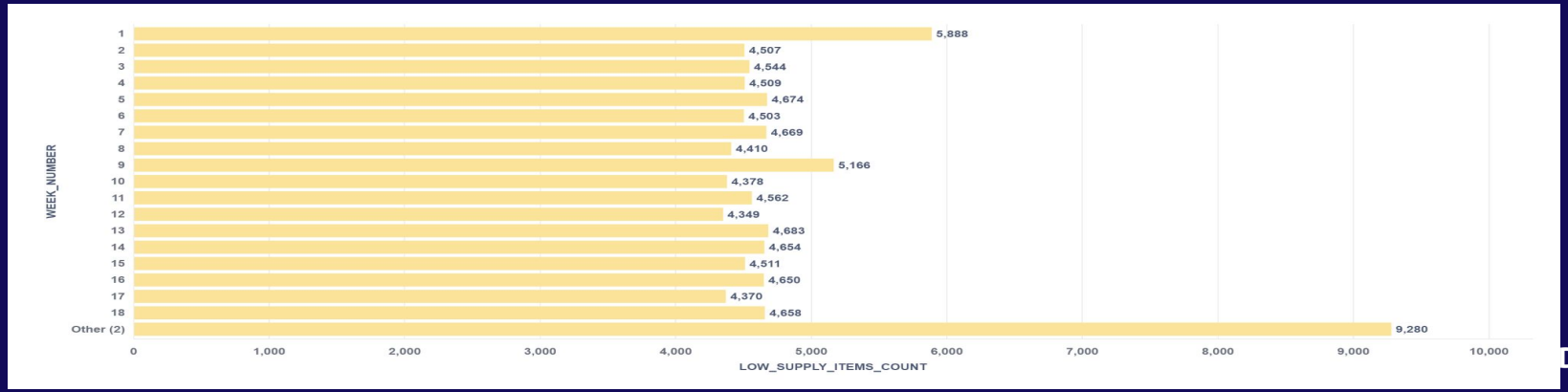
1. Analyze sales quantities to identify the highest performing items of the week.(2022,2023)

WeCloudData

Project Achievement

1. Displaying Items with Low Supply Levels for each week, in (2022,2023)





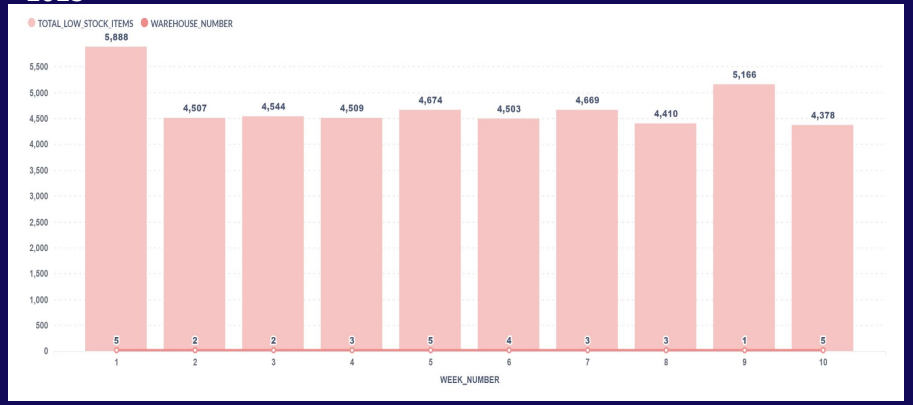
Data

Project Achievement

1. Detecting Items with Low Stock Levels, along with their corresponding week and warehouse numbers, marked as True.

2022

2023



WeCloudData

Future

- Exploring advanced techniques for handling large data volumes or complex transformations.
- Change the date data type from number to date.
- Data cleaning.

Thank you

WeCloudData