## A close-up of a cup of coffee Description automatically generated

**COFFEE SCIENCE**

**GROUP 14**

**Marta Manevska 20220056**

**Efstathia Styliagkatzi 20220078**

**Félix Funes 20220306**

[1](#_Toc154677738)

[1. Presentation of the Business & the Problem Scenario 4](#_Toc154677739)

[1.b. Problem Scenario 4](#_Toc154677740)

[1.c. Business Needs 4](#_Toc154677741)

[2. Original Data Sources 6](#_Toc154677742)

[3. Staging Area 7](#_Toc154677743)

[4. Data Warehouse 8](#_Toc154677744)

[4.a. Fact Tables of the Data Warehouse 9](#_Toc154677745)

[4.c. Hierarchies 12](#_Toc154677746)

[4.d. Star Schema Model 13](#_Toc154677747)

[4.e. Adequacy for the Business Problem 14](#_Toc154677748)

[5. ETL Processes 15](#_Toc154677749)

[5.a. ETL Process in Coffee Science Data Integration 15](#_Toc154677750)

[5.b. Transforming and Loading the Dimension Tables 16](#_Toc154677751)

[5.b.1. dim\_customer 16](#_Toc154677752)

[5.b.2. dim\_staff 16](#_Toc154677753)

[5.b.3. dim\_date 16](#_Toc154677754)

[5.b.4. dim\_sales\_outlet 16](#_Toc154677755)

[5.b.5. dim\_product 17](#_Toc154677756)

[5.c. Transforming and Loading the Fact Tables 17](#_Toc154677757)

[5.c.1. fact\_sales\_receipts 17](#_Toc154677758)

[5.c.2. fact\_pastry\_inventory 18](#_Toc154677759)

[5.c.3. fact\_sales\_target 18](#_Toc154677760)

[5.d. Validation in Coffee Science Staging Area 19](#_Toc154677761)

[5.e. Challenges Detected Post-Validation and Resolution Strategies 21](#_Toc154677762)

[5.f. Loading the Data Warehouse 23](#_Toc154677763)

[6. Extra ETL Work 24](#_Toc154677764)

[6.a. Extra Development of the long\_quality\_check Table 24](#_Toc154677765)

[6.b. Development of Conditional Load for the Data Warehouse 25](#_Toc154677766)

[6.c. Master Pipeline 28](#_Toc154677767)

[7. Summary of Lessons Learned from the Coffee Science Data Integration Project 29](#_Toc154677768)

[7.a. Main Difficulties 29](#_Toc154677769)

[7.b. Positive Lessons Learned 30](#_Toc154677770)

[8. Conclusions 31](#_Toc154677771)

# Presentation of the Business & the Problem Scenario

1.a. Presentation of the Organization

Coffee Science, a coffee chain founded by Amber and Sandeep, has established its main branch in Long Island City in New York, with a secondary branch strategically placed in Νew York City. The journey of Coffee Science has been one marked by dedication to delivering premium coffee and pastries to its customers.

Amber and Sandeep, the co-founders, share an unwavering passion for providing high-quality culinary delights to their customers. This commitment to excellence has not only garnered a loyal customer base but has also propelled the coffee chain toward steady growth. However, amidst their success, a critical concern has emerged. Despite the increasing sales figures, the chain has observed a significant amount of food waste, particularly associated with their delectable pastries.

Every second day, an array of pastries is baked; however, unsold pastries are discarded at the close of the 2nd business day. This realization has prompted the co-founders to reassess their operations and seek innovative solutions to minimize food waste while concurrently boosting sales. The essential step in this direction involves recognizing the need for a robust Business Intelligence (BI) solution to streamline inventory management processes across all branches.

## 1.b. Problem Scenario

Even though Coffee Science has noticed growth and success, it has some challenges to deal with. The most pressing issue revolves around the efficient management of the pastry inventory. Despite the commitment to delivering premium products, the company grapples with maintaining optimal inventory levels, leading to persistent challenges in operational efficiency.

Overstocking, a consequence of these challenges, incurs unnecessary expenditures and poses a threat to profitability. The impact is not only financial but also ecological, with the substantial food waste generated. This problem has spurred the co-founders into action, recognizing the need for a strategic intervention.

To address these challenges head-on, Coffee Science is embarking on a mission to implement a comprehensive BI solution. This solution aims to provide invaluable insights into inventory patterns, identify periods of low pastry sales, and guide strategic decisions to optimize production. The overarching goal is clear – to enhance revenue generation capabilities while concurrently reducing operational costs through intelligent inventory management.

## 1.c. Business Needs

When consulted about the needs of their business, the coffeeshop owners identified the following questions:

1. What is the sales performance of each pastry product per month/quarter/year?
2. What type of product experiences the highest and lowest wastage rates?
3. Is there a noticeable increase in wastage during any particular quarter?
4. Is wastage higher at any store location?
5. How does the waste percentage for each product relate to its unit price?
6. What is the financial impact of the products wasted at the end of each day?
7. Which store has the most sales by product category and product type across different quarters?
8. What were the top 10 product types sold in each quarter (by sales quantity) across the two sales outlets?
9. Which products have the highest profit margins and which ones the lowest?
10. Which products demonstrate the highest and lowest sales based on customer age?
11. Are there specific customer demographics that notably contribute to pastry sales?
12. Have the goals been achieved, and if not, how many sales are required per product category to meet the set targets?
13. What were the highest ten sales amounts per customer for each employee annually, classified by product?

Once the business needs were defined, the next step was to identify which of the available information it would be necessary to address them:

|  |  |
| --- | --- |
| Questions | Measure |
| What is the sales performance of each pastry product per month/quarter/year? | Quantity Sold per month/quarter/year |
| What type of product experiences the highest and lowest wastage rates? | Waste per Product |
| Is there a noticeable increase in wastage during any particular quarter? | Waste per Quarter |
| Is wastage higher at any store location? | Waste per Sales Outlet |
| How does the waste percentage for each product relate to its unit price? | Waste Percentage per Unit Price |
| What is the financial impact of the products wasted at the end of each day? | Monetary Value per Wasted Product per Transaction Date |
| Which store has the most sales by product category and product type across different quarters? | Quantity Sold per Product Category per Product Type per Sales Outlet |
| What were the top 10 product types sold in each quarter (by sales quantity) across the two sales outlets? | Quantity Sold per Store per Quarter |
| Which products have the highest profit margins and which ones the lowest? | Wholesale Price & Retail Price |
| Which products demonstrate the highest and lowest sales based on customer age? | Quantity Sold per Customer Age |
| Are there specific customer demographics that notably contribute to pastry sales? | Quantity Sold per Customer Age, Gender and Store Outlet |
| Have the goals been achieved, and if not, how many sales are required per product category to meet the set targets? | Sales Target Achievement |
| What were the highest ten sales amounts per customer for each employee annually, classified by product? | Sales Amount per Product per Employee per Year |

# 2. Original Data Sources

The company records day-to-day operations related to retail transactions with customers, presenting the information in the form of CSV files. These files are derived from the company's transactional database, a repository that documents transactions, stock levels, customer details, product information, sale locations, and performance targets. The available dataset encompasses transactional records spanning the years 2019 and 2020.

Here is a detailed account of all the tables provided and the information in them:

**Sales Receipts:** This file contains the details of all orders that were placed between 1-1-2019 and 30-11-2020. It has information regarding the date and time of when an order was placed, which client made the order in which store, and what products were bought.

**Pastry Inventory:** This table provides information on each branch, detailing the baked products, dates of production, sales quantities, and the corresponding amount of wasted food.

**Product:** This table describes the products that the company offers, including details such as product category and type. There is also monetary information for each product, like the retail and wholesale price, the profit margin, and whether the product is exempt from taxes.

**Date:** This file holds data describing the dates between 1-1-2019 and 30-11-2020, such as the month, the quarter, and the season.

**Sales Outlet:** This file contains information about the location of each store, including the city, state, address, and contact ways.

**Customer:** This table describes the customers who bought from the coffeeshop. It has information about their name, gender and age, as well as their home store and when they became a customer.

**Staff:** Contained within this file are comprehensive details regarding the staff members affiliated with Coffee Science. This information encompasses their names, respective job roles, store locations, and tenure within the company.

**Sales Target:** This file holds data about the targets that the company set for their quarterly sales. There is a general goal, as well as goals for each of the main types of products that the coffee shop carries: beverages, beans, food, and merchandise.

# 3. Staging Area

The staging area is an intermediate storage area in between the data sources (OLTP), in this case, the dataset given by Coffee Science, and the final destination, the data warehouse. In a Power BI project, the use of a staging area is crucial because it allows the internal processing of data. More specifically, it is a place where data transformation and cleaning can be done. Raw data often needs cleaning and transformation before it's ready for analysis. The staging area allows for data to be processed, cleaned, and transformed without affecting the original data source. It provides a safe environment for these operations. For instance, in this work, it was essential to standardize formats for the date dimension different date formats (YYYY-MM-DD, MM/DD/YYYY) into a uniform format (MM/DD/YYYY) and also to create or calculate new fields or metrics which be required for the analysis, such as customer or staff tenure, customer age and the dates attributes (day, month, semester, quarter, year).

In addition, Coffee Science and general organizations might have data coming from multiple sources, often in different formats or structures. The staging area acts as a common ground where data from various sources can be integrated, standardized, and prepared for analysis. This integration ensures consistency and accuracy in reporting. However, the data files of Coffee Science, which are provided for the analysis, were all in Excel Spreadsheets in CSV format. At the same time, performing complex transformations directly on the live operational database can affect its performance. Although the database for the company is not live operational, the staging area is crucial since it allows for optimization processes to be applied without impacting the performance of the source system.

Another crucial reason for the staging area's importance is the data validation and quality assurance. It enables data validation checks and quality assurance processes to be performed before loading the data into the final destination. This ensures that only high-quality, accurate data is used for analysis, reducing errors and improving decision-making. More specifically, validation serves as a cornerstone for safeguarding data quality, bolstering accurate decision-making, and upholding various operational standards and compliance regulations within a data-driven ecosystem. The implementation of four distinct validation rules, executed through SQL scripts or data flows within the staging area, underscores the critical role of meticulous data validation checks and their significance in maintaining a robust data ecosystem.

Going through the first rule, the main goal is to check the integrity of the business key. This validation rule ensures that the Business Key (BK) within Dimension tables adheres to primary key (PK) integrity constraints. Executed individually for each Dimension table, the script inspects the uniqueness of Business Keys. Rows with repeated Business Keys indicate a 'FAIL,' whereas unique Business Keys signify an 'OK' validation result. This emphasizes the need to tailor the rule for each dimension, stressing the importance of consistent adaptation.

Continuing to the second rule, its proposal is to ensure the unique combination of non-business key attributes. Focused on non-Business Key (non-BK) attributes within Dimension tables, this rule verifies the uniqueness of attribute combinations across the entire dimension. By checking for duplicate combinations of non-BK attributes, the validation identifies any repetition, signaling a failure if duplicate combinations are found.

The third rule is about the uniqueness of foreign key combinations in fact tables. This rule ensures the uniqueness of Foreign Key (FK) combinations, verifying their role as composite Primary Keys (PKs) for the Fact table. Rows with duplicated combinations of FKs indicate a 'FAIL,' while unique combinations signify an 'OK' validation result.

The last but not least rule is the validating relationships between fact and dimension tables. This rule establishes the correctness of relationships between Fact and Dimension tables, confirming that each Foreign Key (FK) in a Fact table corresponds to its Business Key (BK) in the parent Dimension table. It scrutinizes missing connections between tables by identifying rows in the Fact table without corresponding Business Keys in the Dimension table, thereby ensuring valid associations between related data entities.

To sum up, the rigorous application of these rules in the staging area enables the delivery of high-quality, error-free data for informed decision-making and seamless data integration, fostering trust and reliability in the organization's data ecosystem. Using a staging area ensures that the data utilized for reporting and analysis is accurate, consistent, and appropriately transformed for business requirements. It provides a structured and controlled environment for data preparation before presenting insights through Power BI reports and dashboards.

# 4. Data Warehouse

To effectively address Coffee Science's challenges of minimizing food waste and optimizing sales strategies, the integration of a Data Warehouse (DW) solution is essential. The complexities within the business across numerous branches, diverse product lines, customer demographics, and sales intricacies, require a centralized data repository that facilitates extracting insights from the data.

A Data Warehouse enables the combination of data from various sources, including transactional databases and sales records, into a unified platform. This consolidation enables meticulous analysis and pattern recognition, which is critical for identifying relationships across variables like product sales, wastage percentages, customer demographics, and pricing structures.

The use of a data warehouse is essential for extracting actionable insights, refining inventory management, evaluating the financial impact of waste, understanding customer behavior that influences sales, and driving informed decision-making. The data warehouse is a centralized hub that contains historically integrated data from diverse sources within the coffee chain's operations. It is structured to support reporting and analytical activities, facilitating the creation of comprehensive reports delivering insights across the company's operations.

The Data Warehouse design chosen for Coffee Science follows a 3-Fact Star Schema, featuring three primary fact tables connected to several dimension tables. This schema is selected for its efficiency in managing intricate data relationships within a diverse business setup. Utilizing Kimball's methodology, which stresses a bottom-up approach in Data Warehouse development, the emphasis is on creating distinct data marts. These data marts are customized to address exact business requirements, encompassing aspects like sales, inventory, customer and staff information, and product performance.

## 4.a. Fact Tables of the Data Warehouse

The fact tables within the Data Warehouse are essential repositories capturing measurements or metrics concerning specific organizational processes. Each row in these tables records crucial information about individual transactions. In this context, there are three primary fact tables: sales receipts, sales target outlet and pastry inventory. These tables are chosen for their direct relevance to the central issues.

The "Sales Receipts" table stores data about sales outlet, date, time, product, customer details and staff. This granularity is deliberately selected to address core challenges, offering insights into sales performance, customer behavior, and staff involvement. The "Sales Target" table contains the specific sales objectives assigned to each individual store. Meanwhile, the "Pastry Inventory" table captures information about sales outlet, date, and product specifics, focusing on pastry-related transactions. This level of detail facilitates analysis of sales patterns, wastage trends, and inventory management for pastries. The utilization of these specific fact tables with distinct granularity aids in thorough analysis of challenges faced by Coffee Science. They provide the necessary data for informed decisions and strategic improvements, forming the backbone of the Data Warehouse.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sales Receipts** | | **Sales Target** | | **Pastry Inventory** | |
| **Column Names** | **Key Data Type** | **Column Names** | **Key Data Type** | **Column Names** | **Key Data Type** |
| fk\_sales\_outlet | int | fk\_sales\_outlet | int | fk\_sales\_outlet | int |
| fk\_date | date | sales\_target\_year | int | fk\_date | int |
| fk\_product | int | sales\_target\_quarter | int | fk\_product | int |
| product\_quantity | int | beans\_goal | int | quantity\_baked\_products | int |
| product\_unit\_price | decimal | beverage\_goal | int | quantity\_sold | int |
| fk\_customer | int | food\_goal | int | food\_waste\_quantity | int |
| fk\_staff | int | merchandise\_goal | int | food\_waste\_quantity | int |
|  |  | total\_goal | int |  |  |

4.b. Dimensions of the Data Warehouse

Dimension tables are essential components surrounding the fact tables within the Data Warehouse, offering descriptive attributes that provide context to the data in the fact tables. These attributes, known as dimensions, shape and categorize the information captured in the fact tables. In this scenario, several dimension tables play essential roles in delineating various aspects of the organizational data.

Firstly, the "Product" dimension table serves as a comprehensive repository with numerous descriptive attributes. It includes details like product name, type, category, group, quantity, unit price, wholesale price, and promotion information. This dimension includes a hierarchy with a depth of four levels, offering a nuanced understanding of the product data and allowing for multifaceted categorization.

The "Date" dimension table holds significant importance, containing numerous attributes crucial for temporal analysis. These attributes, including date, month day number, weekday number, weekday name, month number, month name, quarter number, quarter name, season name, year number, and bake day, are integral in dissecting time-related data. This dimension table includes a hierarchy with a depth of five levels, enabling detailed temporal analysis and facilitating comprehensive time-based categorization. It is evident that certain attributes require creation as the raw data lacks this level of detailed information.

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | | **Product** | |
| **Column Names** | **Key Data Type** | **Column Names** | **Key Data Type** |
| sk\_date | int | sk\_product | int |
| proper\_date | int | bk\_product | int |
| full\_day | varchar | product\_name | varchar |
| monthday\_number | int | product\_type | varchar |
| weekday\_number | int | product\_category | varchar |
| weekday\_name | varchar | product\_group | varchar |
| day\_name\_short | varchar | unit\_of\_measure | varchar |
| weekday\_type | varchar | product\_unitprice | decimal |
| is\_special | varchar | product\_wholesale\_price | decimal |
| month\_number | int | product\_promotion | varchar |
| month\_name | varchar |  |  |
| quarter\_number | int |  |  |
| quarter\_name | varchar |  |  |
| quarter\_name\_short | varchar |  |  |
| semester\_number | int |  |  |
| semester\_name | varchar |  |  |
| semester\_name\_short | varchar |  |  |
| season\_name | varchar |  |  |
| year\_number | int |  |  |
| bake\_day | varchar |  |  |

Additionally, dimension tables such as "Sales Outlet," "Customer," and "Staff” provide further contextual information about specific aspects of the business operations. These tables include attributes detailing sales outlet details, customer information and staff detail, respectively. Each dimension table aligns with the organizational needs and contributes crucial contextual information for a comprehensive understanding of the business processes.

**Sales Outlet**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sales Outlet** | | **Customer** | | **Staff** | |
| **Column Names** | **Key Data Type** | **Column Names** | **Key Data Type** | **Column Names** | **Key Data Type** |
| sk\_sales\_outlet | int | sk\_customer | int | sk\_staff | int |
| bk\_sales\_outlet | int | bk\_customer | int | bk\_staff | int |
| sales\_outlet\_manager | int | home\_store | varchar | staff\_first\_name | varchar |
| sales\_outlet\_address | varchar | customer\_name | varchar | staff\_last\_name | varchar |
| sales\_outlet\_neighborhood | varchar | customer\_since | int | staff\_position | varchar |
| sales\_outlet\_city | varchar | customer\_tenure | int | staff\_start\_day | int |
| sales\_outlet\_state | varchar | customer\_age | int | staff\_tenure | int |
|  |  | customer\_gender | varchar |  |  |

In conclusion, the dimension tables, with their extensive array of attributes and well-organized hierarchies, form an indispensable component of the Data Warehouse framework. These tables add substantial depth and context to the data ingrained within the fact tables. By meticulously categorizing and detailing various facets across multiple dimensions, they serve as the foundational structure of the analytical capabilities. This framework empowers detailed insights, informed decision-making, and strategic advancements, facilitating a comprehensive understanding of the business operations and fostering continual growth and enhancement within Coffee Science.

## 4.c. Hierarchies

The hierarchical structures provide a clear organizational framework within each dimension table, delineating the relationships and hierarchical order between different attributes or categories. For instance, in the "Dates" dimension, the hierarchy progresses from the smallest unit, the "Day," to broader units such as "Week," "Month," "Quarter," "Season" and "Year," facilitating a structured analysis of temporal data. Similarly, in the "Product" dimension, the hierarchy categorizes products from specific items to broader categories and groups, allowing for detailed categorization and analysis. This hierarchical organization aids in facilitating comprehensive analysis and understanding of the data across various dimensions within the Data Warehouse.

**Dates:**

* Day ➔ Week ➔ Month ➔ Quarter ➔Season ➔Year

**Product:**

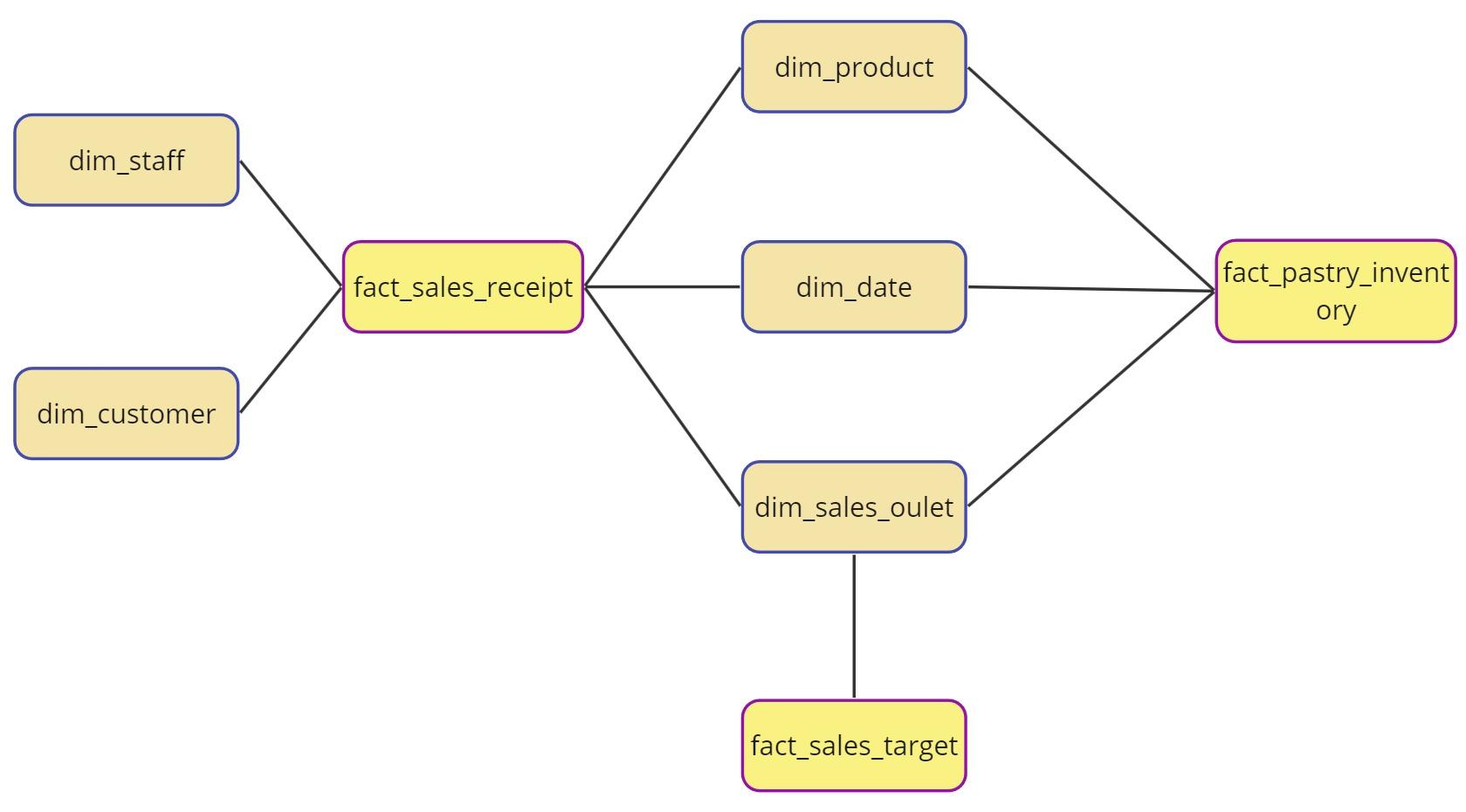
* Product Name➔ Product Type ➔ Product Category ➔ Product Group

**Sales Outlet:**

* Store Address ➔ Store Neighborhood ➔ Store City ➔ Store State

## 4.d. Star Schema Model

The schema architecture arranges the data into a star-like structure, where the fact table resides at the center, surrounded by multiple dimension tables directly linked to it. Each dimension table corresponds to a specific aspect of the business, such as product details, date attributes, sales outlets, customers, and staff information for the fact table sales receipts, sales outlet for the sales targets and date, product and sales outlets from the pastry inventory table. These dimension tables, interconnected with the fact table, collectively create a constellation-like pattern, hence the term "star" schema.



*Figure 1 Star Schema Model of Coffee Science*

The primary advantage of this schema lies in its simplicity and ease of use. By minimizing complex joins and maintaining a straightforward structure, the star schema simplifies both the design process and data querying. Its user-friendly nature enables easier navigation, making it more accessible to extract meaningful insights without encountering complexities. Moreover, this schema’s streamlined design significantly improves query performance. With fewer joins required compared to other schema models like the snowflake schema, the star schema optimizes query execution. This optimization translates into faster response times for analytical queries, enabling swift access to critical information necessary for making informed business decisions.

## 4.e. Adequacy for the Business Problem

The proposed data warehouse solution is well-suited to address the business challenges faced by Coffee Science. More specifically, the 3-Fact Star Schema design effectively captures and organizes the necessary data to analyze pastry sales performance, identify wastage patterns, and optimize inventory management. The use of Kimball's methodology ensures a structured and efficient approach to data warehouse development, aligning with Coffee Science's specific business needs.

As regards granularity, the three fact tables, "Sales Receipts", “Sales Targets” and "Pastry Inventory," capture detailed transaction-level data, allowing for in-depth analysis of sales trends, wastage patterns, and customer behavior. This granularity is crucial for identifying specific areas for improvement.

The proposed data warehouse solution directly addresses the business needs identified by Coffee Science:

* **Sales Performance Analysis:** The data warehouse provides insights into the sales performance of each pastry product, allowing Coffee Science to identify top-selling items and potential areas for improvement.
* **Wastage Pattern Identification:** The data warehouse enables Coffee Science to track wastage rates for different pastry products and identify periods of high wastage. This information can be used to optimize production and reduce food waste.
* **Inventory Optimization:** By analyzing sales trends and wastage patterns, Coffee Science can optimize inventory levels for pastries, ensuring adequate stock without incurring unnecessary costs.
* **Customer Segmentation:** The data warehouse allows for customer segmentation based on purchase patterns, enabling Coffee Science to tailor marketing campaigns and product offerings to specific customer segments.
* **Financial Performance Evaluation:** The data warehouse provides insights into the financial impact of pastry sales and wastage, allowing Coffee Science to evaluate the profitability of different products and make informed pricing decisions.

Overall, the proposed data warehouse solution is an adequate and effective tool for addressing Coffee Science's business challenges and enabling data-driven decision-making to improve sales performance, reduce wastage, and enhance profitability.

# 5. ETL Processes

The Extract, Transform, Load (ETL) process stands as a pivotal stage, as described in Section 3, enabling the transition of data from diverse sources into a unified form suitable for analysis. The ETL pipeline, named "PL\_COFFEE\_SCIENCE\_LOAD\_STG," is a conduit for moving data into the staging area, a crucial intermediary space for processing and refining raw datasets before their utilization in the data warehouse. The data will be loaded into the dimension and fact tables in the staging area warehouse that is called "STG\_COFFEE\_SCIENCE".

## 5.a. ETL Process in Coffee Science Data Integration

To initiate this process, data is loaded into the staging area via dataflows. However, a preliminary SQL activity is incorporated within the pipeline to truncate existing data in the staging area upon every pipeline execution. This measure ensures the consistency and integrity of the staged data and is particularly significant due to the iterative nature of pipeline execution aimed at achieving optimal performance and data readiness for the data warehouse.

Subsequently, a wait activity is introduced to conditionally proceed with the subsequent loading steps contingent upon the successful execution of the SQL activity. The loading sequence commences with the dimension tables, followed by the fact tables, prioritizing this sequence to ensure the logical flow of data.

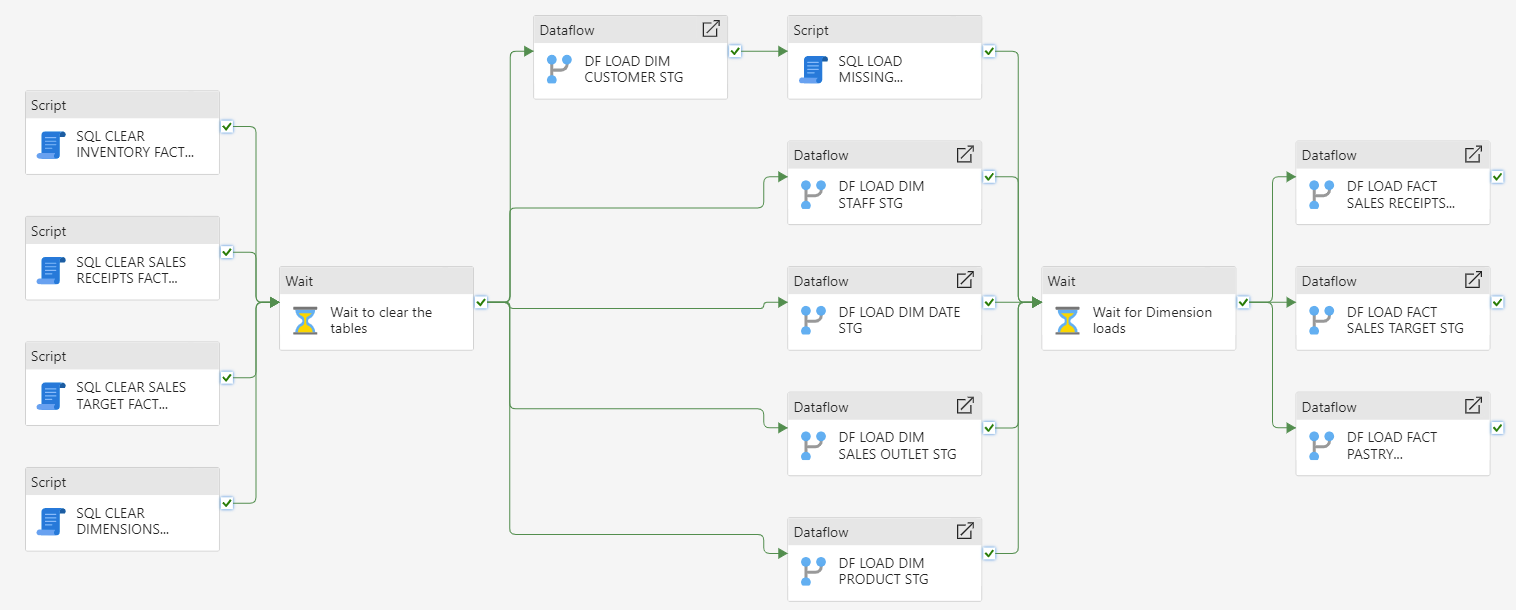


Figure 2 loading Staging Area

## 5.b. Transforming and Loading the Dimension Tables

### 5.b.1. dim\_customer

The first dimension table processed is the **Customer Dimension**. Data is sourced from the lakehouse, where company files are located. Upon initial inspection, the columns are recognized as raw text, necessitating a step to set the first row as headers for column names. Manual intervention follows for attribute typing, aligning the data with the intended attribute types, and setting the locale as English (United States) to standardize data representation. Further refinement involves splitting the 'customer\_first-name' column into distinct 'customer\_first\_name' and 'customer\_last\_name' columns for enhanced organization. Identification of missing surnames led to their replacement of null rows with ‘Unknown last name’. Additional columns for 'customer\_age' and 'customer\_tenure' are introduced, calculated to fulfill specific business requirements. Ultimately, the processed data is directed to the destination, i.e., the staging area, ensuring the preparedness of the data for subsequent processing.

### 5.b.2. dim\_staff

Similar procedures are replicated for the **Staff Dimension**, including data acquisition, attribute type adjustments, and the creation of a new attribute, 'Staff Tenure.' This was done using the ‘Age’ feature, which calculates the years passed since the specified date. Since the number given had decimals, it was rounded down to represent only years that were fully completed. Finally, this process culminated by mapping the columns in this table to those in the staging area table.

### 5.b.3. dim\_date

For the **Date Dimension**, challenges encountered in the existing file prompt the creation of new date columns, meticulously designed to align with the SQL staging table structure. Since the company had already provided a table with some information about the dates, the team proceeded to extract the day of the month and day of the week, as well as the names of the days and the abbreviations of the days and months. The corresponding semester for each was also obtained, together with their full and abbreviated forms. Finally, using the information on the day column, it was specified whether each day is a weekday or part of the weekend, and whether it is a holiday.

### 5.b.4. dim\_sales\_outlet

The **Sales Outlet** table undergoes manual attribute type corrections, with a filter applied to exclude irrelevant data, ensuring relevance in subsequent analyses. The prepared data is loaded into the staging area, aligning with the 'Sales Outlet' table structure.

### 5.b.5. dim\_product

Similar to prior procedures, the **Product Dimension** is structured according to the SQL staging table definition. Attribute types are adjusted manually, and the staged data is directed to its designated destination in the staging area.

## 5.c. Transforming and Loading the Fact Tables

Upon completion of the dimension loading, a wait activity is initiated to ensure the completion of all dimension loading processes before progressing to the subsequent dataflows. Following the completion of the wait activity, the ETL pipeline progresses by incorporating three additional dataflows dedicated to handling the fact tables. This stage mirrors the approach adopted for the dimension tables, encapsulating a meticulous series of procedures to transform, validate, and load data into the staging area.

Each dataflow dedicated to fact tables begins by sourcing data from the lakehouse, ensuring comprehensive acquisition of requisite datasets. Similar to the procedures applied in handling dimension tables, meticulous attention is given to accurately defining attribute types, validating data consistency, and ensuring alignment with the expected data types.

### 5.c.1. fact\_sales\_receipts

In the treatment of the **Sales Receipt** fact table, specific attributes, such as time-related elements deemed irrelevant for the project's analytical scope and non-measurable attributes, are excluded from the loading process. This discernment aligns with project objectives, ensuring the inclusion of only pertinent, measurable attributes relevant to the project's analytical objectives.

There were also over 30.000 empty rows in this fact table that were deleted and one important issue with the dates that was rectified. There was a mix of formats in the date column. Dates up to the 12 of each month were in the American format (MM/DD/YYYY), while dates from the 13 onward were in the European format (DD/MM/YYYY). To solve this issue, this column was split into three. Then, conditional columns were used to fix the mix between months and days.

For example, to obtain the months, the rows that had a month value equal to or higher than 13 were identified. When this was the case, this value was replaced with the value in the day column.

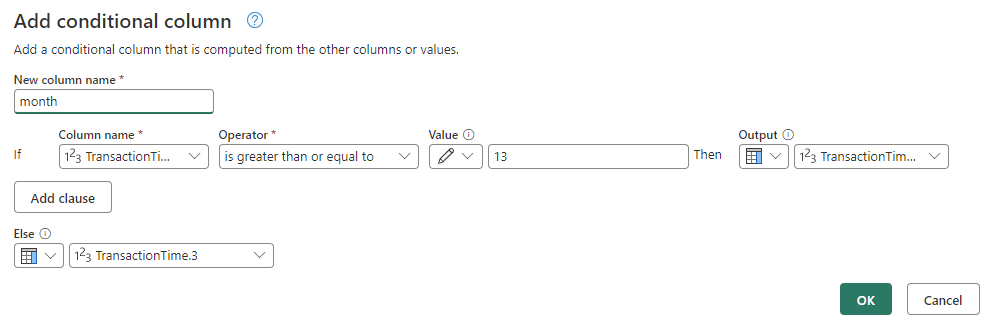


Figure 3 replacing the values

Similarly, to obtain the correct days, the values that were identical to those in the month column were selected. In this case, the right value was obtained from the original column.

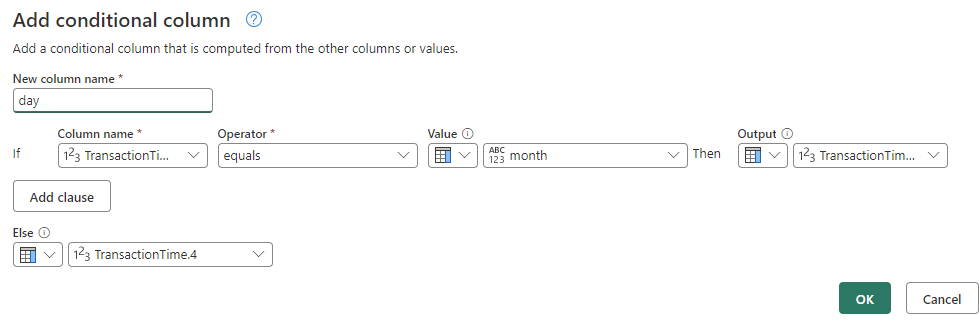


Figure 4 replacing the values

Finally, all the columns containing days, months and years were merged again in the American format, since Coffee Science is a business based in the US.

### 5.c.2. fact\_pastry\_inventory

In the case of the **Pastry Inventory** fact table, there was also a problem with the dates in the ‘BakeDate’ column. They were in the format ‘DD-MM-YYYY,’ which PowerQuery recognized as strings, since this format is not supported as date. To solve this issue, the column was split using ‘-’ as a delimiter. Then, the new columns were joined into a new custom column following the right format and using ‘/’ so the platform could recognize the date type.

### 5.c.3. fact\_sales\_target

Finally, for the **Sales Target** fact table a mismatch in information was rectified. In the data sent by the company, the branches of the coffee shop were referenced using abbreviations such as “LI” for “Long Island” and “NYC” for New York City instead of the corresponding business keys. This is why these values were replaced with the correct BKs. Also, the quarters were referenced using the abbreviated form “Q1” instead of “1”, which generated a mismatch in the expected data type. To fix this, the “Q” was removed, maintaining only the corresponding number.

Upon refining and validating the data types, the staged data is seamlessly directed to its designated destination within the staging area, ensuring the conformity of fact table data with the project's analytical requirements and aligning it for subsequent integration and analysis within the reporting environment.

## 5.d. Validation in Coffee Science Staging Area

In the context of the Coffee Science project, the staging area plays a pivotal role not only in housing data but also in ensuring its integrity through meticulous validation processes. To facilitate this, a distinct table named "log\_quality\_checks" was created within the staging area, dedicated to storing validation results and ensuring data fidelity. A specialized pipeline, "PL\_COFFEE\_SCIENCE\_VALIDATE\_STG," was also developed to execute SQL scripts and data flows intended for validation purposes.

As referred to in Chapter 3, validation is a cornerstone for guaranteeing data reliability, facilitating informed decision-making, meeting regulatory requirements, streamlining operational efficiency, and preventing data corruption. The formulation of validation rules involved the creation of specific SQL scripts or dataflows, organized into distinct rules for meticulous checks performed on the tables within the staging area. These rules spanned across four crucial areas:

Check Integrity of Business Key: Rule one, known as "Check Integrity of Business Key", ensures the conformity of Business Keys (BK) within Dimension tables to primary key (PK) constraints. This verification entails individualized checks for uniqueness of Business Keys, employing SQL for the dimension tables date, customer, staff, and sales outlet, and dataflow for the product dimension. For the SQL approach, the primary query utilizes a subquery to count occurrences of repeated Business Key values within the specified Dimension table (such as 'stg\_dim\_customer'). Grouping these rows based on the Business Key (bk\_customer) using the GROUP BY clause, the HAVING clause then filters groups where the count of rows with the same Business Key is not greater than 1, ensuring their uniqueness. The validation outcome is marked as 'FAIL' if there are rows with repeated Business Keys (count(row\_count) > 0); otherwise, it's marked as 'OK.' The implementation of this rule emphasizes the importance of adapting it for each dimension, as underscored by the necessity to "always adjust the rule, each time!". The process concludes with the insertion of pertinent results into the 'log\_quality\_checks' table within the staging area, providing a timestamped record of the validation execution for reference and auditing purposes. The same validation concept is employed within the dataflow for the product dimension, utilizing filtering mechanisms to retain duplicates within the 'bk\_product’ and consequently updating the 'log\_quality\_checks' table in the staging area.

Unique Combination of Non-Business Key Attributes: The second validation rule aims at ensuring the uniqueness of attribute combinations across Dimension tables that are not Business Keys (non-BK). The procedure utilizes subqueries to examine duplicated attribute combinations and identify instances where uniqueness is violated. For instance, in the case of the 'stg\_dim\_shipper' Dimension table, the main query employs subqueries to count rows containing duplicate combinations of non-BK attributes such as 'shipper\_name,' 'shipper\_coverage,' and 'shipper\_low\_temp.' Employing the GROUP BY clause to group rows by these non-BK attributes and subsequently using the HAVING clause to filter out instances where the count of rows with identical combinations exceeds 1, ensures the uniqueness of each combination. The validation outcome is 'FAIL' if any duplicated rows are identified, indicating non-uniqueness; otherwise, it's marked as 'OK.' Following this validation process, relevant outcomes are inserted into the 'log\_quality\_checks' table in the staging area, recording the validation execution's timestamp for comprehensive reference and audit tracking. This validation methodology is replicated within the dataflow for the product dimension, leveraging filtering techniques to retain duplicates across all columns except 'bk\_product,' thereby updating the 'log\_quality\_checks' table in the staging area with the validation results.

Uniqueness of Foreign Key Combinations in Fact Tables: The third validation rule addresses the confirmation of Foreign Key (FK) combinations' uniqueness within Fact tables, ensuring their suitability as composite Primary Keys (PKs). This verification process involves the identification of duplicate combinations of FKs using subqueries. The primary query employs a subquery to count rows containing duplicated combinations of FKs within the specified "stg\_fact\_sales\_receipts" and "stg\_fact\_sales\_outlet" tables. Rows are grouped based on the Foreign Keys present in the table, and the HAVING clause filters out groups where the count of rows with identical combinations of FKs exceeds 1, indicating non-uniqueness. By identifying rows with duplicated combinations of FKs (count(row\_count) > 0), the validation outcome is marked as 'FAIL'; otherwise, it is marked as 'OK.' These validation outcomes are systematically recorded by inserting relevant results into the 'log\_quality\_checks' table within the staging area. This validation methodology is mirrored in the dataflow process for the 'pastry inventory' Fact table, utilizing filtering mechanisms across columns containing FKs like 'fk\_date,' 'fk\_sales\_outlet,' and 'fk\_product.' Consequently, the 'log\_quality\_checks' table in the staging area is updated with the validation results specific to this fact table.

Validating Relationships between Fact and Dimension Tables: The fourth validation rule concentrates on ensuring accurate associations between Fact and Dimension tables by validating the Foreign Key (FK) linkage in each Fact table to its corresponding Business Key (BK) in the Dimension table. The rule's SQL script executes a check to verify if each Fact table's FK can be linked back to its respective BK in the parent Dimension table ('stg\_dim\_sales\_outlet' in this instance). By utilizing a LEFT JOIN operation between the 'stg\_fact\_sales\_receipts' Fact table and the 'stg\_dim\_sales\_outlet' Dimension table, the script identifies rows in the Fact table where the FK does not have a corresponding BK in the Dimension table. The validation outcome, indicating the number of rows lacking a parent key, is captured and stored in the 'log\_quality\_checks' table within the staging area. Similar checks are performed for other Fact tables, requiring separate script activities for each FK, ensuring the existence of the parent key across different Fact tables. Moreover, for the 'pastry inventory' Fact table, due to its three FKs, three distinct dataflows are executed to validate the FK linkage. These dataflows merge queries between the Dimension table's respective BK and the Fact table's FK, verifying the existence of the parent key. Subsequently, the 'log\_quality\_checks' table in the staging area is updated with the specific validation outcomes for this Fact table.

## 5.e. Challenges Detected Post-Validation and Resolution Strategies

Following the validation process conducted within the Coffee Science data, several anomalies were uncovered, prompting immediate attention and resolution. Predominantly, these discrepancies centered on inconsistencies within customer information embedded in the fact tables, presenting challenges to data coherence and integrity.

Notably, failures occurred in rules 3 and 4, concerning the fact tables of sales targets and sales receipts, indicating duplicate combinations between their Foreign Keys (FKs) and corresponding dimensions. Specifically, there were 2 instances of repeated Primary Keys (PKs) in the sales target table and 457 in the sales receipts table. Moreover, rule 4's failure in the pastry inventory table revealed 5 rows without a parent key in the date dimension, signifying dates absent in the date dimension but referenced in the pastry dimension. However, the most critical challenge emerged from the rule 4 failure in the sales receipts fact table, exposing 14,704 rows without a parent key associated with the customer dimension. This indicated a substantial deficit in customer data within the customer table.

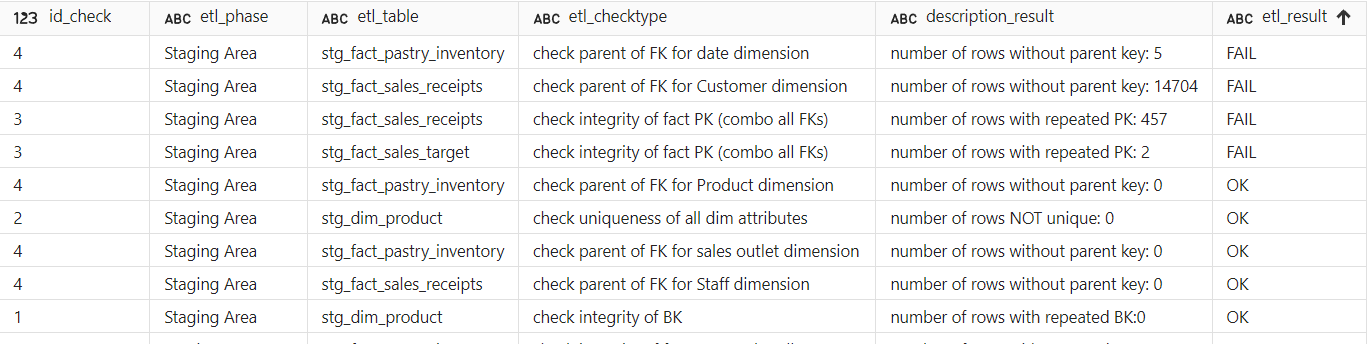


Figure 5 results of the validation quality check

Addressing these issues necessitated revisiting the staging area pipeline for resolution. To rectify rule 3's discrepancy, the solution involved amending the validation scope. Previously, the code only scrutinized FK columns for duplicates, neglecting other crucial metrics. Enhancing the rule's execution, adjustments were made to incorporate checks across all non-BK columns, ensuring comprehensive validation, such as considering sales goals and quarters in addition to sales outlets.

Regarding rule 4's challenge in the pastry inventory, the solution involved eliminating redundant data entries that referenced December 2020, a period not covered in the fact tables. Given the incongruity and limited scope of these entries, their removal from the date dimension was deemed a viable solution. This was done using a filter in the existing data flow that excluded dates after November 30, 2020.

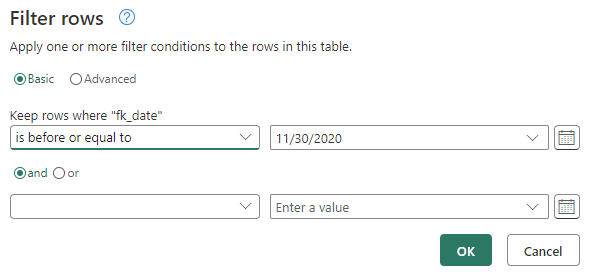


Figure 6 Eliminating redundant data

Resolving the critical issue stemming from rule 4 between sales receipts and the customer dimension involved adding a placeholder row in the customer dimension for a non-existent ID (0). As this ID was referenced in the fact table but absent in the customer dimension, the addition of this row acted as an identifier for a customer for whom there was no information available. This was executed using an SQL script activity, bridging the gap between non-existent customer IDs in the fact table and the customer dimension.

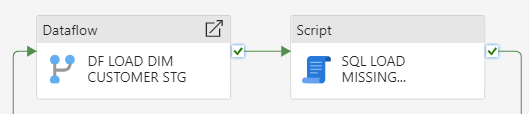


Figure 7 finding the missing customer

Another discrepancy involved a customer ID (5000) in the sales receipts table without a corresponding customer in the customer table. Following communication with the company regarding this missing client information, the necessary details are acquired. Subsequently, to address this disparity, a new customer entry was created using SQL, incorporating the information obtained to fill the void in the customer database. This comprehensive approach to error resolution ensures data consistency and rectification across the Coffee Science data environment.

Upon scrutiny, an error was detected within the SQL script associated with Rule 3 concerning the fact table 'sales receipt.' Rectifying this script led to the inadvertent generation of duplicate entries upon subsequent execution. To address this issue, the SQL activity was replaced by a dataflow process. In the dataflow, procedures were implemented to eliminate duplicates, ensuring a count of zero duplicates. Subsequently, new columns were introduced in the log quality check to accommodate this refined data verification process.

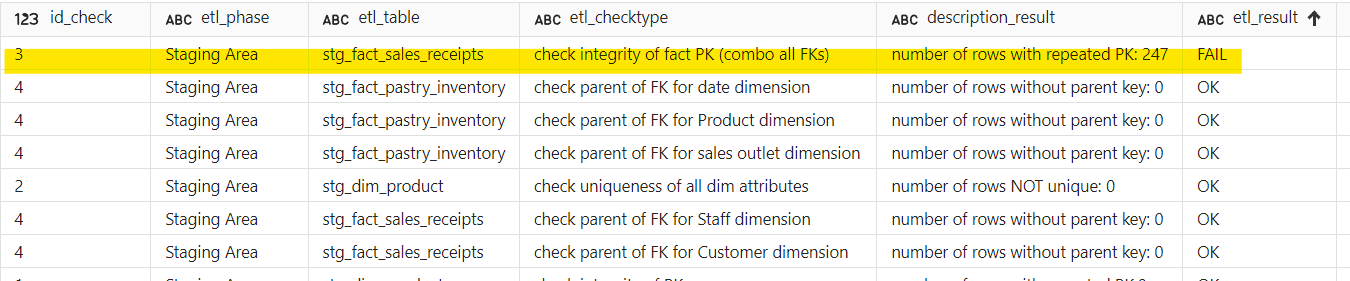


Figure 8 Issue with rule 3

## 5.f. Loading the Data Warehouse

Once all the data was transformed and loaded to the staging area, and the problems identified in the validation stage were fixed, the team proceeded to transfer all the information to its final destination: the data warehouse. For this, a specified pipeline was created, called ‘PL\_COFFEE\_SCIENCE\_LOAD\_DW’.

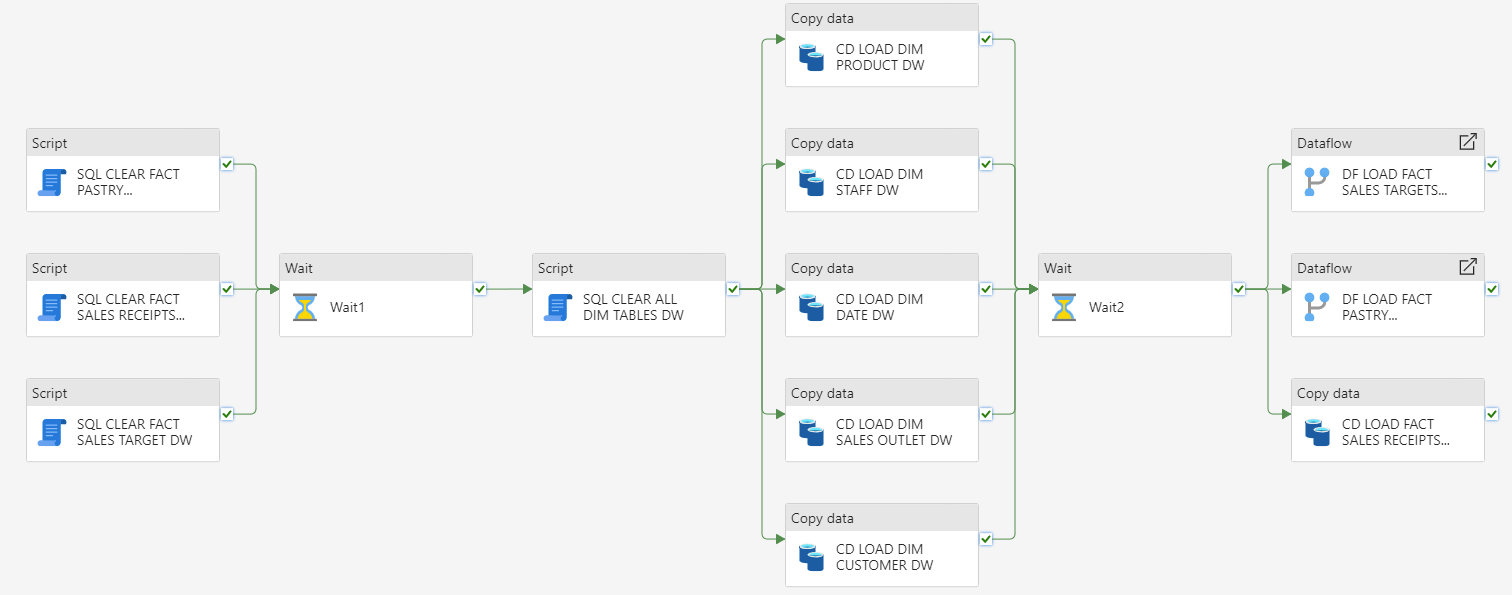


Figure 9 Loading into the data warehouse

Though this process looks very similar to the one carried out for loading the staging area, there are some key differences. The main one is that now all the data has already been cleaned and fix, which means that the only remaining transformation is adding a surrogate key to all the dimension tables. This was done using a SQL script within the copy data activities. In this way, it can be ensured that any inconsistencies in the business keys will not affect the relationship between tables.

Another important transformation is replacing the business keys with the new surrogate keys in the fact tables. This is done using data flows and copy data activities. In the case of the former, all the dimensions related to the fact table are loaded so that the tables can be joined, and the new surrogate keys are extracted.

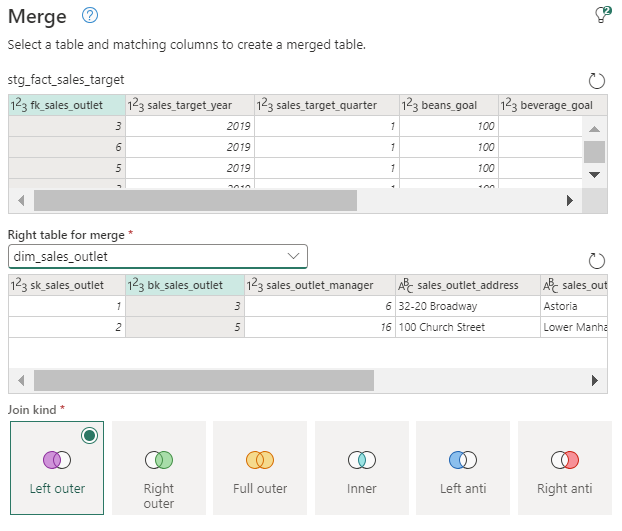


Figure 10 joining tables

In the case of the fact table using a copy data activity, the tables were joined using SQL and the surrogate keys were obtained by including them in the select statement. Finally, in all cases, the surrogate keys were selected as the new foreign keys in the mapping stage.

# 6. Extra ETL Work

In order to fully refine the ETL process for Coffee Science, and to meet some needs that arose as the project was being developed, the following features were added to the project. In this way, the final result to be handed to the company is more polished as it provides a better user experience.

## 6.a. Extra Development of the long\_quality\_check Table

The introduction of two new columns, namely for date and time, within the 'log quality check' table, stemmed from the necessity to precisely document the execution timeline of validation rules. This addition was prompted by the imperative need to maintain a comprehensive log that accounts for the exact moment when these quality checks were performed. To achieve this, adjustments were made in the validation pipeline, necessitating the integration of two distinct columns dedicated to capturing the date and time of execution. Firstly, two new columns were added in the dataflows of the validation pipeline and secondly, the SQL script was modified to incorporate commands such as "CONVERT(DATE, GETDATE()) AS execution\_date" for recording the present date and "CONVERT(TIME, GETDATE()) AS execution\_time" to denote the current time.

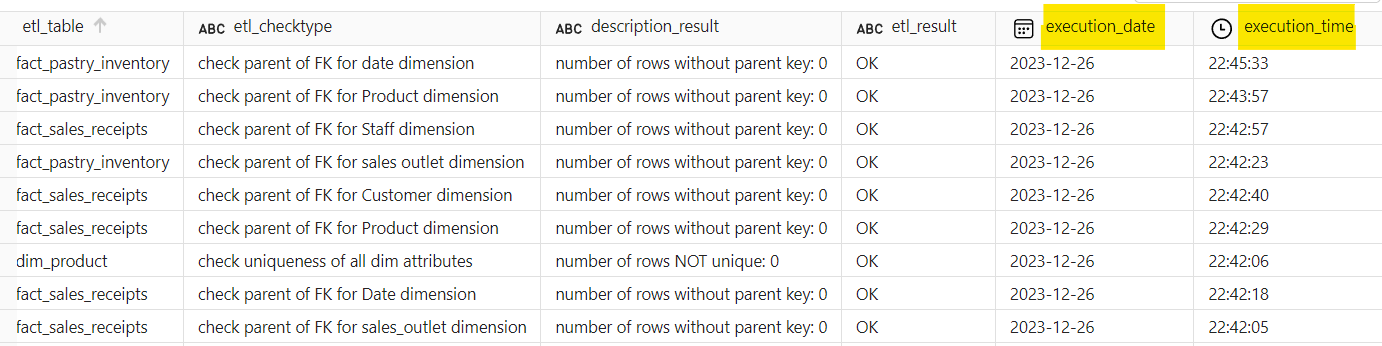


Figure 1 creating two distinct columns dedicated to capturing the date and time of execution

## 6.b. Development of Conditional Load for the Data Warehouse

In an effort to test the integrity and effectiveness of the validation rules, deliberate duplication of data was introduced into the validation pipeline. The duplication process was applied across all dimensions to assess the functionality of rules 1 and 2. Additionally, the fact tables underwent augmentation with new rows utilizing SQL to introduce fresh foreign keys that deliberately lacked corresponding matches in the dimension tables. This augmentation also entailed the creation of duplicates within the foreign keys.

Upon executing the pipeline, observations revealed that Rule 2, specifically applied to the date dimension, and Rule 4, targeted at scrutinizing the foreign key connections in the pastry inventory fact table against the product dimension, did not register as failed validations. Subsequently, upon revisiting the pipeline, meticulous scrutiny was undertaken to pinpoint the root causes of the discrepancies.

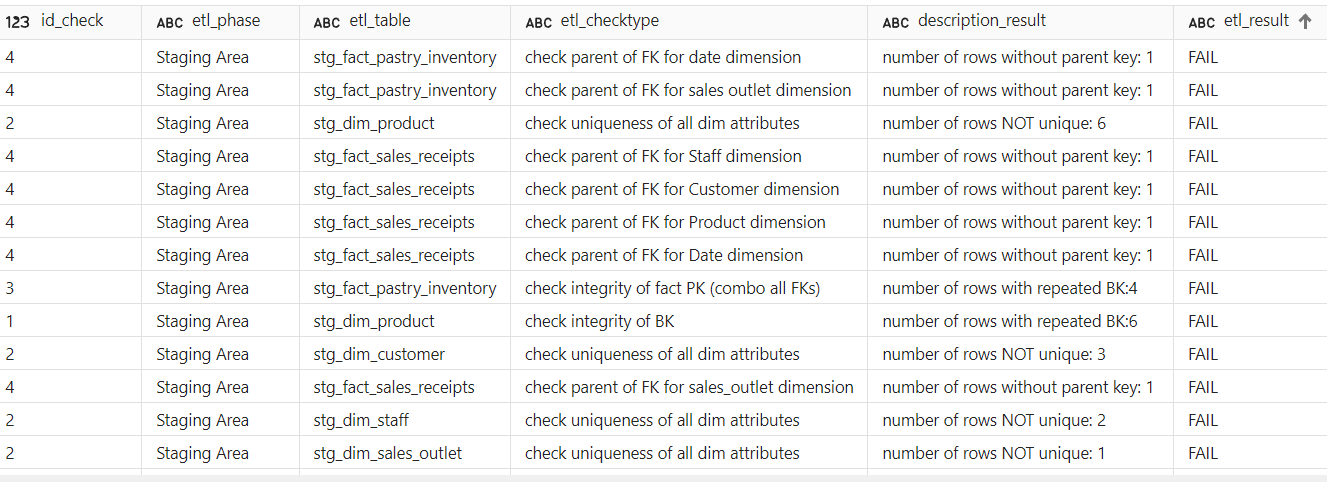


Figure 1 checking if the validation made is correct

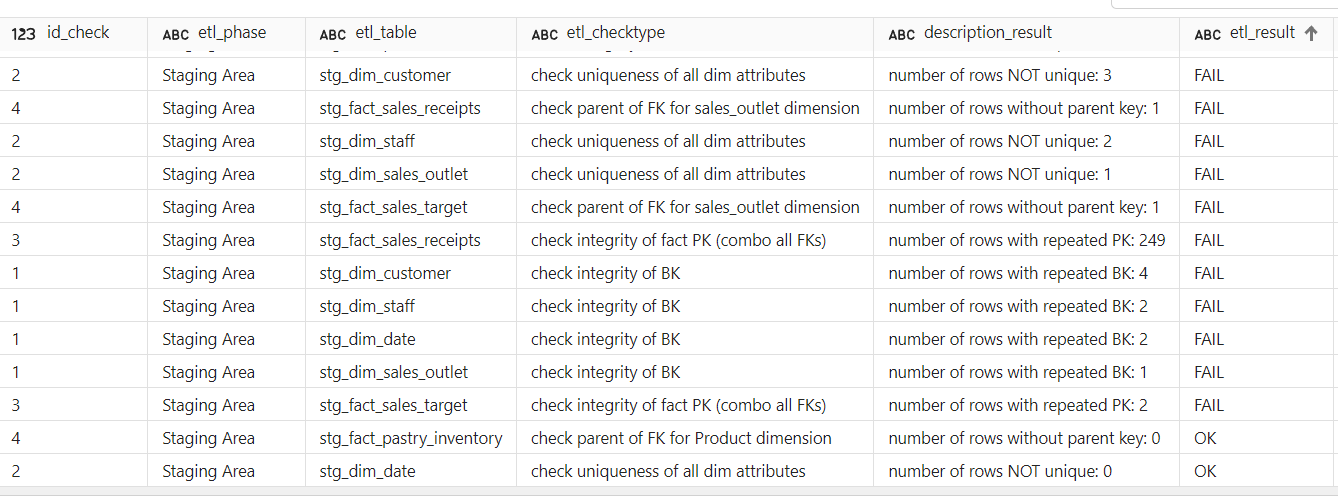


Figure 1 checking if the validation rules are correct.

Regarding the date dimension, it was identified that an oversight occurred during the SQL query generation for the duplicates. Rectification involved an update to the duplication process for new rows. On the other hand, for the fact tables, an anomaly was uncovered within the dataflow, as the system was unable to interpret the 'NULL' value, requiring a conversion to a 'null' value. Post addressing these issues and rerunning the pipeline, the comprehensive set of results indicated the validation rules were now generating the intended failures. The rectifications and subsequent execution provided assurance that the quality checks were operating effectively, thereby ensuring the robustness of the validation framework.

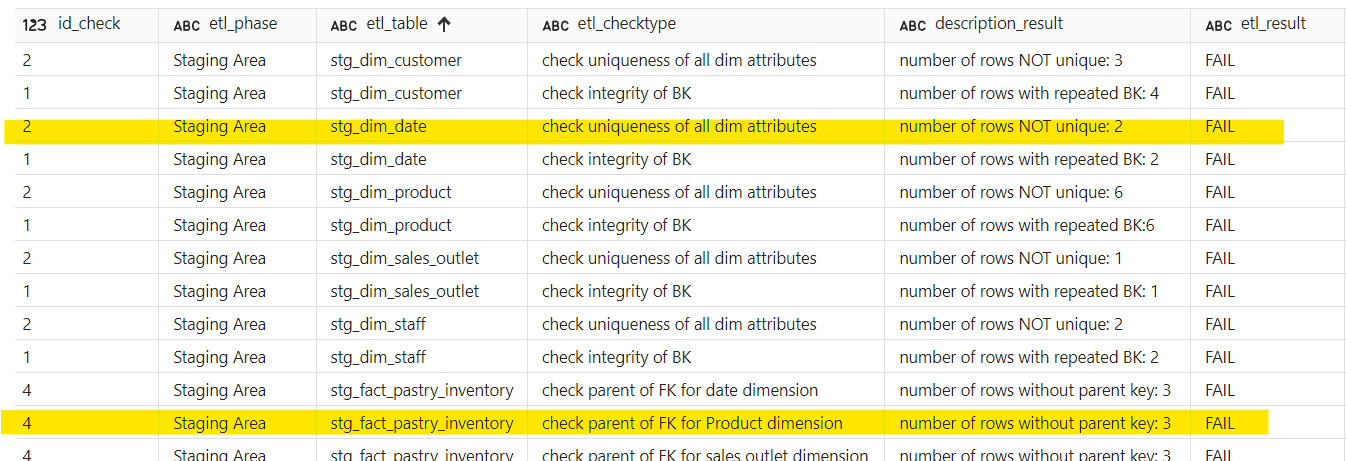


Figure 14 incorrect validation for pastry and date

Subsequently, an initiative was conceived to implement a new workflow feature that correlates the status of the validation process with the content of the 'log quality check' table. If the table contains even a single failure in its results, the pipeline status is marked as a failure. Conversely, if all results in the table are 'OK,' the status is designated as a success. Given the complexity of the existing validation pipeline and the fact that the limit of 40 activities per pipeline had already been reached, a new dedicated pipeline titled 'PL\_COFFEE\_SCIENCE\_DATA\_QUALITY\_VERIFICATION' was devised for this purpose. This pipeline orchestrated the integration of the validation process with an invoke activity, executing a precisely crafted SQL script as described. The following images depict the contrasting outcomes when the 'log quality table' contains at least one failure and when it doesn't.

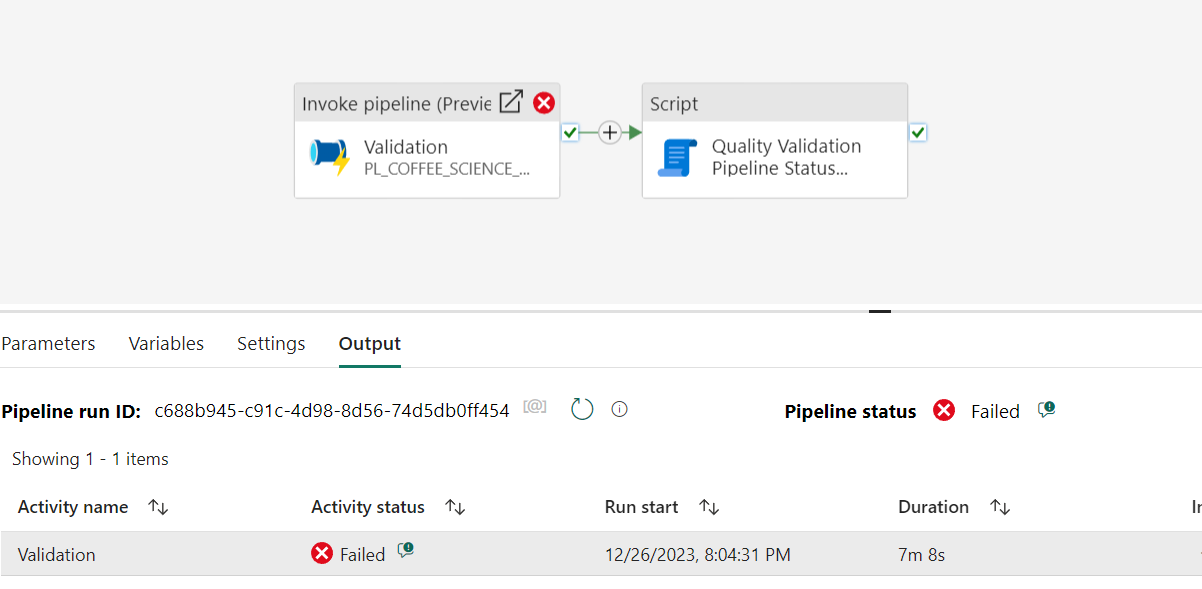


Figure 15 Invoke pipeline - there is at least one incorrect activity.

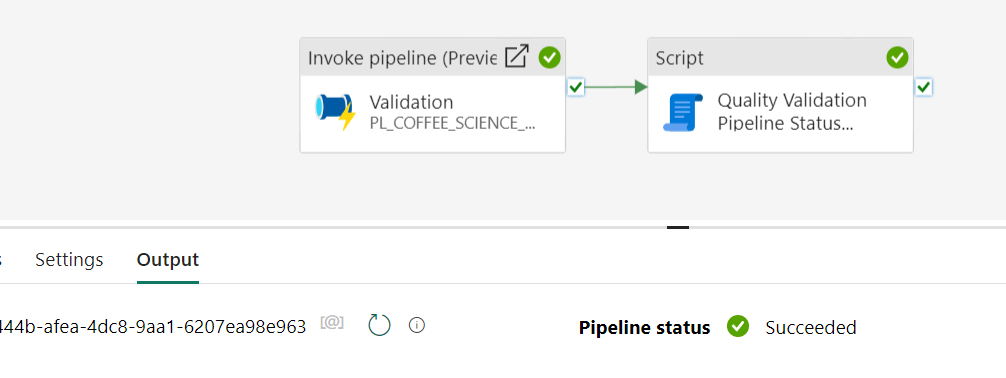


Figure 16 invoke pipeline - all activities are correct.

## 

## 6.c. Master Pipeline

To streamline the data warehouse loading process and minimize disruptions to the OLAP system during operational hours, a master pipeline called 'PL\_COFFEE\_SCIENCE\_MASTER\_ORCHESTRATION\_PIPELINE' was implemented. This automated pipeline can be scheduled to run when the OLAP system is not being used.

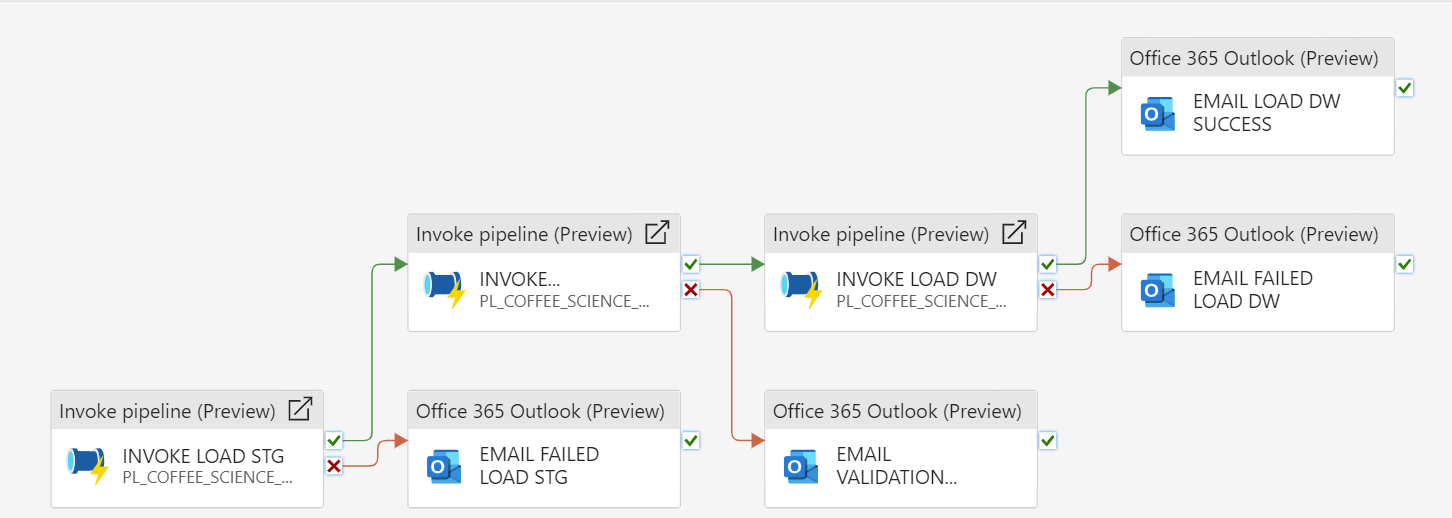


Figure 16 master pipeline

When run, the pipeline initiates data transformation, loading the data into the staging area, and seamlessly progressing to validation. If no issues arise during the validation, the data is promptly uploaded to the data warehouse. Email alerts are triggered whenever an error arises in any part of the process.

1. The pipeline starts by invoking the pipeline ‘PL\_COFFEE\_SCIENCE\_LOAD\_STG’.
2. To handle errors, an email alert is triggered if issues arise during data transformation, facilitating swift issue resolution.

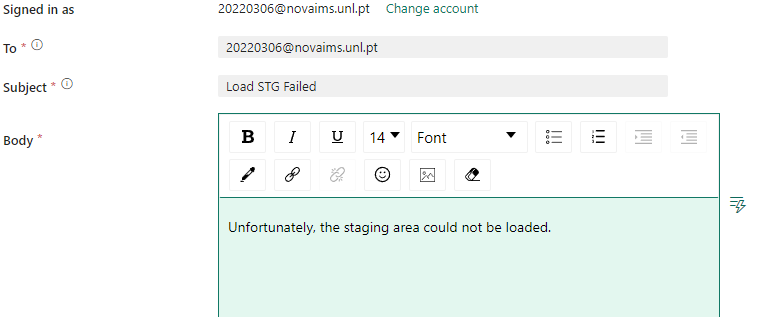


Figure 17 receiving mail if the staging area is not loaded

1. In the case that no issue comes up, the process continues by invoking ‘PL\_COFFEE\_SCIENCE\_DATA\_QUALITY\_VERIFICATION’ and the same process is repeated with intelligent error triggering email alerts. This means that if the pipeline fails or if there is any of the rules in the validation fails, an email is sent. Otherwise, the process continues.
2. Finally, once the validation succeeds, the data reaches the data warehouse through the invocation of the pipeline ‘PL\_COFFEE\_SCIENCE\_LOAD\_DW’. The process finishes by sending an automated email indicating whether final loading activity succeeded or not.

This whole process can be automated using the ‘Schedule’ function in the ‘Home’ menu.



Figure 18 Schedule Button

In essence, the master pipeline ensures a reliable, automated, and timed data integration process, aligning with Coffee Science's goal of maximizing data-driven decision-making capabilities.

# 7. Summary of Lessons Learned from the Coffee Science Data Integration Project

The Coffee Science data integration project provided the group with valuable lessons and experiences, encompassing both challenges and positive takeaways.

## 7.a. Main Difficulties

Data Quality Challenges: Dealing with inconsistencies and quality issues within the dataset posed a significant challenge. The need to rectify discrepancies in date formats, handle missing information, and address anomalies in customer data highlighted the importance of thorough data validation and cleaning processes.

Complexity of Validation Rules: Implementing and validating complex rules, especially those related to foreign key relationships and unique attribute combinations, proved challenging. The intricacies of ensuring data integrity across multiple tables required meticulous attention and adaptation of validation rules for each dimension and fact table.

Identification of Anomalies Post-Validation: Uncovering anomalies in the data post-validation, such as duplicate keys and missing relationships, added complexity to the project. Addressing these issues demanded quick problem-solving and an agile approach to ensure the integrity of the data before loading it into the Data Warehouse.

## 7.b. Positive Lessons Learned

Importance of Staging Area: The project emphasized the critical role of the staging area as an intermediate storage space for processing and refining data. It served as a safe environment for data transformation, cleaning, and validation without affecting the original data source. The staging area played a key role in ensuring data quality before loading it into the Data Warehouse.

Thorough ETL Process Understanding: Going through the Extract, Transform, Load (ETL) process provided the group with a deep understanding of the intricacies involved in moving data from diverse sources to a unified form suitable for analysis. The step-by-step explanation of transforming and loading both dimension and fact tables enhanced the group's knowledge of data integration workflows.

Flexibility and Problem-Solving Skills: The project underscored the importance of flexibility and problem-solving skills in the face of unexpected challenges. Identifying and resolving issues post-validation required adaptability and a collaborative approach to ensure the success of the data integration process.

Transparent Communication about Challenges: Transparency in communication about challenges, as evidenced by the identification and discussion of anomalies, emerged as a positive aspect. Open communication regarding difficulties fosters a collaborative environment and enables the group to collectively address issues for a more robust solution.

Data Warehouse Design Considerations: The design of the Data Warehouse using a 3-Fact Star Schema following Kimball's methodology provided insights into efficient data management in a diverse business setup. Understanding the importance of distinct data marts tailored to specific business requirements was a positive outcome.

In conclusion, the Coffee Science data integration project served as a valuable learning experience for the group. The challenges encountered underscored the importance of data quality, thorough validation processes, and adaptability. Positive lessons included the critical role of the staging area, a deep understanding of the ETL process, problem-solving skills, transparent communication, and considerations for effective Data Warehouse design.

# 8. Conclusions

In conclusion, the implementation of Coffee Science's new Business Intelligence (BI) platform marks a transformative milestone in the company's journey toward operational excellence and strategic growth. Through the rigorous data integration project, the team has not only addressed the immediate challenges of inventory management and food wastage but have also laid the groundwork for sustainable and data-driven decision-making.

The BI platform's role in optimizing inventory management, precisely controlling pastry sales performance, and minimizing unnecessary costs is poised to make a substantial impact. By leveraging the power of data analytics, Coffee Science can now implement data-driven sales strategies, segment customers effectively, and evaluate financial performance to enhance marketing tactics and pricing decisions. This holistic approach not only improves profitability but also positions the company as a more sustainable and socially responsible player in the market by curbing food wastage.

The centralized analytics of the BI platform promises to enhance operational efficiency, allowing Coffee Science to uncover and address operational bottlenecks. This, in turn, guides outlets to meet realistic sales targets, fostering adaptability and customer satisfaction. The strategic implementation of the BI platform empowers proactive growth, providing the agility needed to thrive in a dynamic market environment.

In summary, the Coffee Science data integration project, culminating in the implementation of the BI platform, goes beyond resolving immediate challenges. It unlocks strategic insights that drive growth, financial efficiency, and operational excellence. The lessons learned from this endeavor position the company to navigate future data challenges with confidence, ensuring a resilient and forward-thinking approach in the ever-evolving landscape of business intelligence and data management.