

Data Engineer Bootcamp



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



Capstone Project



01 Project Objects

02 Data Overview

03 Project Steps

04 Project Achievement

05 Future



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:



Amazon
EC2



Airbyte



RDS



PostgreSQL



Metabase



Amazon S3



AWS Lambda



snowflake



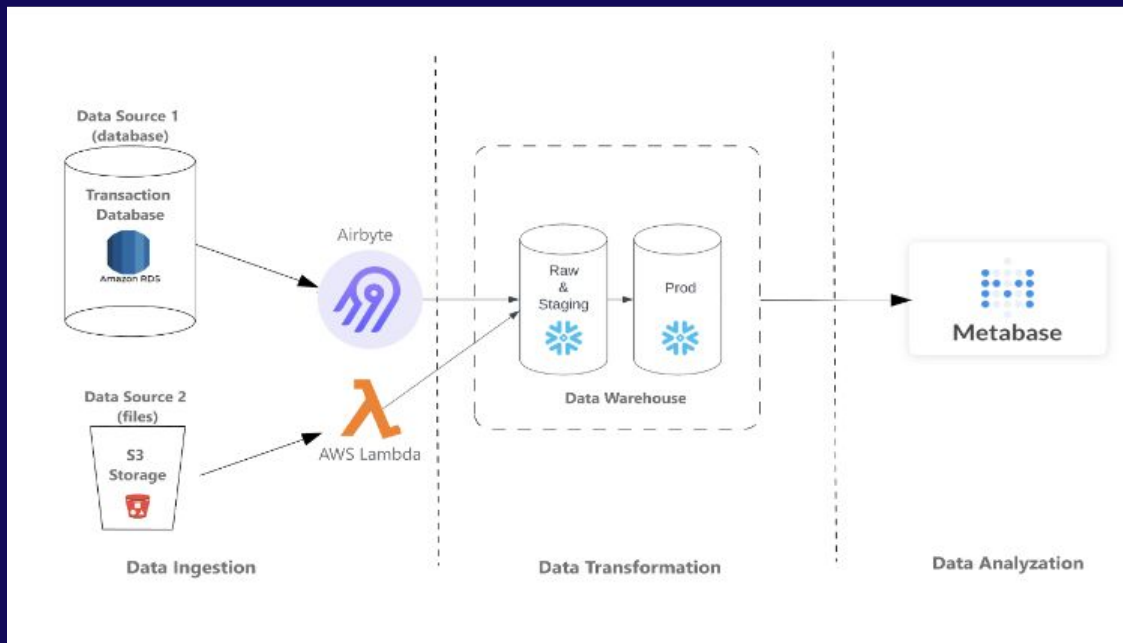
docker



WeCloudData

Project Objectives

- The project process



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



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.

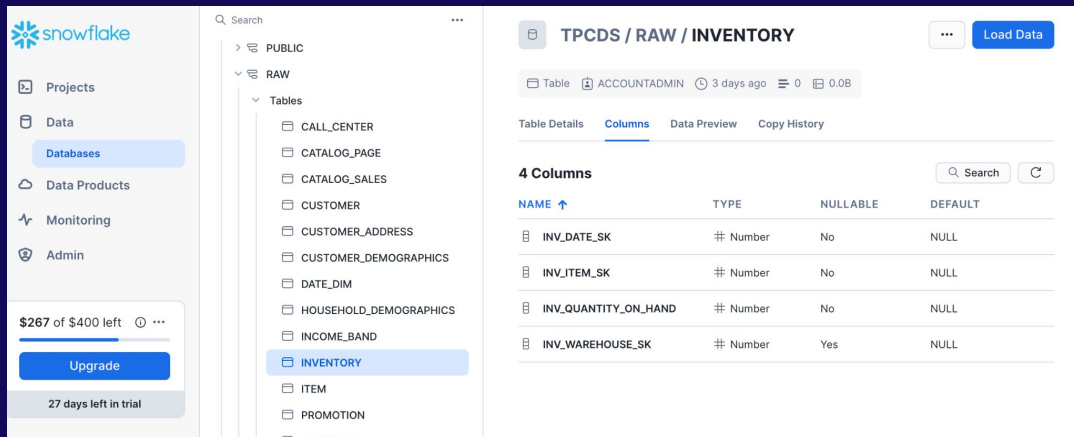


Project Steps (cont)..

We will walk through the process of building:

Part 1: Data Ingestion

-Create a Database on Snowflake to store the data.








The screenshot displays the Snowflake web interface. On the left, the navigation sidebar includes 'Projects', 'Data', 'Databases', 'Data Products', 'Monitoring', and 'Admin'. The 'Databases' section is active, showing a list of databases including 'INVENTORY'. The main panel shows the 'TPCDS / RAW / INVENTORY' table details. The 'Columns' tab is selected, displaying a table with 4 columns: 'NAME', 'TYPE', 'NULLABLE', and 'DEFAULT'. The table lists four columns: 'INV_DATE_SK' (Number, No, NULL), 'INV_ITEM_SK' (Number, No, NULL), 'INV_QUANTITY_ON_HAND' (Number, No, NULL), and 'INV_WAREHOUSE_SK' (Number, Yes, NULL). A 'Load Data' button is visible in the top right corner of the table details panel.

NAME	TYPE	NULLABLE	DEFAULT
INV_DATE_SK	# Number	No	NULL
INV_ITEM_SK	# Number	No	NULL
INV_QUANTITY_ON_HAND	# Number	No	NULL
INV_WAREHOUSE_SK	# Number	Yes	NULL

Project Steps (cont)..

-Launching 2 Ubuntu EC2 Instances in AWS Console,
one for **Airbyte** and one for **Metabase**.

<input type="checkbox"/>	Name  ▲	Instance ID	Instance state ▼	Instance type ▼	Status check	Alarm status	Availability Zone ▼
<input type="checkbox"/>	Airbyte	i-06d391f927c505642	⊖ Stopped  	t2.large	–	View alarms +	us-east-1e
<input type="checkbox"/>	Metabase	i-0f0a7301cb2879878	⊖ Stopped  	t2.small	–	View alarms +	us-east-1c

-Install Docker on both EC2 instances.

Project Steps, (cont)..

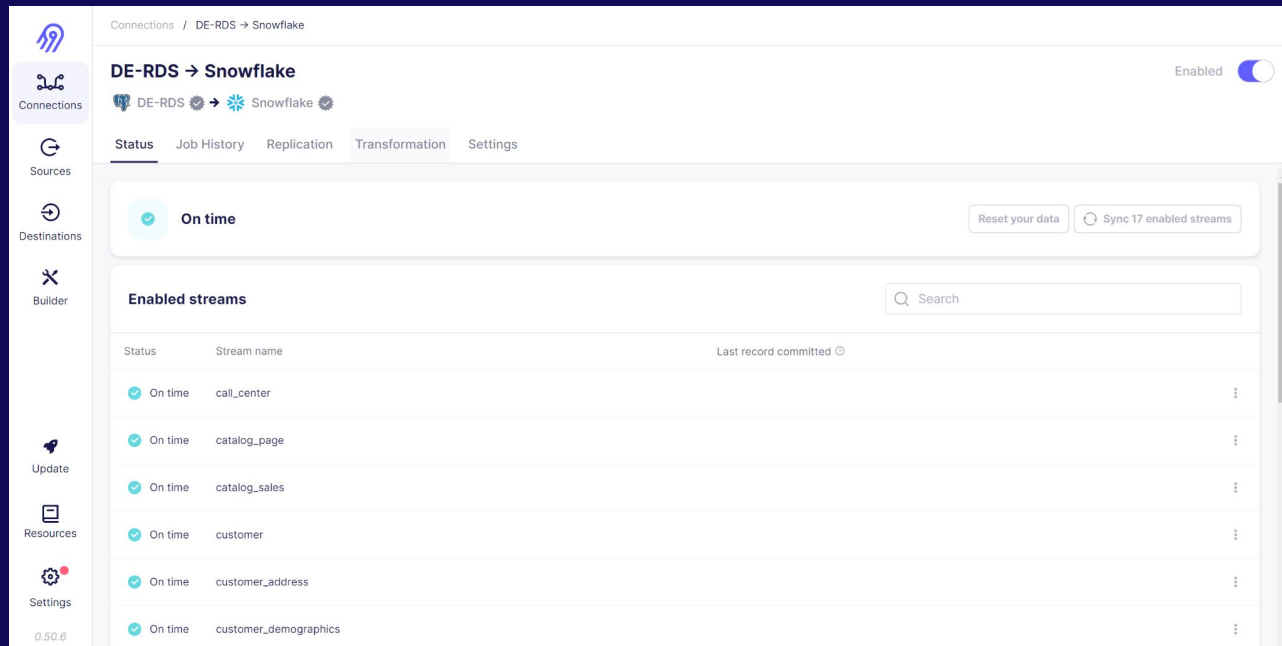
-Create an AWS Lambda Function

```
File Edit Find View Go Tools Window Test Deploy
lambda_function x Environment Var x +
Environment
1
2 import os
3 import boto3
4 import requests
5 import snowflake.connector as sf
6
7
8 def lambda_handler(event, context):
9
10     url = 'https://de-materials-tpcds.s3.ca-central-1.amazonaws.com/inventory.csv'
11     destination_folder = '/tmp'
12     file_name = 'inventory.csv'
13     local_file_path = '/tmp/inventory.csv'
14
15     # Snowflake connection parameters
16     account = 'OPSJOQP-CI49455'
17     warehouse = 'COMPUTE_WH'
18     database = 'TPCDS'
19     schema = 'RAW'
20     table = 'inventory'
21     user = 'shathal'
22     password = 'AbS@123456789'
23     role = 'accountadmin'
24     stage_name = 'inv_Stage'
25
26     # Download the data from the API endpoint
27     response = requests.get(url)
28     response.raise_for_status()
29
30
31
32     # Save the data to the destination file in /tmp directory
33     file_path = os.path.join(destination_folder, file_name)
```

```
lambda_function x Environment Var x +
35
36 # Establish Snowflake connection
37 conn = sf.connect(user = user, password = password, \
38                 account = account, warehouse=warehouse, \
39                 database=database, schema=schema, role=role)
40
41
42 cursor = conn.cursor()
43
44 # use schema
45 use_schema = f"use schema {schema};"
46 cursor.execute(use_schema)
47
48 # create CSV format
49 create_csv_format = f"CREATE OR REPLACE FILE FORMAT COMMA_CSV TYPE = 'CSV' FIELD_DELIMITER = ',';"
50 cursor.execute(create_csv_format)
51
52
53 create_stage_query = f"CREATE OR REPLACE STAGE {stage_name} FILE_FORMAT =COMMA_CSV "
54 cursor.execute(create_stage_query)
55
56 # Copy the file from local to the stage
57 copy_into_stage_query = f"PUT 'file://{local_file_path}' @{stage_name}"
58 cursor.execute(copy_into_stage_query)
59
60 # List the stage
61 list_stage_query = f"LIST @{stage_name}"
62 cursor.execute(list_stage_query)
63
64 # truncate table
65 truncate_table = f"truncate table {schema}.{table};"
66 cursor.execute(truncate_table)
67
68
69 # Load the data from the stage into a table (example)
70 copy_into_query = f"COPY INTO {schema}.{table} FROM @{stage_name}/{file_name} FILE_FORMAT =COMMA_CSV"
71 cursor.execute(copy_into_query)
```

Project Steps, (cont)..

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



The screenshot displays the Airbyte web interface for a connection named 'DE-RDS to Snowflake'. The interface is divided into a left sidebar and a main content area. The sidebar contains navigation links: Connections, Sources, Destinations, Builder, Update, Resources, and Settings. The main content area shows the connection status as 'Enabled' with a toggle switch. Below this, there are tabs for Status, Job History, Replication, Transformation, and Settings. The 'Status' tab is active, showing a green checkmark and the text 'On time'. To the right of this status are buttons for 'Reset your data' and 'Sync 17 enabled streams'. Below the status bar is a section titled 'Enabled streams' with a search bar. A table lists the enabled streams with columns for Status, Stream name, and Last record committed. The table contains six rows of data.

Status	Stream name	Last record committed
On time	call_center	
On time	catalog_page	
On time	catalog_sales	
On time	customer	
On time	customer_address	
On time	customer_demographics	

Project Steps, (cont)..

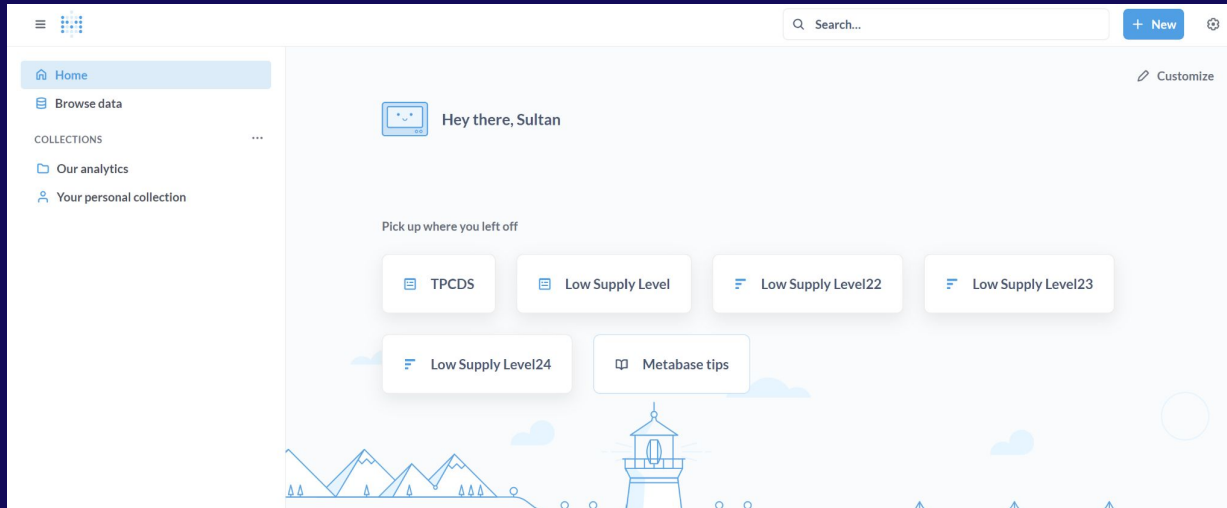
Part 2: Data Modeling

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.

Project Achievement:

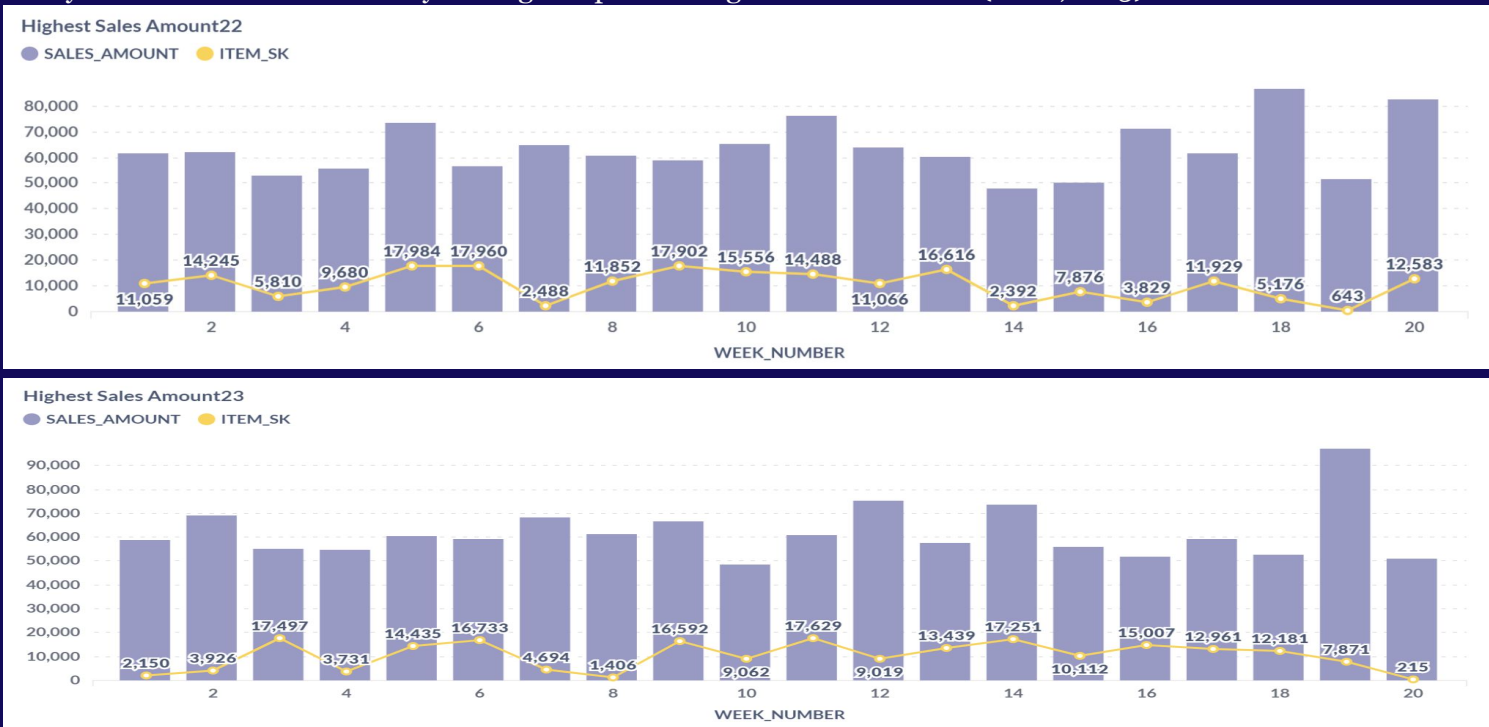
Final Part: Data Visualization:

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



Project Achievement

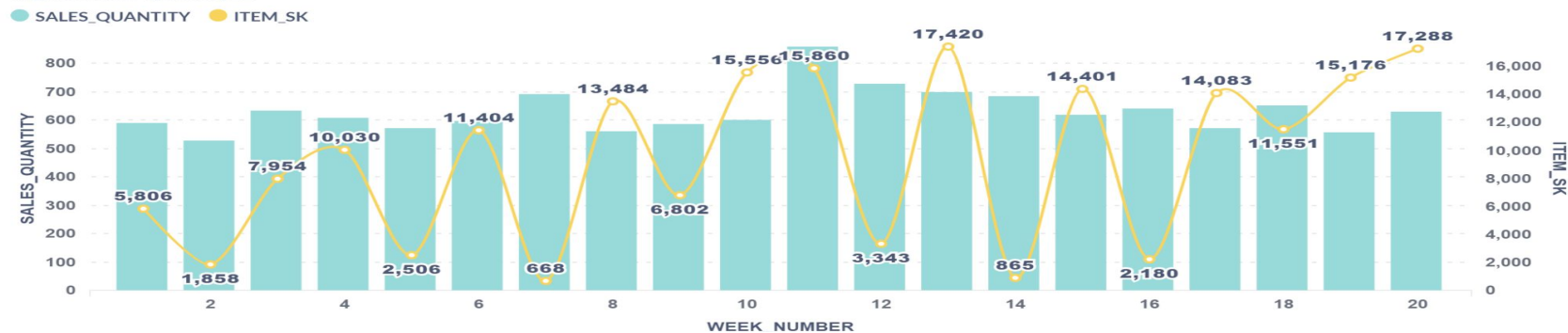
1. Analyze sales amounts to identify the highest performing items of the week.(2022,2023)



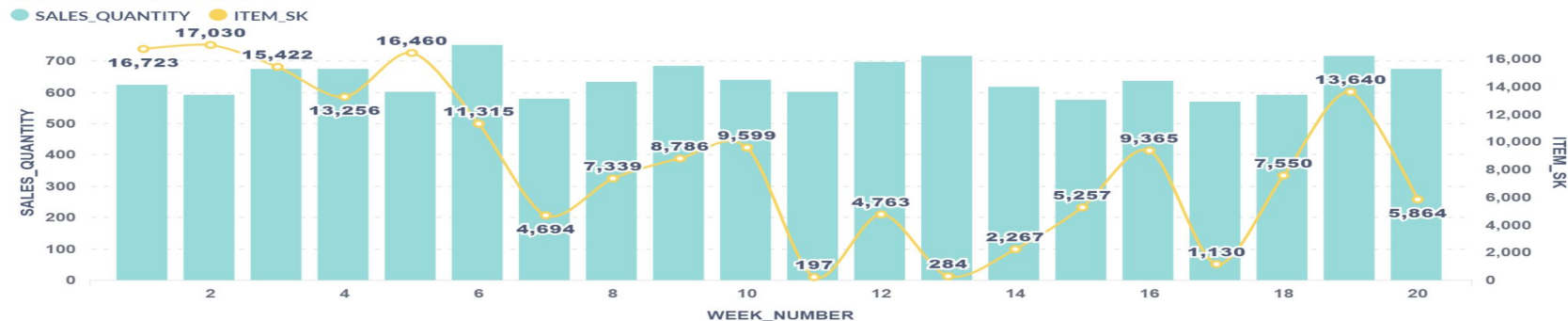
Project Achievement

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

Highest Sales Quantity22

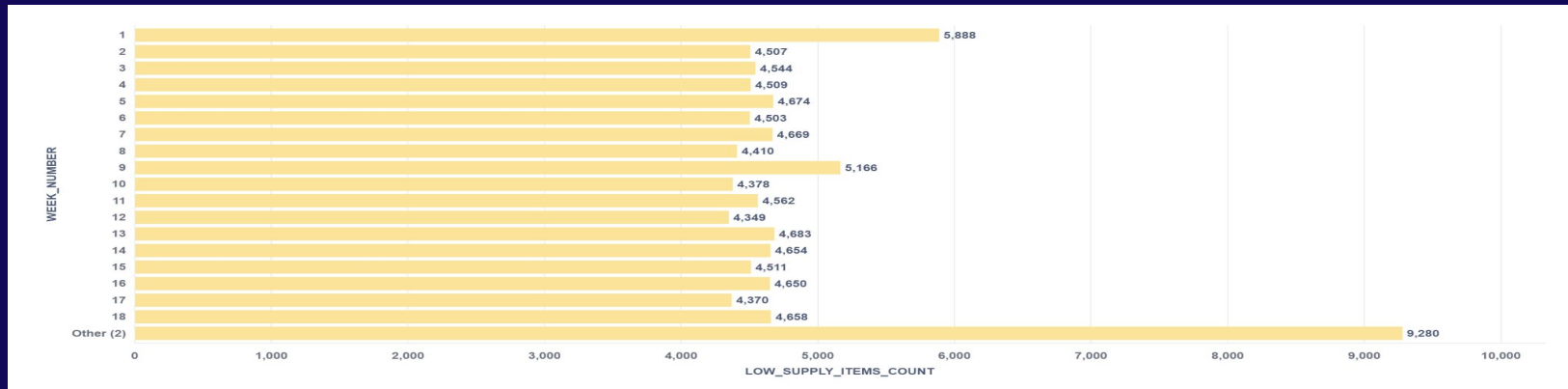
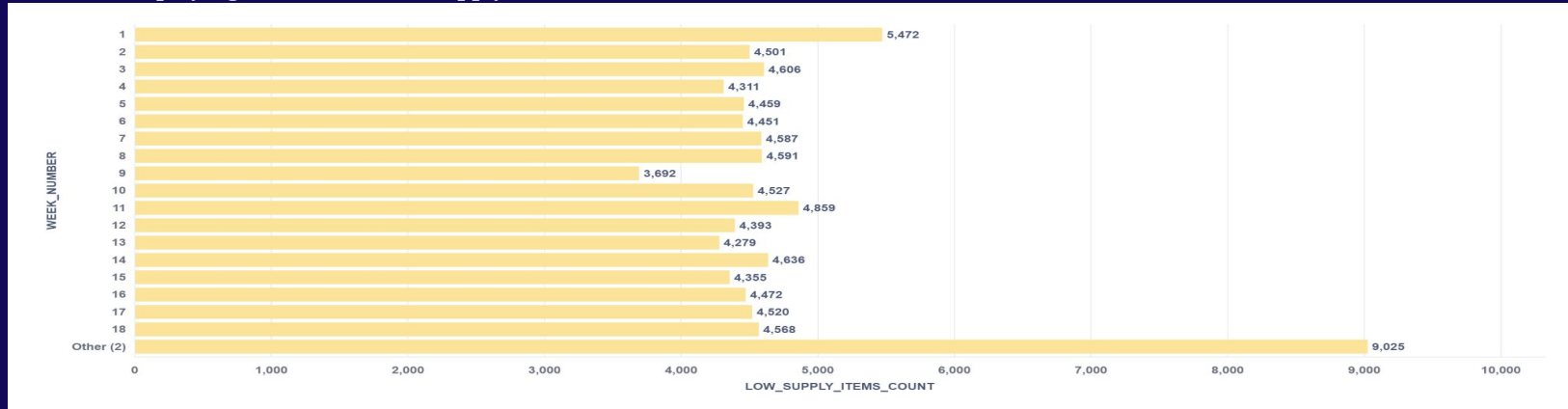


Highest Sales Quantity23



Project Achievement

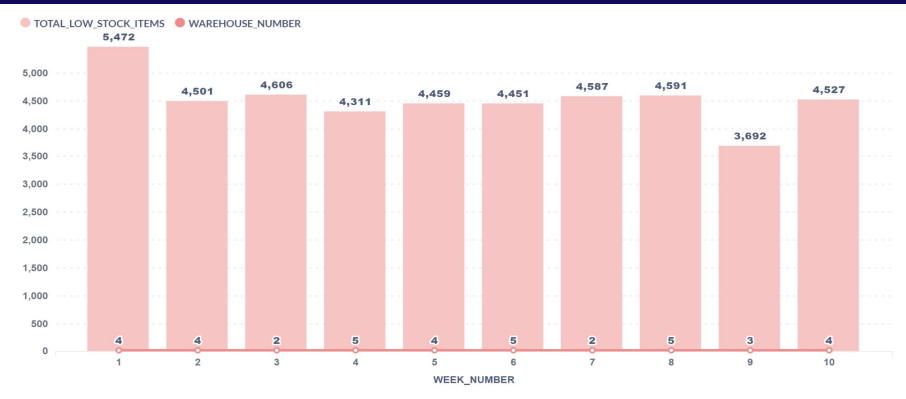
1. Displaying Items with Low Supply Levels for each week, in (2022,2023)



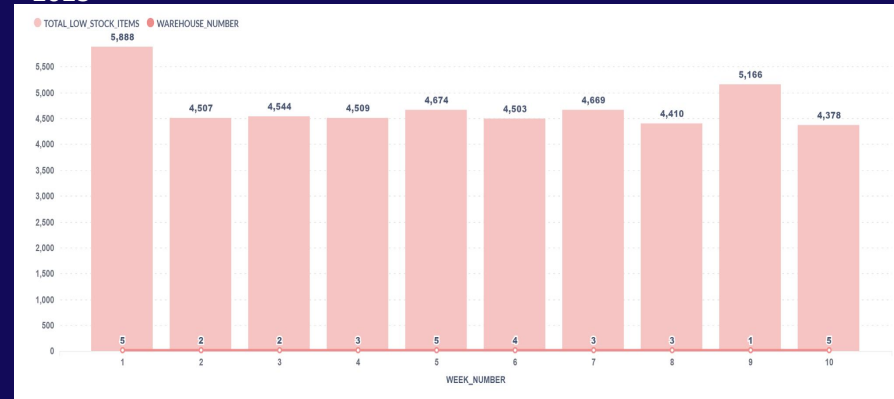
Project Achievement

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

2022



2023



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