Designing and Implementing a Data Warehouse

Background

Galleria Holdings is a fast-food chain which originated in Italy but has now acquired a number of businesses in the UK. The company believes that the UK menus are too large and inconsistent across the outlets they have acquired. They also believe that the menus should be consolidated to suit the preferences of the customers and therefore reduce unnecessary purchases of food which is not chosen by customers.

In order to achieve this, the parent company wishes to create a data warehouse for Business Intelligence purposes.

They wish to analyse sales to determine the most **popular menu items** and also those **which produce the most revenue**. They also wish to find out the most **popular product groups** and those which **provide the most revenue**. Finally, they wish to **construct a “league table”** of the total **weekly and monthly sales** of the various outlets so that their sales performance can be monitored easily.

**The project brief is as follows:**

* Review the case study and identify the business requirements from a reporting perspective. Your overall objective is to design a data warehouse capable of reporting on the questions the business wish to answer.
* Perform an assessment of the initial dataset to:
  + Understand the data items and how they relate together.
  + Identify any data quality issues which could affect your implementation.
* Design the schema to satisfy the requirements.
* Implement the ETL to create and populate the tables in PowerBI ensuring all counts reconcile.
* Create a set of visualisations to meet the reporting requirements contained in the brief.

**Detailed requirements**

In its simplest terms, your task can be broken down into six parts which should be evidenced as a report that includes screen shots and commentary.

Part 1: Problem Domain Understanding

**Task 1**

Given the nature of the business you have been asked to conduct a comparative analysis of on-premises, cloud-based, and hybrid data engineering solutions e.g., ETL and ELT.

The evaluation criteria should consider the following:

* Security
* Compliance
* Scalability
* Efficiency
* Reliability
* Fidelity
* Flexibility
* Portability

Cloud based solutions are the new, more efficient ways to run a business for handling files and data. For Galleria Holdings, Software as a Service (SaaS) provides a simple yet effective solution that will allow the managers of the company to use, without any coding or technical knowledge.

Table of comparisons (Green = Pros, Red = Cons, Orange = Both good and bad depending on context):

|  |  |  |  |
| --- | --- | --- | --- |
| Evaluation Criteria | **On Prem** | **Cloud (SaaS)** | **Hybrid (SaaS)** |
| **Security** | Company has full control (dependant on your own resources and security levels both physically and virtually) | Your cloud provider will be responsible for the physical security of the datacentre.  SaaS is 'multi-tenat' - multiple end user customers sharing same database and only separated by the internet security. Therefore there is an open risk if this security is not up to date. | Very secure as can allow the company to keep extra ‘high-risk’ files such as customer contacts on prem and have other data stored on cloud. |
| **Compliance** | Need to ensure that you meet all compliance regulations, including with the physical datacentre as well as the data being stored itself. | You are responsible for the type of data that is being stored and should ensure they meet compliance. This Including card details from transactions In outlet stores.  However, the amount you are responsible is less than on prem.  Azure services can also provide advice to ensure your data meets the compliance of the country data is stored in. | You have responsibility for both the data in your on prem and cloud storage, as well as the datacentre storage for on-prem. |
| **Scalability** | There is a lot of capital expenditure for data centres for when the company decides to expand or open new outlets and will handle larger sets of data. This Includees buying the tech as well as ensuring the company has enough office space. | The flexibility of cloud allows you to easily scale up your SaaS's, without extra capital expenditure In comparison to on-premises. This may not only work out cheaper, but also requires less work and a much faster solution. | Hybrid allows you to scale up on the cloud, whilst still having the other benefits that come with on-prem. |
| **Efficiency** | Limits your capacity and solutions based on your own resources and IT team.  Can also be more expensive as there is capital expenditure as well as maintenance. | Cloud solutions can be implemented by the staff who do not have IT technical skills. This will also allow outlet managers to have easy access to their own stats and data (including league tables).  Cloud solutions are also more cost-efficient as there are only running-costs. | When moving from on prem to cloud, hybrid is an efficient solution to keeping the process smooth and allows the company to ease into cloud. |
| **Reliability** | Your data solely relies on the company and its solutions for back up and protection. | Several availability zones means that natural disasters in one area will not affect your data. | There is still running the risk of having issues with your data on your onprem |
| **Fidelity** | Data and solutions will come from and within your premises and company, with none to minimal handling from other services. This should guarantee a high level of fidelity. | Your processes and data will be handled by different services, depending on the nature of your solution. These should still be reliable, but will not fully depend on the quality of your process. | access to cloud solutions, whilst maintaining control and the possibility to apply your own private solutions as needed. |
| **Flexibility** | The flexibility in what you will be able to do is dependable on the skill and expertise of your IT team, as well as expenditure. | Greater range of simple solutions for incoming tasks and variations for individuals with very little IT knowledege. | access to cloud solutions, whilst maintaining control and the possibility to apply your own private solutions.  You are able to decide what stay on prem and what is on the cloud. |
| **Portability** | Datacentres cannot be moved, unless at a very high cost. This may be very inconvenient if the company expands more after changing the menus, as they will not be able to move office very easily. | When using a cloud provider, datacenters do not have to be moved elsewhere In case Galleria Holdings decide to change their office location, as all data Is stored In the cloud datacentre. | Able to move around data when needed, but still would need to keep on prem data in the same location |

**Task 2**

**Review the case study and identify the business requirements from a reporting perspective and identify additional relevant business questions for reporting purposes.**

*Proposed problems for Galleria Holdings:*

- The menus are too large and Inconsistent across the UK outlets

- The company Is making unnecessary purchases of food

- The company does not know their most popular food Items and Items that produce the most revenue

- The company does not know their most popular food groups and groups that produce the most revenue

*Objectives:*

- Reduce the number of unnecessary purchases of food by the company that make Items that are not popular with customers

- Determine the most and least popular food Items and groups based on number of purchases and revenue

- Find what outlets do better (in sales) than others on a weekly and monthly basis to keep track of top performing outlets

*Deliverables:*

- Reduce food wastage

- Advertise popular Items on new menus

- Create a new menu that can be consistently used across all outlets, which Include the most popular Items and food groups

- Create a league tables that record the outlets' sales on a weekly basis

- Create a league table that records outlet sales on a monthly basis

*Scope:*

- Should take approx 4-5 days

- Complete by 10/10/2022

*Resources Needed:*

- Data Engineer

- Data Analyst

- Sales Representative

Part 2: Data Understanding

**Task 1**

**Perform an assessment of the initial dataset to:**

* **Understand the data items and how they relate together.**

The original data has columns mentioned below and my understanding of what data they present.

*SaleDate*: When items were sold

*Ticket Number:* ID for the sale

*Outlet*: City the branch is in

*Total*: Total price paid by cutsomer

*OrderQty*: quantity of purchased item

*Stock\_code*: unique ID for item

*Name*: name of item

*Description*: describes the item and what ingredients the item includes

*Price*: price of 1 unit

*Product group*: ID for food groups

*Group\_name*: full name for food groups

*CardType*: type of payment card used for transaction

*SaleDate*: when item was sold

The source dataset is a sales transactions dataset, where each row represents the sale of an item, including information of the specific sale as defined above.

* **Identify any data quality issues which could affect your implementation.**

There Is a few columns In the data that Is not relevant to our objectives - Like the card type used to pay for Items and description. The description column also has some NULL values. These columns were therefore deleted.

Graphical user interface, application

Description automatically generated

Some column titles had an underscore between 2 words, however some did not. The second words in some columns were also not capitalised. Therefore, we renamed columns to not have underscores and stay consistent with the format: 'FirstLast'

Renamed “Stock\_Code” -> Stock Code

Product\_Group -> ProductGroup

Group\_name -> GroupName

To make some columns more clear to understand during analysis, columns, the following columns were also renamed:

Name -> ProductName

Total -> Revenue

The different stock codes were given to different sizes of Latte, but were both still titled the same: ‘Latte’. Therefore, we changed the value with the DBA40 stock code to be ‘Latte Large’

Table

Description automatically generated

We then double checked if the item was renamed, as it could be detrimental to affecting our analysis.

Graphical user interface, application

Description automatically generated

We also changed the formatting of the columns to appropriately suit the data. This was completed In the model view -> format -> currency (for price and Revenue)

Graphical user interface, application

Description automatically generated

The outlet location format was changed to city. (In model view, go to Advanced -> Data Category -> City)

Part 3: OLAP Schema Design

**Task 1**

**Identify and justify your OLAP schema design.**

The source data Is a big dataset, with over 40000 rows; which can require a lot of storage as sales Increase. The OLAP schema design will allow us to store the data In a more storage-efficient way.

Our snowflake schema will consist of a central fact table with connecting dimension tables.

The facts of the data are the sales Galleria Holdings made. These can be found In the fact table, which Includes the ID's of the different attributes, and how each attribute combines to create a fact. You can see below the fact table we Intend to create:

Application

Description automatically generated with low confidence

Dimension tables Include the definitions of the ID's and what categories they relate to. The following tables are the dimension tables we Intend to create:

Graphical user interface, table

Description automatically generated

Table

Description automatically generatedTable

Description automatically generated with medium confidenceTable

Description automatically generatedTable

Description automatically generated

- DimDate gives the ID as well as an heirarchy of the date, allowing extractions of specific aspects of the date during analysis.

- DimProduct provides the ID and defines what products relate to the ID's, along with their stock code, price and relating group ID.

- DimOutlet defines what outlets link to the ID's

- DimTicket defines what ticket numbers link to the ID's

- DimGroup defines what Group each ID represents along with their unique code.

As product and food group relate to each other, DimGroup will be connected to DimProduct, where DimProduct Is connected to FactSales.

**Task 2**

**Design and implement the schema in PowerBI to satisfy the requirements.**

The snowflake schema we intend to pursue:

Diagram

Description automatically generated

Where all connections are many to one starting from the fact table and spreading to the dimension tables. The connection from Products to Food Groups is also many to one.

The schemas were then implemented into PowerBI after loading and transforming, where PK = Primary Key for that table and FK = Foreign Key

Diagram

Description automatically generated

Part 4: Data deliverables

**Task 1**

**Produce a document (Source to Target Mapping) that contains the mapping of source system fields to the fields of the target system.**

Graphical user interface, application, table, Excel

Description automatically generated

This diagram portrays the differences made to the original source data columns, either mapped onto the new tables, or deleted.

**Task 2**

**Create the tables in PowerBI.**

The following tables were created using the PowerBI dashboard and then exported as CSV files (From left to right on the image below).

1. DimGroup

2. DimDate

3. DimTicket

4. DimOutlet

5. DimProduct (Please note that an extra column: GroupName was also added to this table before exporting but Is not shown In the picture below)

Table

Description automatically generated

**Task 3**

**Extract Transform and Load the data using PowerBI.**

Importing csv files onto PowerBI:

**Graphical user interface, application, table

Description automatically generated**

Before transforming the data, I ensured that there was no duplicate data by right-clicking on a column title In each table and then removing any duplicates.

Graphical user interface, text, application

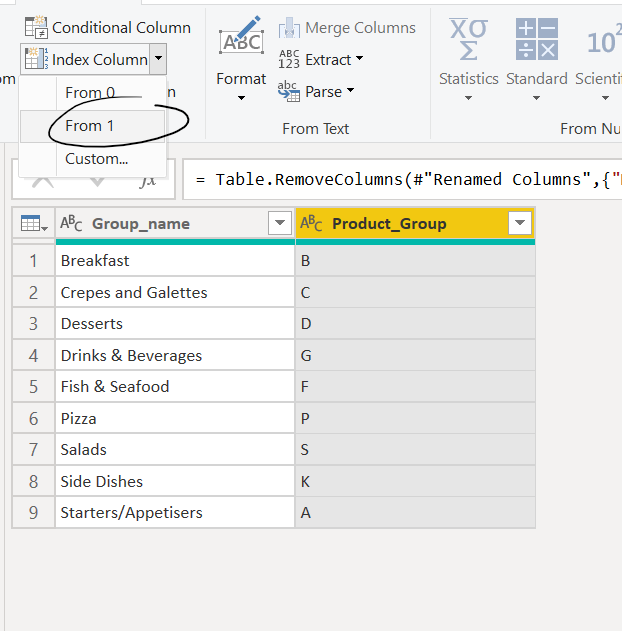
Description automatically generated

The DimDate table was transformed to Include other Information about the date, Including: Year, Quarter, Month Number (MonthNo), MonthName, WeekofYear, WeekBeginning. This was completed by duplicating the original ‘SalesDate’ column and then right-clicking the title of the new column to transform it.

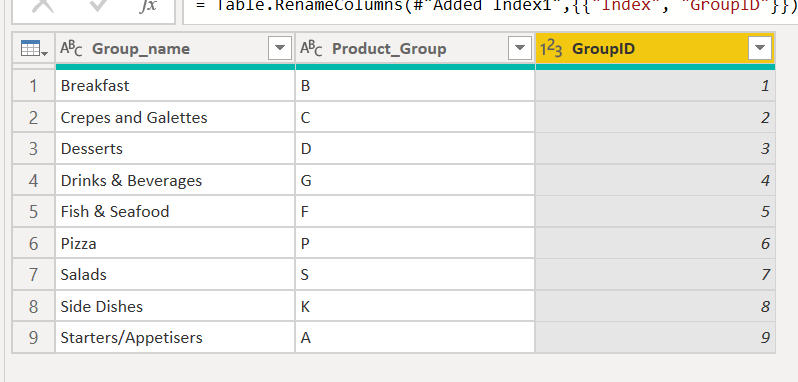
Table

Description automatically generated

For each dimensions table, I added an Index column In the 'add column tab', as shown below:

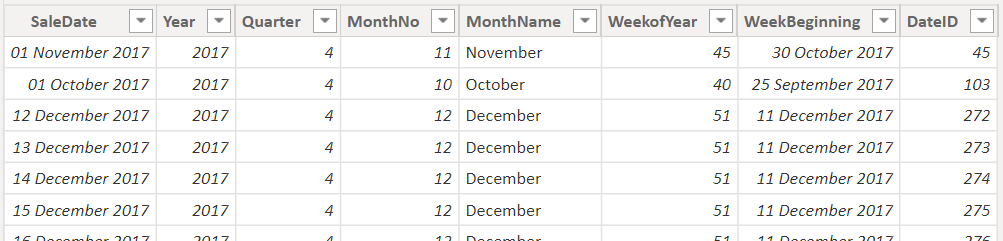


This index column was then renamed to \*name\*ID.

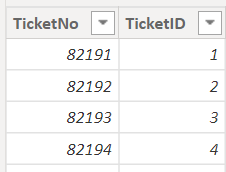


The final dimension tables:

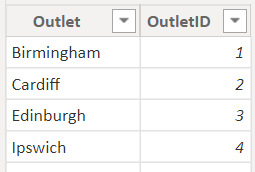
DimDate



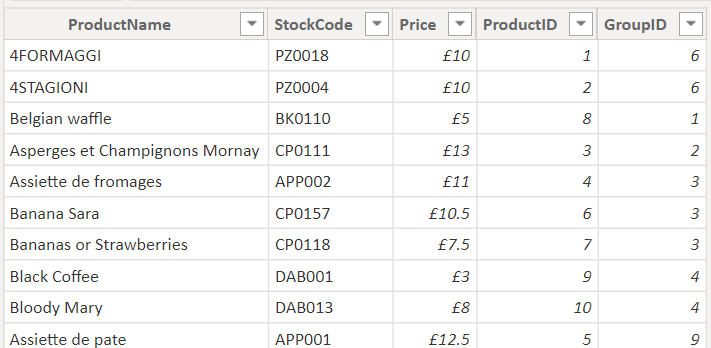
DimTicket



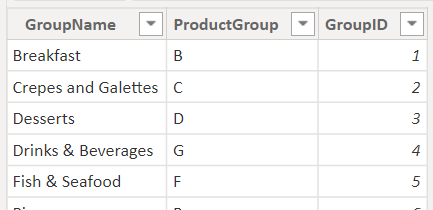
DimOutlet



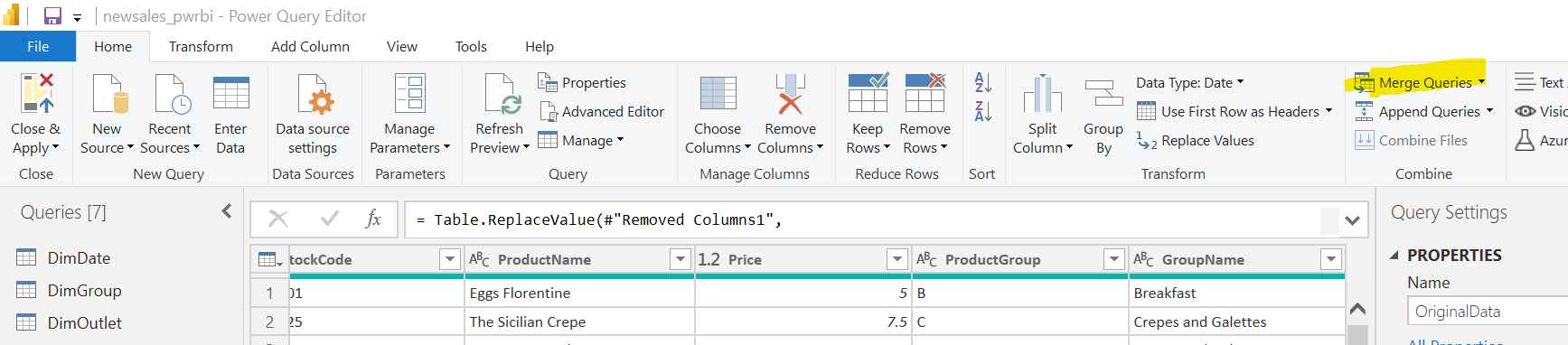
DimProduct (Deleted the GroupName after merging with the DimGroup table to get the GroupID)



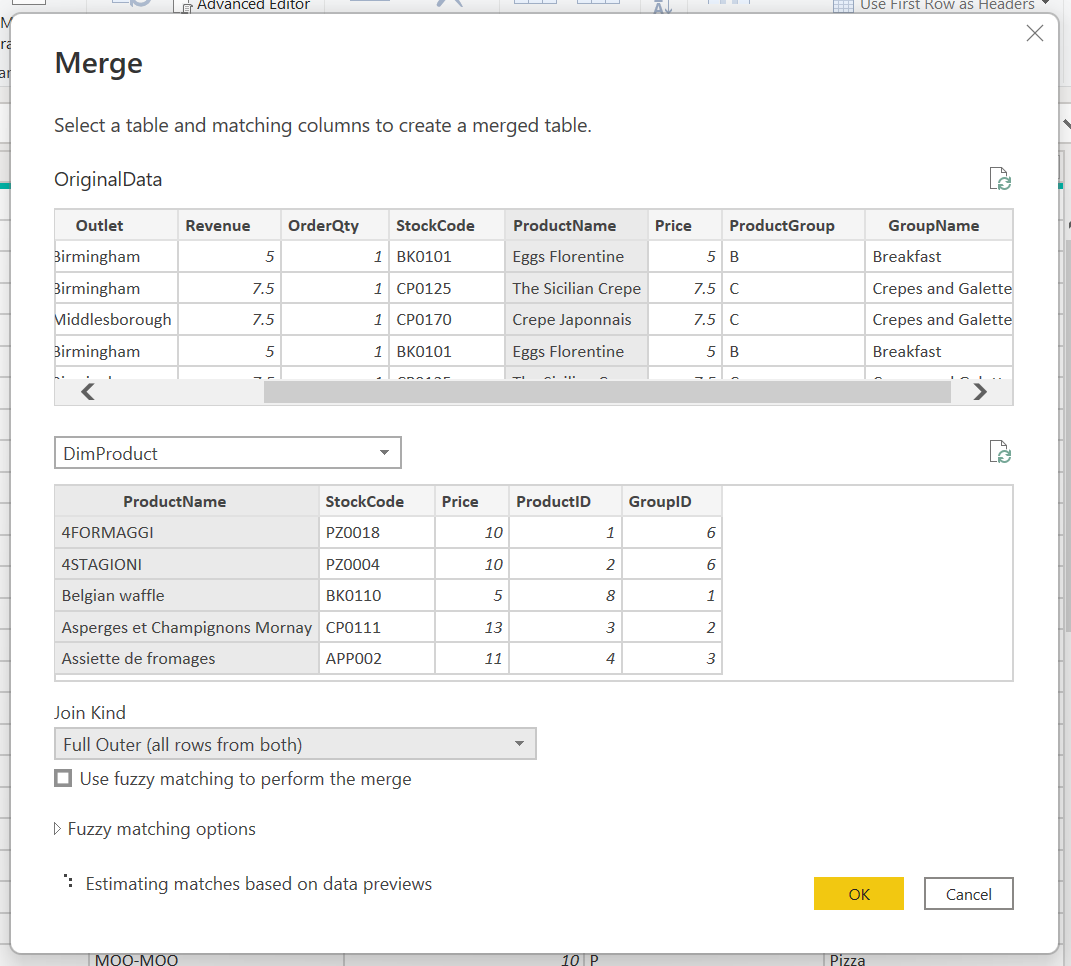
DimGroup



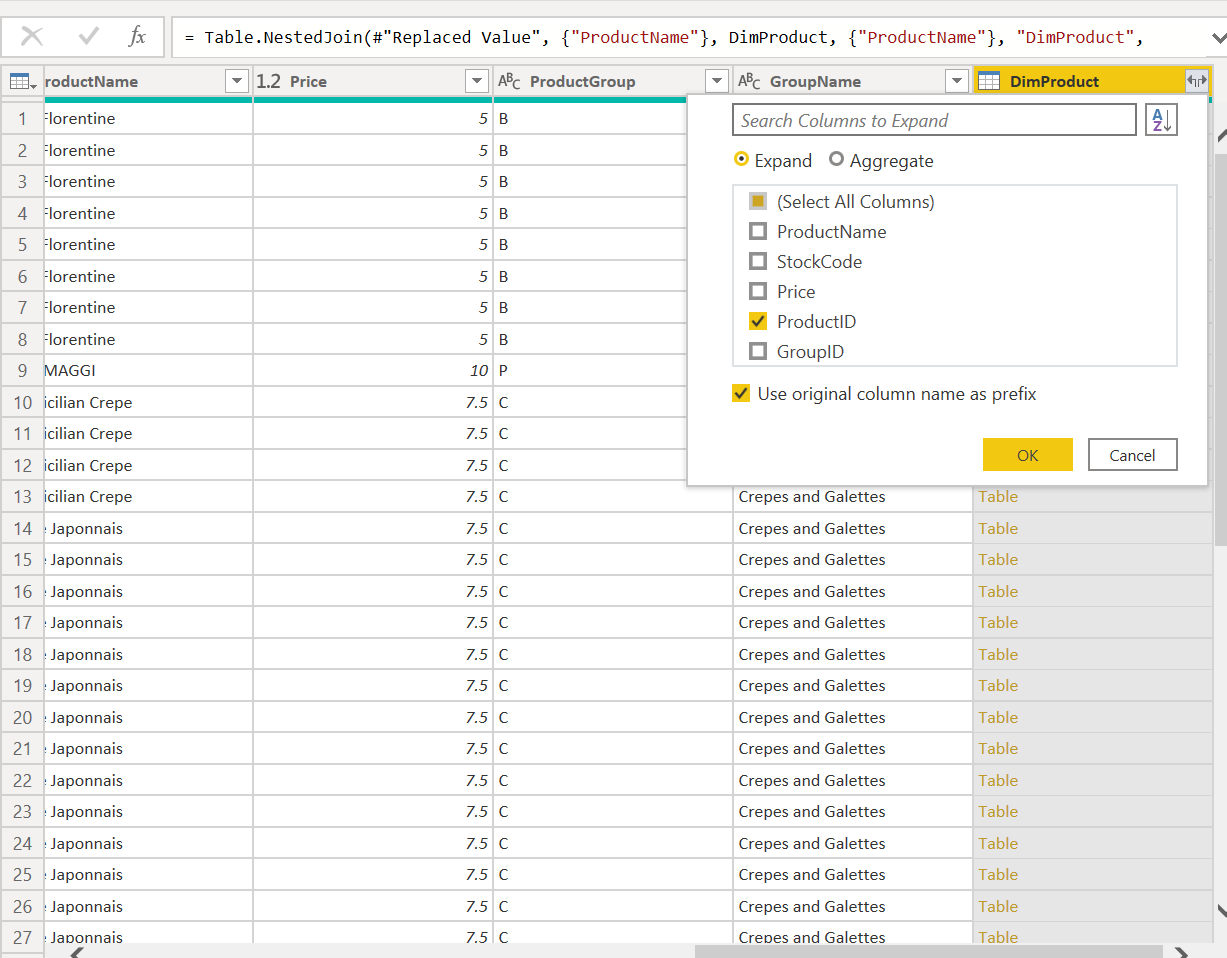
The cleaned and refined source data is duplicated. This duplicate was our fact table. We merged the fact table and the dimension tables (apart from DimGroup) based on a common columns between the two. For example, to connect the DimProduct to FactSales, FactSales was merged with DimProduct based on ProductName, using the following steps:



Use these settings:



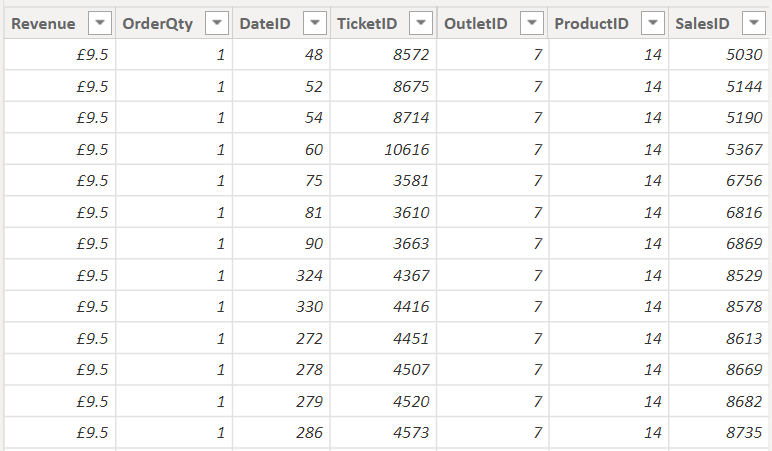
This should create a new column in our FactTable. Select the grey button in the column title and only show the relevant ID. Then rename the column to \*name\*ID (ProductID in this example).



Completed these steps for all dimension tables apart from DimGroup. DimGroup was merged with DimProduct instead of the fact table to create a link between the two dimension tables.

After creating all the new columns in the fact table, I deleted the original columns that weren’t the ID’s, apart from revenue. The connections were then created automatically by powerBI.

The final fact table looked like this:



Part 5: Test approaches

**Task 1**

Perform some tests to check the four different levels of data testing.

**Entity Level:**

**Review the reporting requirements that were listed as part of the initial data warehouse project brief. Can your design support the requirements?**

Our original objectives/ requirements:

* Reduce the number of unnecessary purchases of food by the company that make Items that are not popular with customers
  + Our analysis discovers the bottom 10 products and food groups based on quantity sold. This gives Insight into what products or particular food groups will need to be taken off from the original menu, as these items are not particularly popular with customers. This will reduce the amount of unnecessary food purchases made by Galleria Holdings.
* Determine the most and least popular food Items and groups based on number of purchases and revenue to help create a new, consistent menu.
  + The most popular and least popular Items were determined In our analysis by finding the bottom and top 10 Items and food groups. These top 10 Items will be the most highly recommended to keep on the menu, whereas the bottom 10 will be the least recommended to keep on the menu.
* Find what outlets do better (in sales) than others on a weekly and monthly basis to keep track of top performing outlets
  + Our league tables during analysis shows that the top performing outlet of the month (July 2019) Is Birmingham. Our weekly league tables reveal that the top performing outlet of the last recorded week Is Edinburgh. Further detail Into the wider scope of outlet performance can be found In part 6, task 1.

**Record Level:**

Run some queries to get the counts for each of your populated tables.

Look back at the original source data and analyse what each entity count should be.

Do your counts align? If not, consider what needs to be fixed.

Our counts for the fact table align with the source data at 40705 as the Images below portray:

Table

Description automatically generated

Table

Description automatically generated

Our entity count for dimension tables also align with the source data, when counting unique Items.

Source data entity count:

Graphical user interface, application, Word

Description automatically generated

Dimension table entity counts:

Count of Date dimension table:

Table

Description automatically generated

Count of Outlets:

Graphical user interface

Description automatically generated with medium confidence

Count of Products: 78

These do not align with original source data as we changed the title for one of the items which adds 1 more unique count from 77 to 78.

Graphical user interface

Description automatically generated with medium confidence

Count of Product Groups: 9

Graphical user interface, application

Description automatically generated

Count of Tickets: 16708

Table

Description automatically generated

**Column Level:**

Check your table structures against the original source data. Have you accounted for all the data items?

No, not all data Items are Included In order to Improve the quality and clarity of the data. For example, the CardType and Description of Items were removed.

However, other than these columns, all other data items can be found In our snowflake schema model.

Diagram

Description automatically generated

The original data had the following columns:

Table

Description automatically generated

**Column Value Level:**

Run some queries to get frequency counts of your columns. As a minimum, pick one column per table and do a frequency count of that column.

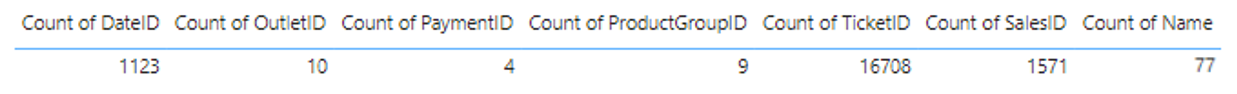
Do the same in your source data. Do the values match?

Yes, the values match. Please see below the queries we ran to test whether the counts for our dimension and fact tables matched up with the source data.

Source data counts:



Dimension and fact table counts:



Part 6: Measures and Visualisations

**Task 1**

Create visualisations to meet the original reporting requirements. Include any additional visualisations to support the extra business questions you identified in Part 1.

Top 10 food Items based on revenue and quantity sold:

Chart, bar chart

Description automatically generated

The food Item that produces the most revenue Is Crepe Japonnais, at approx £62k.

Chart, bar chart

Description automatically generated

The food Item that is ordered the most by customers is Crepe Japonnais, at approx 8.2k.

This analysis shows that the Crepe Japonnais Is currently the best, most popular item on the menu. Furthermore, there Is a lot of common Items between the two graphs, Indicating that these Items should be kept on the new menu across all outlets.

Bottom 10 products by revenue and sales made:

Chart, bar chart

Description automatically generated

The product to produce the least revenue Is the beef bourguinon, only having a total revenue of approx £230.

Chart, bar chart

Description automatically generated

The worst performing product Is the beef bourguinon, only having a sales of 17.

This analysis shows that the beef bourguinon Is currently the least popular item on the menu and will not be worth keeping on the new menu anymore.

Product groups and their revenue:

Chart, bar chart

Description automatically generated

Product Groups and how much Is being sold:

**Chart, bar chart

Description automatically generated**

Our analysis shows that the most popular food groups Is Pizza and Crepes and Galettes. The least popular food groups include Fish & Seafood as well as Starters/ Appetisers. Therefore, on the new menu, the company should focus on having Items from these two most popular groups and consider removing Items that belong to the unpopular groups as mentioned.

Monthly outlook of 2019 on Outlet Sales (League Table and graph):

A picture containing table

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

This graph shows that the outlet that often outperforms other branches Is Middlesborough and the worst performing Is Cardiff on a monthly basis.

Weekly outlook of the last 10 weeks recorded on Outlet Sales (League table and Graph):

Table

Description automatically generated

A picture containing application

Description automatically generated

The data shows a jump In sales for all outlets In week 22 of the year, more than the general average that can be found In this 10 week scope. On a weekly basis, there Is no one particular branch that consistently outperforms others based on our analysis.

**Task 2**

Identify and create measures in Power BI for non-additive and semi-additive facts, so they can be used in visualisations and reports.

**Non-additive facts:**

The ‘Distinct products sold per month’ column represents the proportion of distinct products that were sold per month out of a total of 78.

Table

Description automatically generated

Marking Scheme

The skills evaluated within this project are described within the SFIA 8 framework; please see https://sfia-online.org/en/framework for further information.

The skills which this project will evaluate are the following:

* Data Engineering
  + Evaluating, designing, and implementing on-premises, cloud-based, and hybrid data engineering solutions.
  + Structuring and storing data for uses including — but not limited to — analytics, machine learning, data mining, and sharing with applications and organisations.
  + Integrating, consolidating, and cleansing data.
  + Building in security, compliance, scalability, efficiency, reliability, fidelity, flexibility, and portability.
* Data Visualisation
  + Presenting findings and data insights in creative ways to facilitate the understanding of data across a range of technical and non-technical audiences.

Submissions

Completed project report should be emailed to: [AcademyQAC@qa.com](mailto:AcademyQAC@qa.com).

The email should be titled “DFEDATA2 Final Project”.