Vishwakarma Institute Of Information Technology, Pune.

Department Computer Engineering (Software Engineering)

(Data Warehouse and Data Analytics Lab)

EXPERIMENT NO – 01

Aim : Choose a set of business processes like Sales, Customer Services, Accounting, Production, Marketing processes etc. for any organization and design star, snowflake and fact constellation schema. Also using ETL tools ,extract data from various sources and perform transform and load operations on data. (Use Power BI/RapidMiner).

Objective: To design dimensional models (star, snowflake, and fact constellation schemas) for a set of business processes and implement an ETL process to populate the data warehouse using a chosen ETL tool (Power BI or RapidMiner).

Theory :

Scenario: Choose an organization (real or hypothetical) and select at least three business processes within that organization. Examples:

- Retail: Sales, Customer Service, Marketing
- Manufacturing: Production, Inventory, Sales
- Healthcare: Patient Care, Billing, Admissions
- Education: Student Enrollment, Course Management, Alumni Relations

Tasks:

1. **Business Process Definition:**
   - Clearly define the chosen business processes.
   - Describe the key metrics and dimensions associated with each process.
   - Identify the data sources for each process (e.g., CRM, ERP, databases, files).
2. **Dimensional Modeling:**
   - **Star Schema:** Design a star schema for *one* of the chosen business processes. Include a fact table and dimension tables. Clearly define the grain of the fact table.
   - **Snowflake Schema:** Design a snowflake schema for the *same* business process as the star schema. Normalize at least one dimension table. Explain the benefits of the snowflake schema in this context.
   - **Fact Constellation Schema:** Design a fact constellation schema that integrates *all* chosen business processes. Identify shared dimension tables. Explain how this schema supports cross-process analysis.
3. **ETL Process Design and Implementation:**
   - **Tool Selection:** Choose either Power BI or RapidMiner for the ETL implementation. Justify your choice.
   - **Data Sources:** Describe the simulated or actual data sources you will use for your implementation. If using simulated data, explain how you generated it. If using real data, ensure you have the necessary permissions and anonymize any sensitive information.
   - **ETL Steps:** Detail the steps involved in your ETL process:
     - **Extraction:** Describe how you will extract data from the various sources.
     - **Transformation:** Explain and implement the transformations you will perform:

        - Data Cleaning: Handle missing values (using appropriate techniques like mean/median imputation, mode imputation for categorical data), remove duplicates, correct inconsistencies. Justify your chosen methods.
        - Data Type Conversion: Ensure proper data types.
        - Calculations: Derive new measures (e.g., total sales, average interaction time).
        - Data Integration: Join data from different sources.
        - Surrogate Key Generation: Create surrogate keys for dimension tables.
        - Date Dimension Creation: Create a date dimension table.
        - Data Transformation: Include log transformation for skewed numerical data (if applicable) and one-hot encoding for categorical variables.
    - **Loading:** Describe how you will load the transformed data into the data warehouse (Power BI data model or RapidMiner repository).

4. **Data Warehouse Implementation (Power BI or RapidMiner):**
    - Implement the designed schemas in your chosen tool.
    - Establish relationships between fact tables and dimension tables.
5. **Analysis and Reporting (Power BI or RapidMiner):**
    - Develop at least three reports or visualizations that demonstrate the analytical capabilities of your data warehouse. These should showcase insights related to the chosen business processes and, ideally, cross-process analysis enabled by the fact constellation schema.

Power BI Tool:

Power BI is a Data Visualization and Business Intelligence tool that converts data from different data sources to interactive dashboards and BI reports. In Power BI we can load the data from sources like csv, excel files and also can transform that data. In here we can easily design different styles of data mart schemas which are most widely used to develop data warehouses.

Dataset Used:

For this assignment we have used one dataset. It is the **Supermarket sales dataset**. We have downloaded these datasets from the Kaggle.

Loading Supermarket Sales Dataset In Power

After performing preprocessing on the dataset we moved towards the next step which is the “Loading” step. We have opened the Power BI tool and there is the Get Data option present in the Home tab. Here we have given the Supermarket sales dataset file that we have to load and after giving the file path it shows a preview of the dataset in the window and then click on the Load button to load the dataset into Power BI. We successfully loaded the Dataset in Power BI Tool.

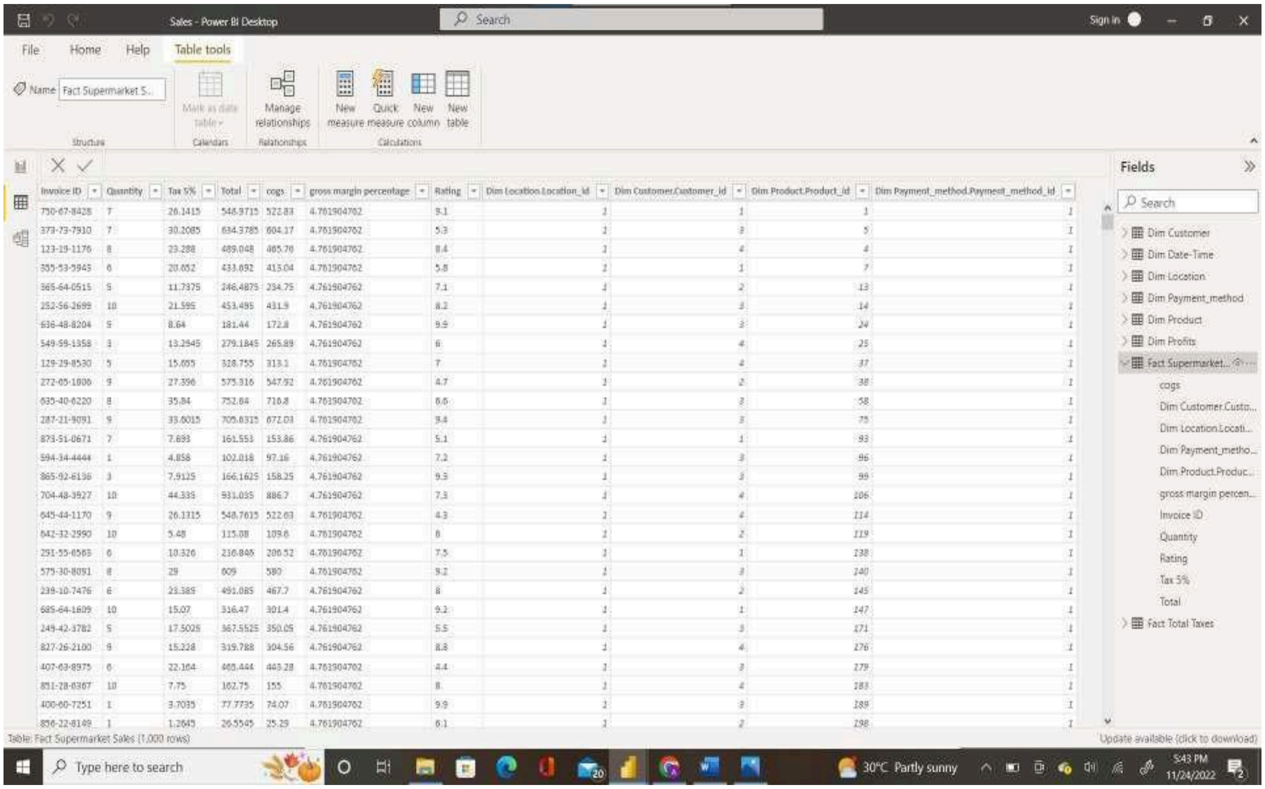
Star Schema:

The Star schema is the simplest style of data fomat schema which is most widely used to develop data warehouses. It consists of a fact table in the middle of the schema connecting to a set of dimension tables. The fact table and dimension table hold the columns that store the data for the model.

Fact table mainly consists of the business facts and foreign keys that refer to the primary keys in the dimension table. The Dimension table consists mainly of descriptive attributes.

Creating Star Schema:

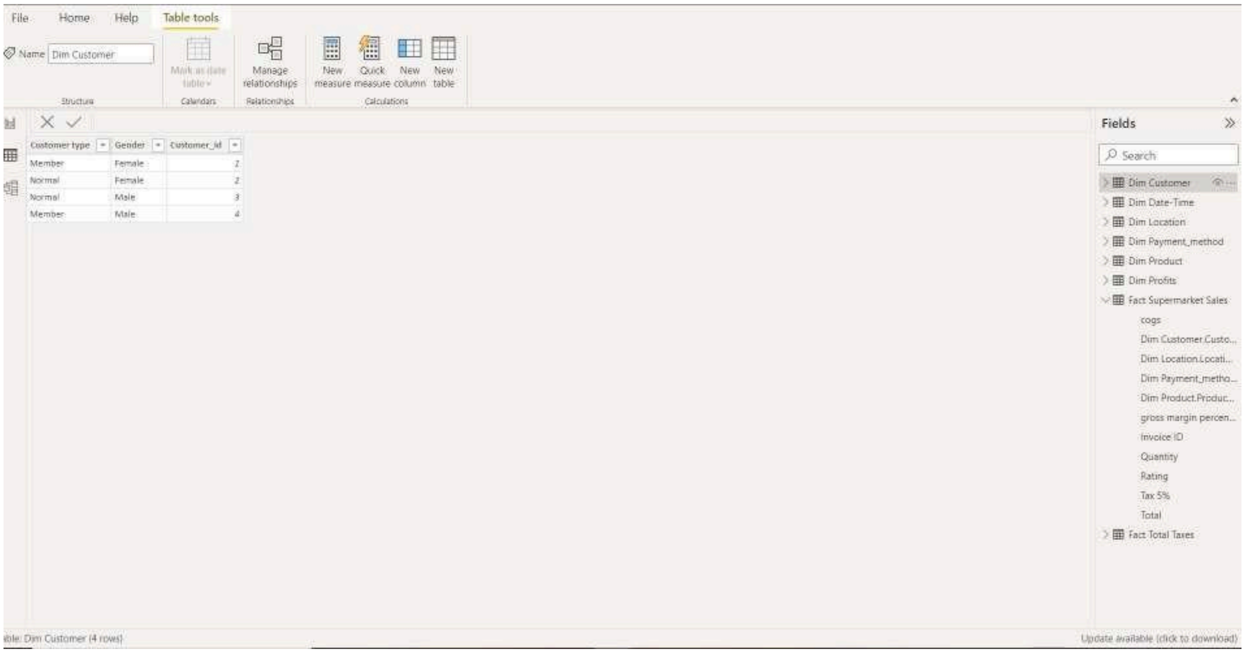
When we have loaded the data then there exists only single table names as Supermarket sales tables which act as our FactTable. We have renamed the table as Fact Supermarket Sales.



Invoice ID	Quantity	Tax 5%	Total	cogs	gross margin percentage	Rating	Dim Location.Location_id	Dim Customer.Customer_id	Dim Product.Product_id	Dim Payment_method.Payment_method_id
750-67-8428	7	26.1415	548.9715	522.83	4.761904762	9.1	1	1	1	1
373-73-7910	7	30.2085	634.3785	604.17	4.761904762	5.3	1	3	5	1
123-19-1176	8	23.288	489.048	465.76	4.761904762	8.4	1	4	4	1
355-53-5943	6	20.652	433.692	413.04	4.761904762	5.8	1	1	7	1
365-64-0515	5	11.7375	246.4875	234.75	4.761904762	7.1	1	2	13	1
252-56-2699	10	21.595	453.495	431.9	4.761904762	8.2	1	3	14	1
636-48-8204	5	8.64	181.44	172.8	4.761904762	9.9	1	3	24	1
549-59-1358	3	13.2945	279.1845	265.89	4.761904762	6	1	4	25	1
129-29-8530	5	15.655	328.755	313.1	4.761904762	7	1	4	37	1
272-65-1806	9	27.396	575.316	547.92	4.761904762	4.7	1	2	38	1
635-40-6220	8	35.84	752.64	716.8	4.761904762	6.6	1	3	58	1
287-21-9091	9	33.6015	705.6315	672.03	4.761904762	9.4	1	3	75	1
873-51-0671	7	7.693	161.553	153.86	4.761904762	5.1	1	1	93	1
594-34-4444	1	4.858	102.018	97.16	4.761904762	7.2	1	3	96	1
865-92-6136	3	7.9125	166.1625	158.25	4.761904762	9.3	1	3	99	1
704-48-3927	10	44.335	931.035	886.7	4.761904762	7.3	1	4	106	1
645-44-1170	9	26.1315	548.7615	522.63	4.761904762	4.3	1	4	114	1
642-32-2990	10	5.48	115.08	109.6	4.761904762	8	1	2	119	1
291-55-6563	6	10.326	216.846	206.52	4.761904762	7.5	1	1	138	1
575-30-8091	8	29	609	580	4.761904762	9.2	1	3	140	1
239-10-7476	6	23.385	491.085	467.7	4.761904762	8	1	2	145	1
685-64-1609	10	15.07	316.47	301.4	4.761904762	9.2	1	1	147	1
249-42-3782	5	17.5025	367.5525	350.05	4.761904762	5.5	1	3	171	1
827-26-2100	9	15.228	319.788	304.56	4.761904762	8.8	1	4	176	1
407-63-8975	6	22.164	465.444	443.28	4.761904762	4.4	1	3	179	1
851-28-6367	10	7.75	162.75	155	4.761904762	8	1	4	183	1
400-60-7251	1	3.7035	77.7735	74.07	4.761904762	9.9	1	3	189	1
856-22-8149	1	1.2645	26.5545	25.29	4.761904762	6.1	1	2	198	1

Table: Fact Supermarket Sales (1,000 rows)

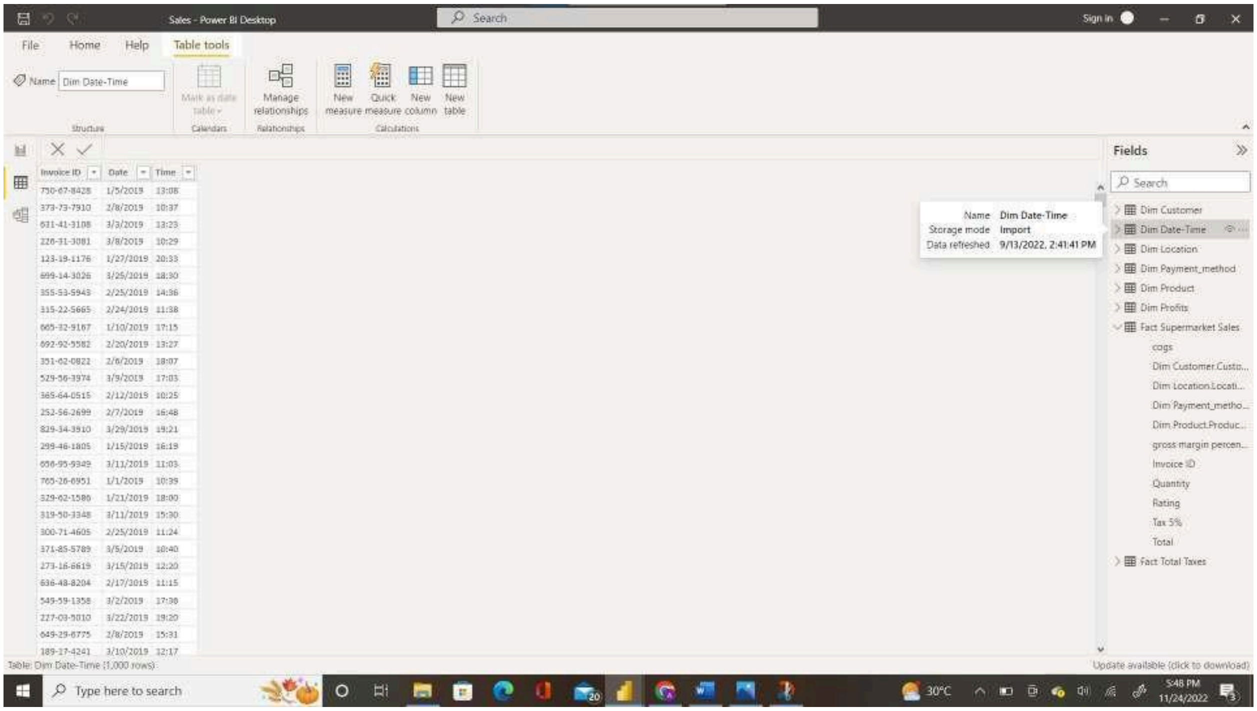
So firstly create a dimension table named **Dim Customer** for the Customers where we have attributes Customer_id, Gender and Customer_Type . After doing that we performed a merge operation on column Customer_id from Supermarket Sales fact table and newly created Customer dimension table and this is how we added a foreign key Dim Customer.Customer_id to Supermarket fact table and deleted the rows Gender and Customer_Type from this table.



Customer type	Gender	Customer_id
Member	Female	1
Normal	Female	2
Normal	Male	3
Member	Male	4

ble: Dim Customer (4 rows)

Similarly we created a dimension table for date named **Dim Date-time** where we added the columns Date and Invoice id, Time to new table and removed the duplicates and given ids to it and merged it with our main table and then deleted the columns from Supermarket Sales fact table

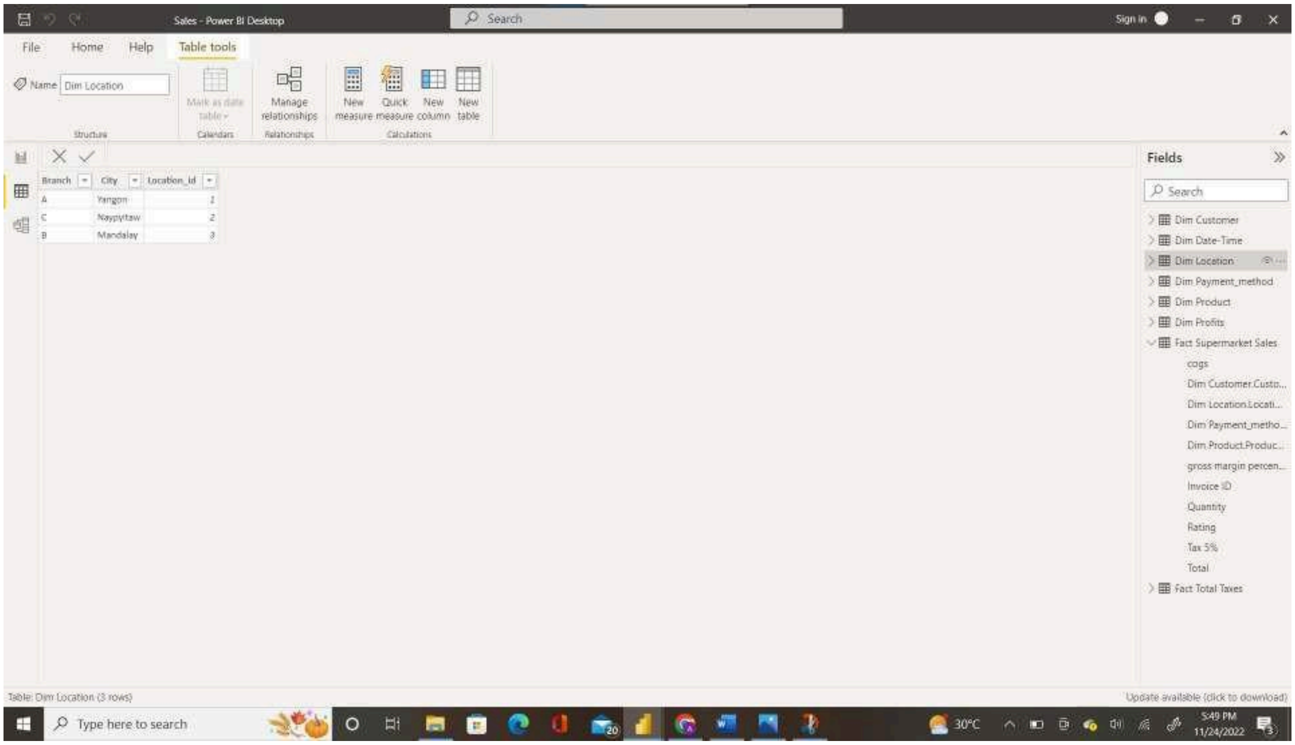


Invoice ID	Date	Time
750-67-8428	1/5/2019	13:08
373-73-7910	2/8/2019	10:37
631-41-3108	3/3/2019	13:23
226-31-3081	3/8/2019	10:29
123-19-1176	1/27/2019	20:33
699-14-3026	3/25/2019	18:30
355-53-5943	2/25/2019	14:36
315-22-5665	2/24/2019	11:38
665-32-9167	1/10/2019	17:15
692-92-5582	2/20/2019	13:27
351-62-0822	2/6/2019	18:07
529-56-3974	3/9/2019	17:03
365-64-0515	2/12/2019	10:25
252-56-2699	2/7/2019	16:48
829-34-3910	3/29/2019	19:21
299-46-1805	1/15/2019	16:19
656-95-9349	3/11/2019	11:03
765-26-6951	1/1/2019	10:39
329-62-1586	1/21/2019	18:00
319-50-3348	3/11/2019	15:30
300-71-4605	2/25/2019	11:24
371-85-5789	3/5/2019	10:40
273-16-6619	3/15/2019	12:20
636-48-8204	2/17/2019	11:15
549-59-1358	3/2/2019	17:38
227-03-5010	3/22/2019	19:20
649-29-6775	2/8/2019	15:31
189-17-4241	3/10/2019	12:17

Table: Dim Date-Time (1,000 rows)

Name Dim Date-Time
Storage mode Import
Data refreshed 9/13/2022, 2:41:41 PM

Similarly we did for Branch and City column and created a new dimension table named **Dim Location** with Branch, City and Location_id.

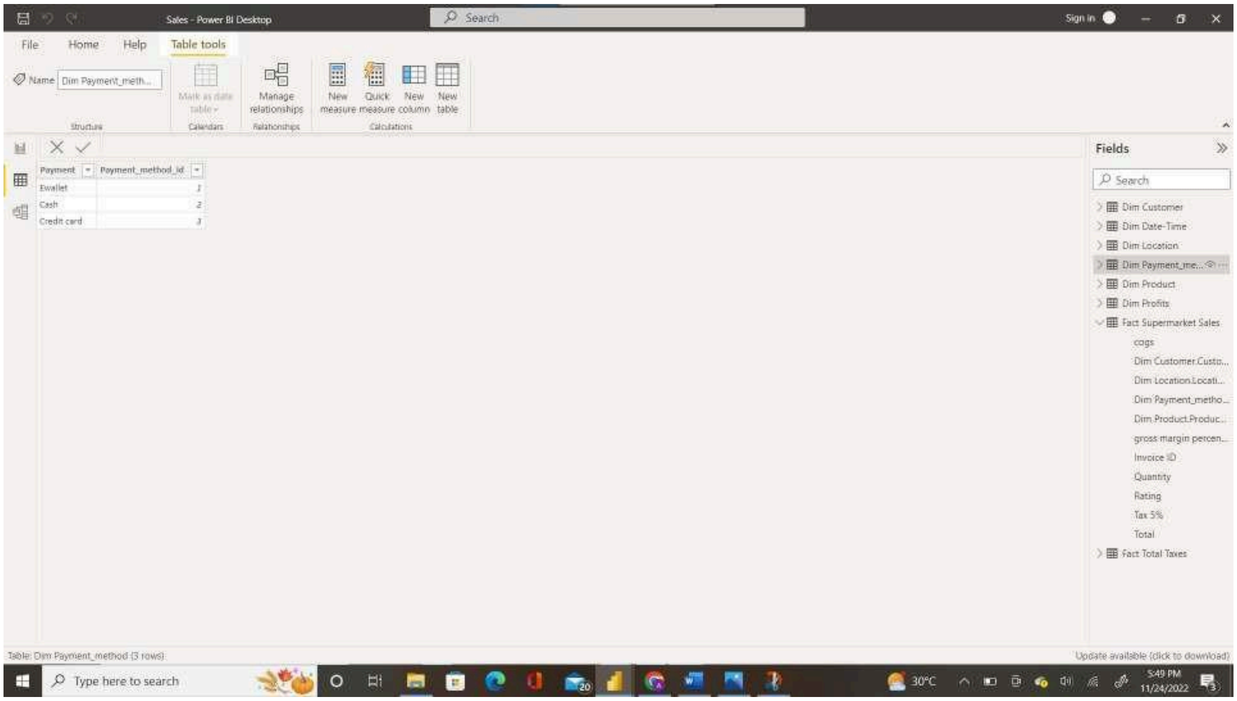


Name: Dim Location

Branch	City	Location_id
A	Yangon	1
C	Naypyitaw	2
B	Mandalay	3

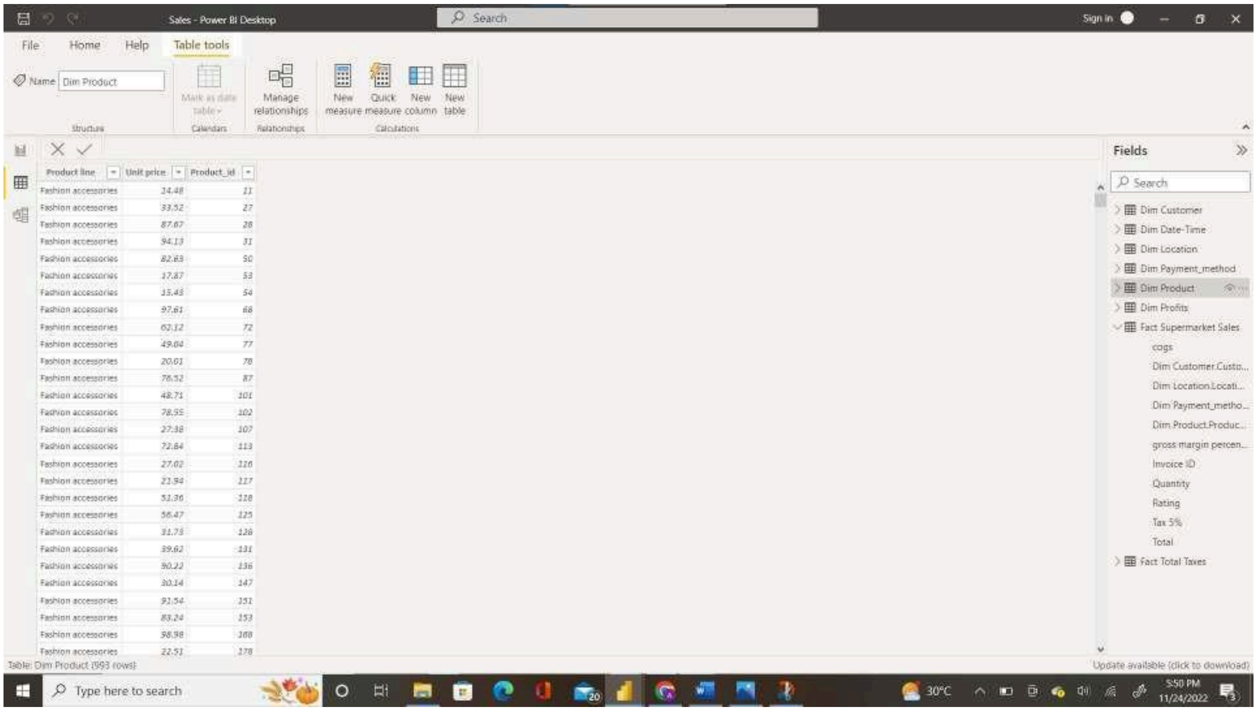
Table: Dim Location (3 rows)

Similarly for payment attributes we have created a new dimension table named **Dim Payment** method with payment and Payment_method_ID as primary key.



Payment	Payment_method_id
Ewallet	1
Cash	2
Credit card	3

We created **Dim Product**, with Product line, Product_id and Unit price as attributes.

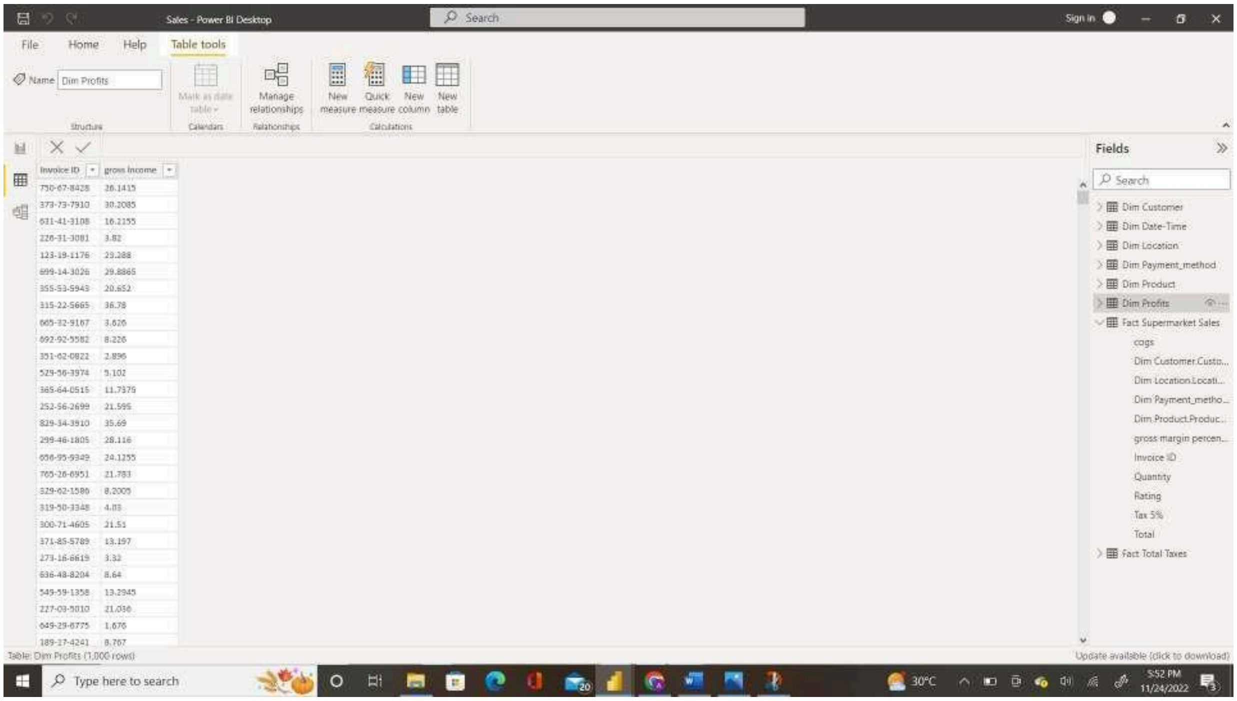


Name: Dim Product

Product line	Unit price	Product_id
Fashion accessories	14.48	11
Fashion accessories	33.52	27
Fashion accessories	87.67	28
Fashion accessories	94.13	31
Fashion accessories	82.63	50
Fashion accessories	17.87	53
Fashion accessories	15.43	54
Fashion accessories	97.61	68
Fashion accessories	62.12	72
Fashion accessories	49.04	77
Fashion accessories	20.01	78
Fashion accessories	76.52	87
Fashion accessories	48.71	101
Fashion accessories	78.55	102
Fashion accessories	27.38	107
Fashion accessories	72.84	113
Fashion accessories	27.02	116
Fashion accessories	21.94	117
Fashion accessories	51.36	118
Fashion accessories	56.47	125
Fashion accessories	31.73	128
Fashion accessories	39.62	131
Fashion accessories	90.22	136
Fashion accessories	30.14	147
Fashion accessories	91.54	151
Fashion accessories	83.24	153
Fashion accessories	98.98	168
Fashion accessories	22.51	178

Table: Dim Product (993 rows)

We created **Dim Profits**, with gross income and Invoice id as attributes.

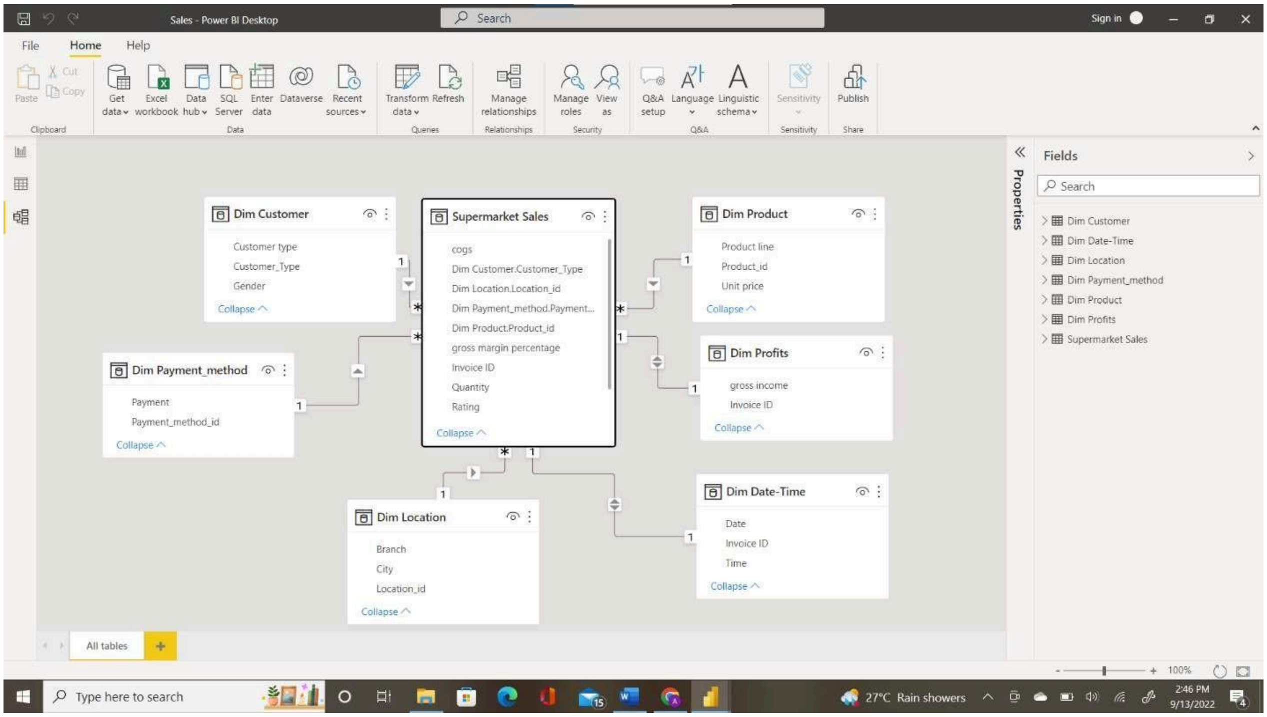


Name: Dim Profits

Invoice ID	gross income
750-67-8428	26.1415
373-73-7910	30.2085
631-41-3108	16.2155
226-31-3081	3.82
123-19-1176	29.288
699-14-3026	29.8865
355-53-5943	20.652
315-22-5665	36.78
665-32-9167	3.626
692-92-5582	8.226
351-62-0822	2.896
529-56-3974	5.102
365-64-0515	11.7375
252-56-2699	21.595
829-34-3910	35.69
299-46-1805	28.116
656-95-9349	24.1255
765-26-6951	21.783
329-62-1586	8.2005
319-50-3348	4.03
300-71-4605	21.51
371-85-5789	13.197
273-16-6619	3.32
636-48-8204	8.64
549-59-1358	13.2945
227-03-5010	21.036
649-29-6775	1.676
189-17-4241	8.767

Table: Dim Profits (1,000 rows)

That completes our Star Schema.



Snowflake Schema:

Snowflake Schema is basically a variation of Star Schema which is represented by centralized Fact Tables which are connected to multiple Dimension Tables. But the difference between Star schema and Snowflake schema is that in Snowflake schema large dimension tables are normalized into multiple tables.

Creating Snowflake Schema:

For creating Snowflake schema we have copied Invoice Id attributes from the Dim Profits table into another table named as Dim Date-Time table. After this we have removed duplicates from attributes and created a new column Invoice ID that acts as the primary key into the table and linked the table with the Dim Date-Time table. That completes our Snowflake Schema.

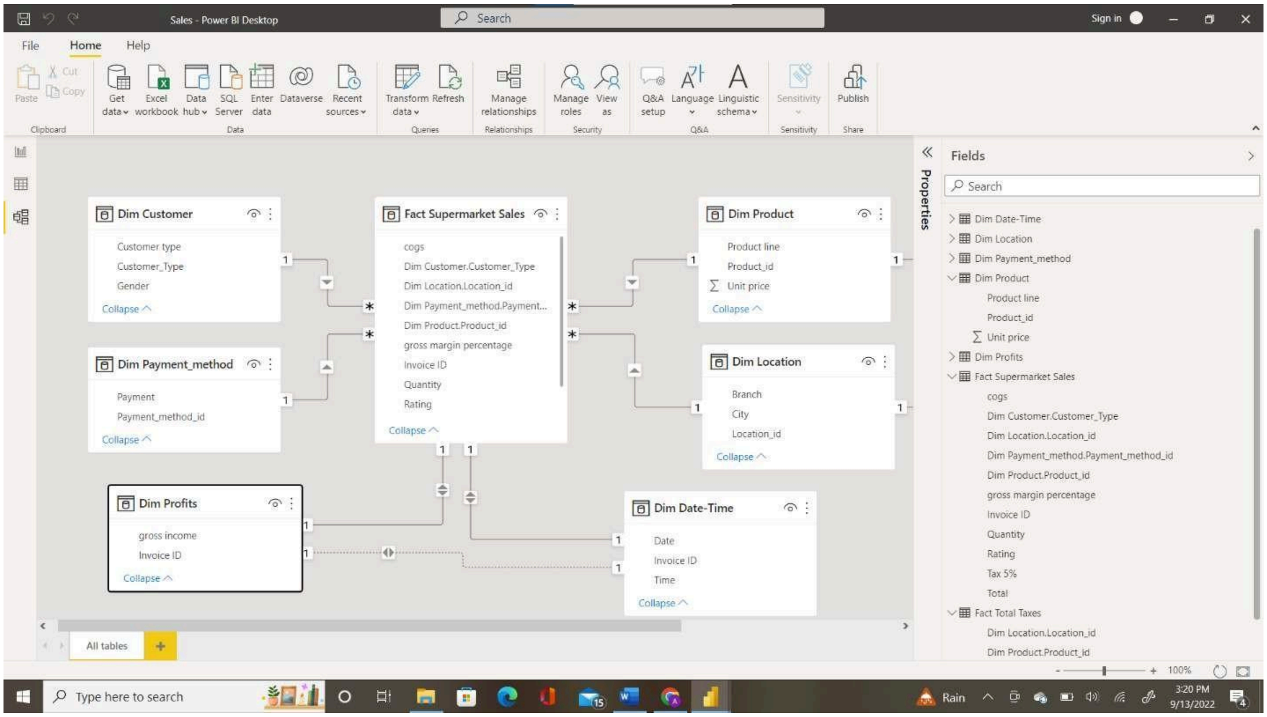
Fact Constellation Schema:

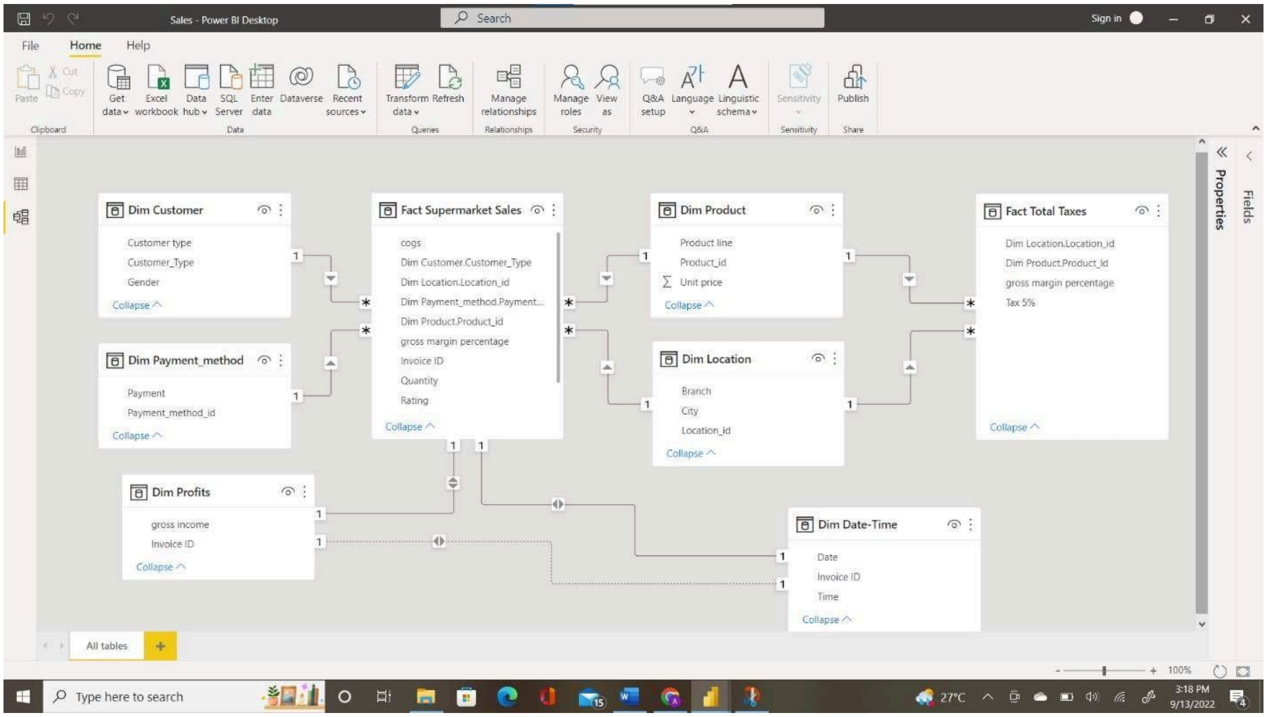
Fact Constellation is a schema for representing a multidimensional model. It is a collection of multiple fact tables having some common dimension tables. It can be viewed as a collection of several star schemas and hence, also known as **Galaxy schema.**

Creating Fact Constellation Schema:

For creating Fact Constellation Schema we have created a new Fact Total taxes, which acts as another Fact Table for our schema. In the Dim Product dimension table which is linked to Fact Supermarket Sales Table there is one Product id attribute which is also present in Fact Total Taxes Table, hence we have linked the Dim Product table to the Fact Total Taxes Table also. That means the Dim Product table is now linked to both Fact Tables.

Similarly In the Dim Location dimension table which is linked to Fact Supermarket Sales Table there is one Location id attribute which is also present in Fact Total Taxes Table, hence we have linked the Dim Location table to the Fact Total Taxes Table also. That means the Dim Location table is now linked to both Fact Tables.





Conclusion:

Thus, we have successfully preprocessed and loaded the dataset into the Power BI tool and designed Star, Snowflake and Fact Constellation Schemas.

RESULTS

1. **Business Process Chosen:**

Health and Lifestyle Analysis in the Food Industry – focusing on analyzing food preferences, eating habits, lifestyle choices, and their impact on student well-being.

2. **Schema Design:**

Designed a **Star Schema** with the central **Fact Table** capturing student food behavior metrics (e.g., calories, meals out, nutritional checks) and **Dimension Tables** including student demographics, eating habits, cuisine preferences, family background, and health indicators.

3. **ETL Process:**

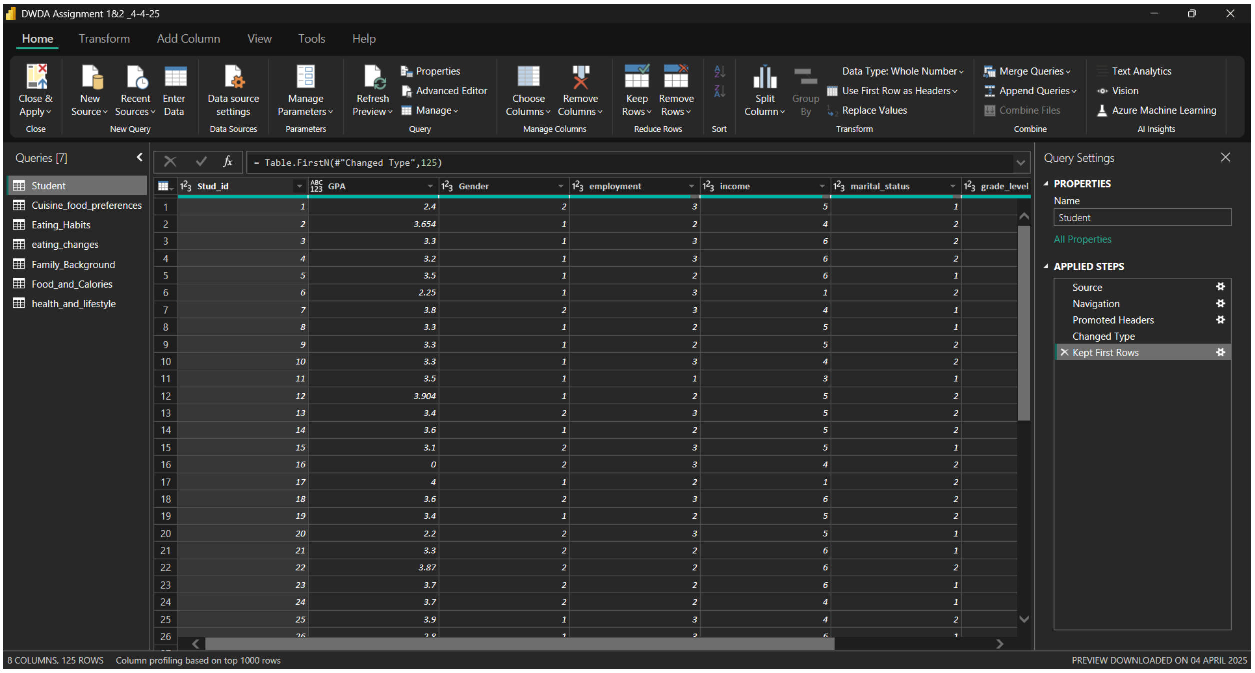
I. **Extracted** data from a publicly available dataset using Power BI's **Power Query Editor.**
II. **Transformed** data by handling nulls, correcting data types (e.g., GPA), removing unnecessary columns, and filtering irrelevant rows (e.g., from row 125 onwards).
III. **Loaded** the clean data into Power BI's model and established relationships across the tables to create an optimized data warehouse structure.

4. **Tools Used:**

I. Power BI (Data loading, transformation, schema design, and visualization)
II. Power Query Editor (for ETL operations)

Results Attached:

I. Data Loading in Power BI

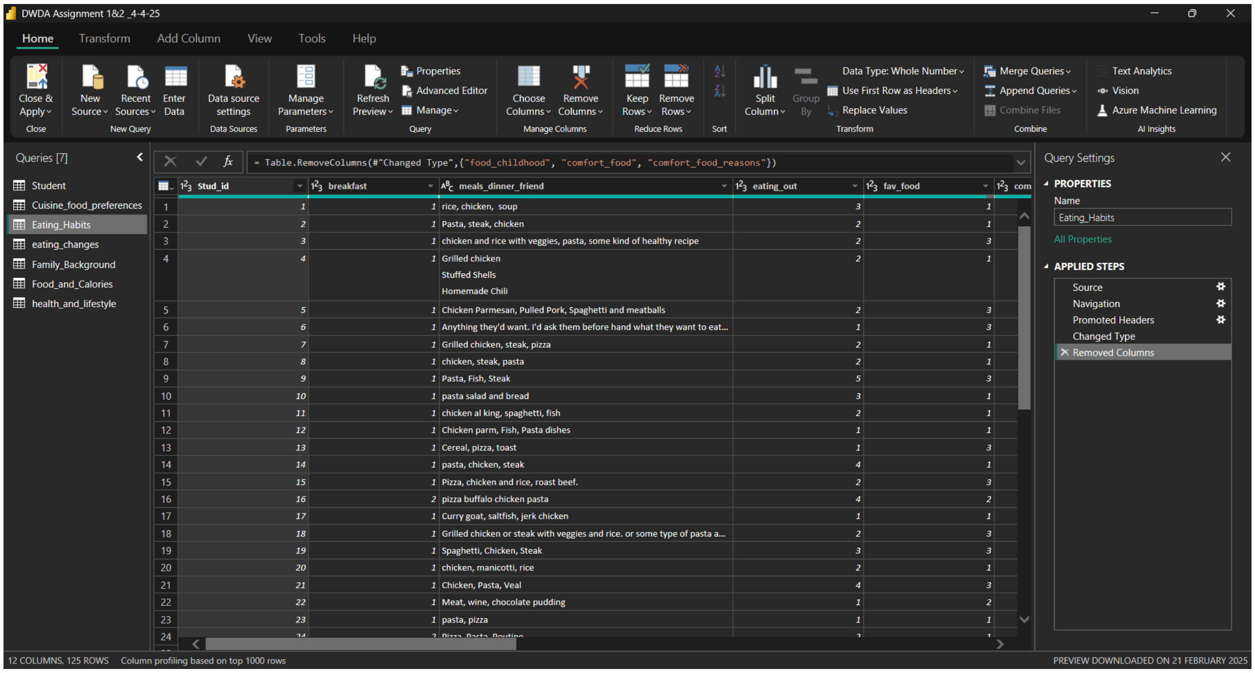


```
= Table.FirstN(#"Changed Type",125)
```

	Stud_id	GPA	Gender	employment	income	marital_status	grade_level
1	1	2.4	2	3	5	1	
2	2	3.654	1	2	4	2	
3	3	3.3	1	3	6	2	
4	4	3.2	1	3	6	2	
5	5	3.5	1	2	6	1	
6	6	2.25	1	3	1	2	
7	7	3.8	2	3	4	1	
8	8	3.3	1	2	5	1	
9	9	3.3	1	2	5	2	
10	10	3.3	1	3	4	2	
11	11	3.5	1	1	3	1	
12	12	3.904	1	2	5	2	
13	13	3.4	2	3	5	2	
14	14	3.6	1	2	5	2	
15	15	3.1	2	3	5	1	
16	16	0	2	3	4	2	
17	17	4	1	2	1	2	
18	18	3.6	2	3	6	2	
19	19	3.4	1	2	5	2	
20	20	2.2	2	3	5	1	
21	21	3.3	2	2	6	1	
22	22	3.87	2	2	6	2	
23	23	3.7	2	2	6	1	
24	24	3.7	2	2	4	1	
25	25	3.9	1	3	4	2	

8 COLUMNS, 125 ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED ON 04 APRIL 2025

Table : Student_Info



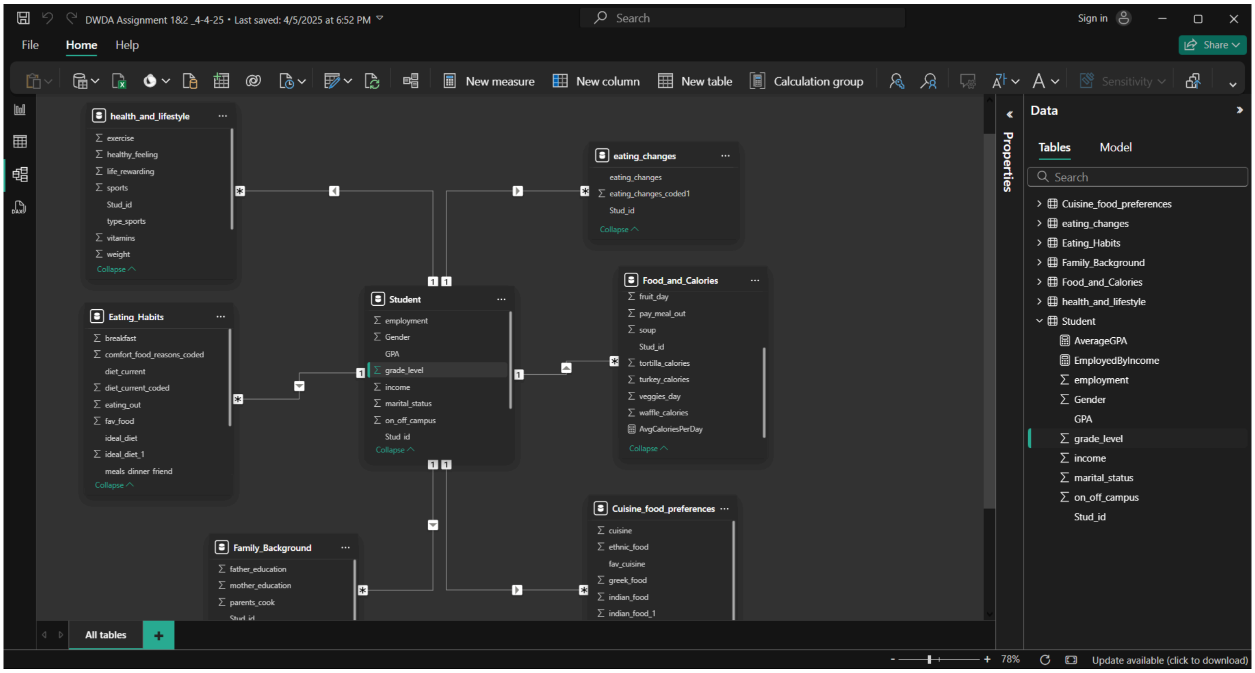
```
= Table.RemoveColumns(#"Changed Type",{"food_childhood", "comfort_food", "comfort_food_reasons"})
```

	Stud_id	breakfast	meals_dinner_friend	eating_out	fav_food	com
1	1	1	rice, chicken, soup	3	1	
2	2	1	Pasta, steak, chicken	2	1	
3	3	1	chicken and rice with veggies, pasta, some kind of healthy recipe	2	3	
4	4	1	Grilled chicken Stuffed Shells Homemade Chili	2	1	
5	5	1	Chicken Parmesan, Pulled Pork, Spaghetti and meatballs	2	3	
6	6	1	Anything they'd want. I'd ask them before hand what they want to eat...	1	3	
7	7	1	Grilled chicken, steak, pizza	2	1	
8	8	1	chicken, steak, pasta	2	1	
9	9	1	Pasta, Fish, Steak	5	3	
10	10	1	pasta salad and bread	3	1	
11	11	1	chicken al king, spaghetti, fish	2	1	
12	12	1	Chicken parm, Fish, Pasta dishes	1	1	
13	13	1	Cereal, pizza, toast	1	3	
14	14	1	pasta, chicken, steak	4	1	
15	15	1	Pizza, chicken and rice, roast beef.	2	3	
16	16	2	pizza buffalo chicken pasta	4	2	
17	17	1	Curry goat, saltfish, jerk chicken	1	1	
18	18	1	Grilled chicken or steak with veggies and rice. or some type of pasta a...	2	3	
19	19	1	Spaghetti, Chicken, Steak	3	3	
20	20	1	chicken, manicotti, rice	2	1	
21	21	1	Chicken, Pasta, Veal	4	3	
22	22	1	Meat, wine, chocolate pudding	1	2	
23	23	1	pasta, pizza	1	1	

12 COLUMNS, 125 ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED ON 21 FEBRUARY 2025

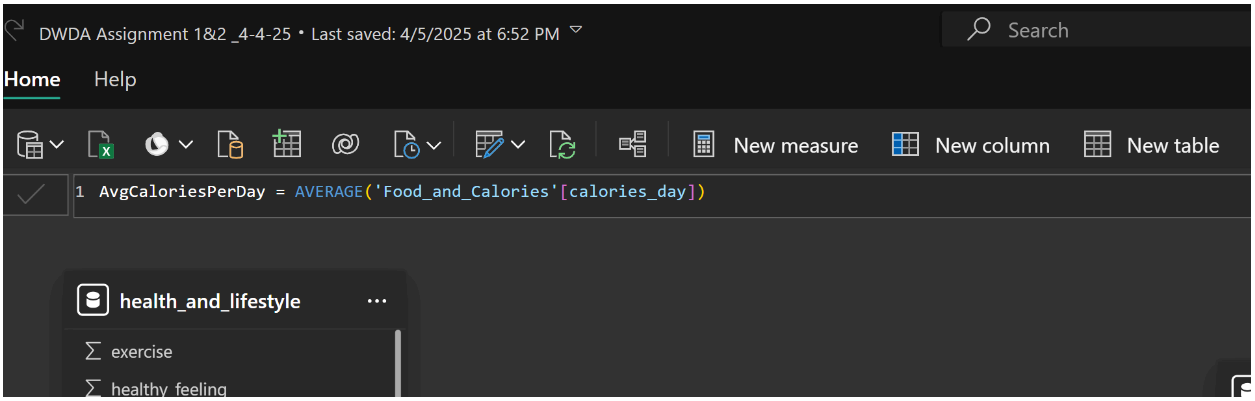
Table : Eating_Habits

II. Star Schema Diagram (Model View)

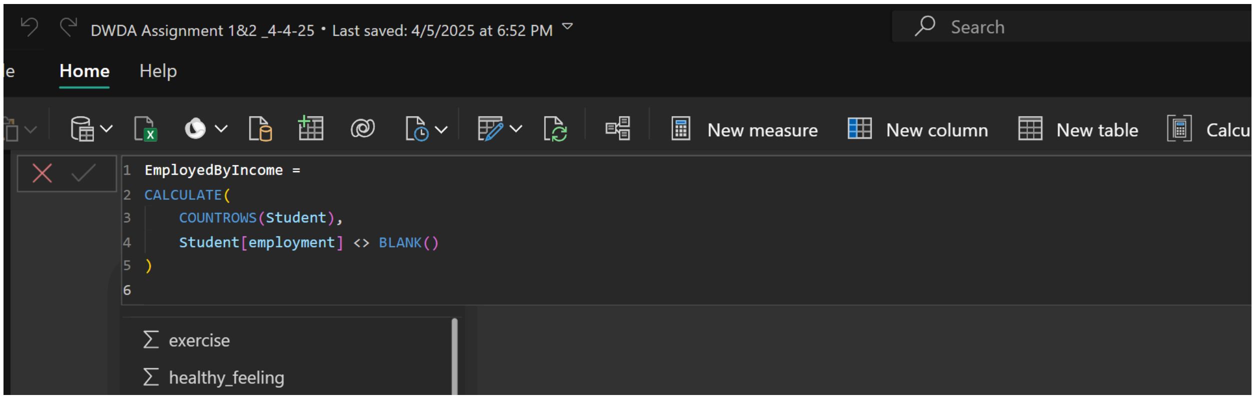


Star Schema

III. Creating new measure using Power Query



```
AvgCaloriesPerDay = AVERAGE('Food_and_Calories'[calories_day])
```



```
EmployedByIncome =
CALCULATE(
    COUNTROWS(Student),
    Student[employment] <> BLANK()
)

```