Specification Document

Sales Department Data Mart

BI33 Final Project

Date: 01.28.2024

Version 2 - Functional Specification

Alex Chagan

Table of Contents

1.1. Project Objective

This project aims to develop a data mart for the Sales Department of Decathlon, a leading sports retail company. The data mart will serve as a centralized repository for sales-related data, enabling efficient and effective data analysis and reporting for decision-making purposes.

1.2. Business Objectives

- In-depth analysis of sales data to identify trends, patterns, and opportunities for revenue growth.
- To optimize sales strategies, and provide insights into top-performing products, sales regions, and customer segments.
- Enable a comprehensive view of customer behavior, preferences, and purchasing patterns.
- Provide detailed analytics on product performance, including sales volume, revenue, and profitability.
- Identify top-performing sales representatives and provide insights into their strategies and techniques.

1.3. Project Contents

Fact Sales - captures detailed information about sales transactions, including the date, employee, customer, product, store, quantity, and revenue. It serves as the central table in the data mart.

Dim Employees - contains information about the employees in Decathlon's Sales Department. It includes employee ID, name, department, role, hire date, and location.

Dim Customers - holds customer-related information, including customer ID, name, address, and geographical details. It allows for customer segmentation and analysis to understand customer behavior.

Dim Products - provides details about the products sold by Decathlon. It includes product ID, name, category, subcategory, brand, and price attributes.

Dim Stores - contains information about the stores operated by Decathlon. It includes store ID, name, and location.

Dim Products History - functions as a repository of alterations to product-related attributes over time, preserving a chronological record of modifications such as updates, inserts, and deletes.

1.4. GANTT - Link

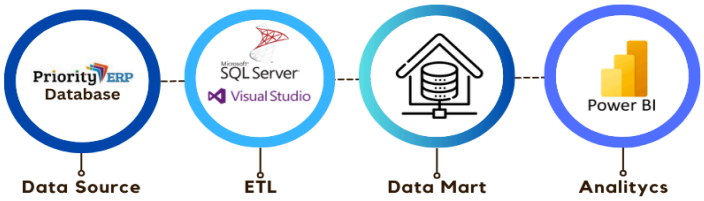
2.1. Prerequisites

System / Process	Description
SQL Server OLTP	Operational DataBase: PriorityERP
SSIS Path	C:\Users\alexc\source\repos\Decathlon\DecSalesDM\DecSalesDM.sln With all the Solution files for each DIM/Fact Table
SQL Server OLAP	Analytical Database (Data Mart): DecSalesDM
PowerBI	

2.2. Source To Target - Link

2.3. ERD - Link

2.4. HLD:



Data collection and exploration from the ERP system will be performed in SQL Server. The data will undergo an ETL process for organization and arrangement into a Data Warehouse using Visual Studio's SSIS. Finally, the measures in reports and visuals will be presented in Power BI.

2.5 Analysis Measures and Reports

The following reports are based on the last 2 years to generate the most accurate and relevant analysis.

Customer Analysis:

- Amount of active customers by Country/Top 5 Regions/Top 5 Cities.
- Number of orders by Country/Top 5 Regions/Top 5 Cities.
- Total Sales and Quantity sold by Country.
- Total Sales and Quantity sold each month for the last year.
- Number of customers who made purchases for each Store.

Employee Performance Analysis:

- Top 5 Employees by orders and total sales
- Total Sales Comparison with the previous year
- Quantity sold distribution in percentage by Product Category
- Top 5 Stores by Total Sales
- Total Sales Comparison between Physical Stores and Internet Stores

Visualization Dashboard:

A visual representation of key business data and metrics, providing a consolidated and interactive view of critical information within an organization. an introduction to the underlying reports and semantic models. It contains key KPIs and graphs that represent the most important aspects of the report.

3.1 ETL Process Overview (SSIS)

The ETL process is designed to extract relevant data from diverse sources, transform it into a standardized format, and load it into the relevant tables. This streamlined approach ensures the availability of accurate and consistent data for analytical purposes.

- Extraction: Retrieve data from the source database, including sales transactions, store details, customer information, employee records, and product data.
- Transformation: Standardize, cleanse, and enrich the extracted data to ensure consistency and accuracy.
- Loading: Populate the FACT and DIM tables with transformed data, ensuring proper relationships to the data model.

3.2 Solutions and SSIS Packages

The ETL processes in SSIS consist of several Solution projects that deploy each stage in the workflow. Each Solution project contains the relevant packages for the deployment purposes of the given stage.

Hierarchy of Solution and Packages:

DecSalesDM_MRR_ETL

- MRR_DIMS - Creation of mirror tables for dimension processes
- MRR_FACTS - Creation of mirror tables for fact purposes

DecSalesDM_Stores_ETL

- STG_Stores
- DIM_Stores

DecSalesDM_Customers_ETL

- STG_Customers
- DIM_Customers

DecSalesDM_Employees_ETL

- STG_Employees
- DIM_Employees

DecSalesDM_Products_ETL

- STG_Proudcts
- DIM_Products

3.3 General Notes

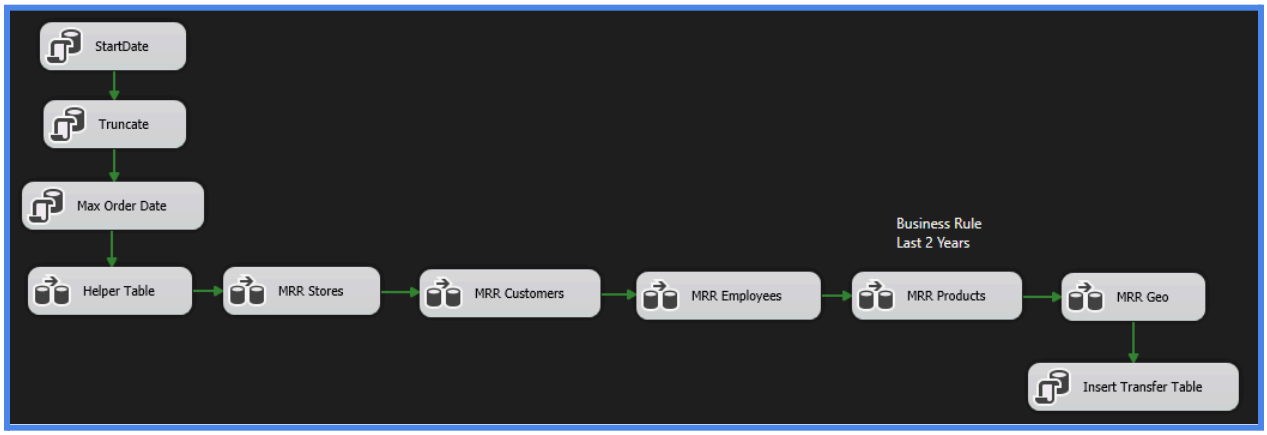
- For each package, we record the date and time of the execution's beginning and end, inserting them into the transfer table for better documentation.
- We also insert the number of rows transferred in each step into the transfer table.
- For the mirroring and staging packages, we perform a truncation on the relevant tables.

3.4. Mirror Stage

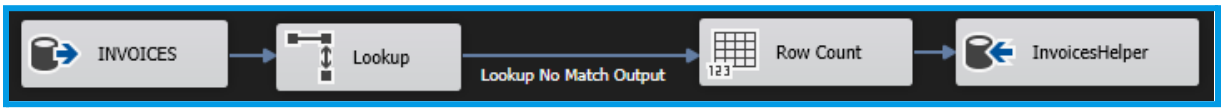
The primary objective of the mirror stage is to replicate and temporarily store raw data exactly as it exists in the source systems. This ensures the integrity of the original data.

Dimension-related Tables Mirroring:

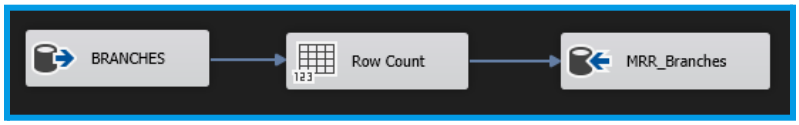
Control Flow:



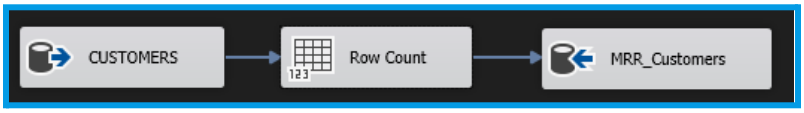
Data flows:



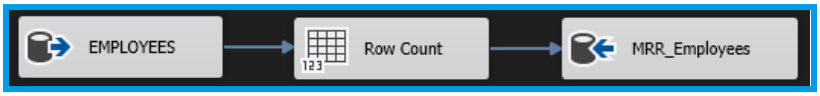
Helper table that saves data from the invoices table of the source database. We only take new data that isn't already in the helper table. In addition, we don't truncate this table because we need all the data to create and update some of the dimension staging tables.



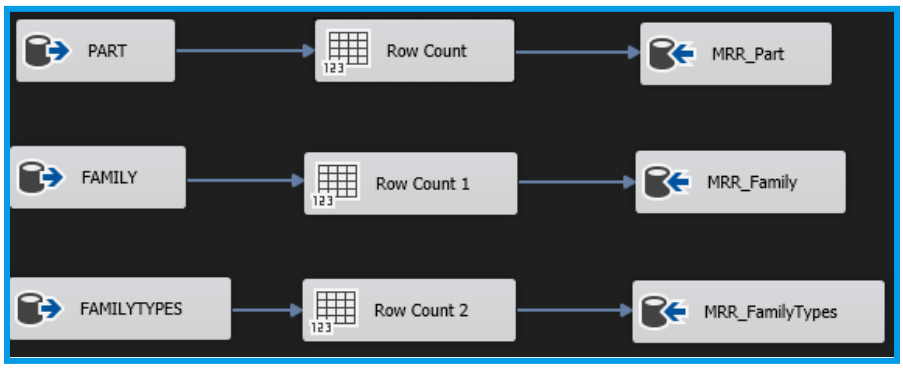
The Stores dimension-related tables from the source database



The Customers' dimension-related tables from the source database

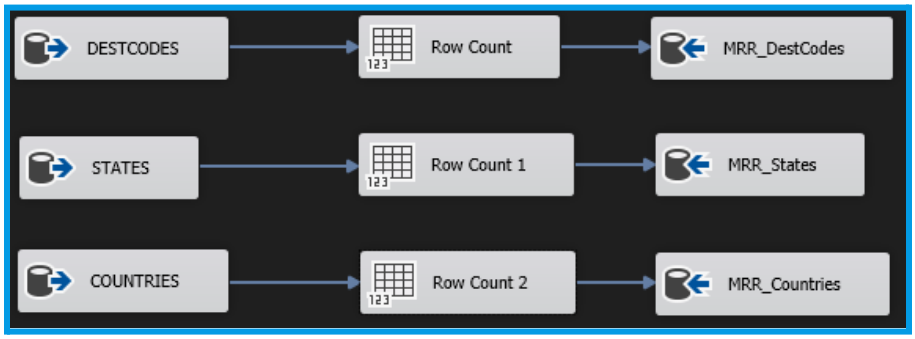


The Employees' dimension-related tables from the source database



The Products' dimension-related tables from the source database

- MRR_Part is affected by a business rule in the form of an SQL query. Only products that were sold in the last 2 years are inserted into the mirror table.

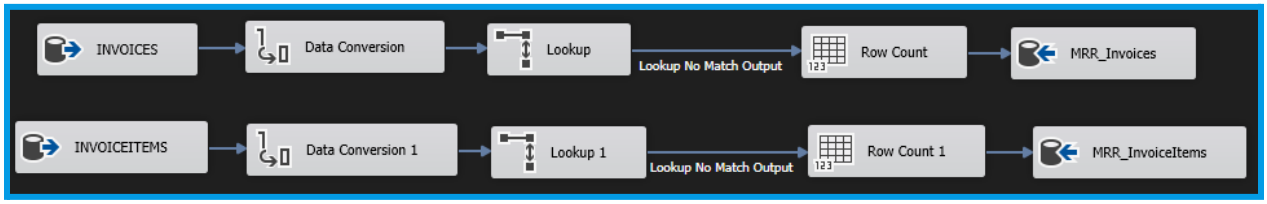


The geography-related tables from the source database

Fact-related tables mirroring

Control Flow:

Data Flow:



The Fact/Sales related tables from the source database

- We use lookup only to extract data that isn't already in the Fact Sales table.

3.5. Staging Stage

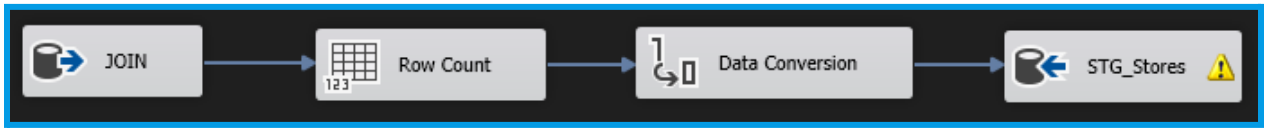
After completing the mirror phase, the ETL process transitions into the staging phase. The data is refined, standardized, and prepared for the subsequent transformations in the ETL process. Staging acts as a critical bridge, ensuring that the data is consistent and reliable before being loaded into the datamart for further analysis and reporting.

For each Dimension/Fact Staging table, we use a JOIN SQL query to retrieve the relevant columns from the correlating mirror tables.

Data Task pipelines

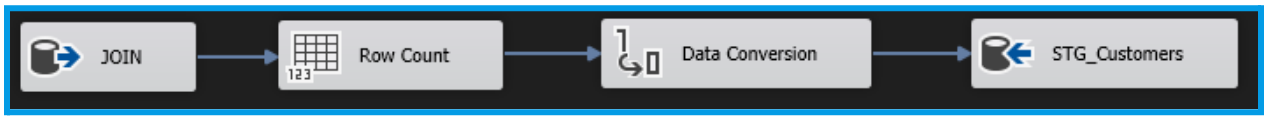
Data Flows:

- **Stores:**



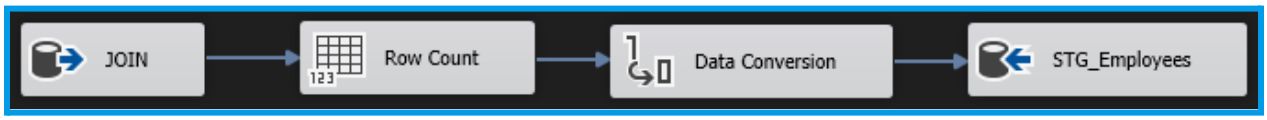
Joined Tables: MRR_Branches, MRR_States, MRR_Countries

- **Customers:**



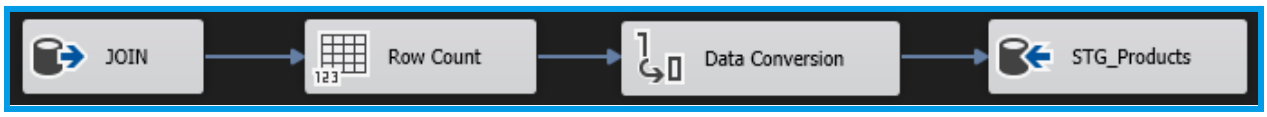
Joined Tables: MRR_Customers, MRR_DestCodes, MRR_States, MRR_Countries, InvoicesHelper

- **Employees:**



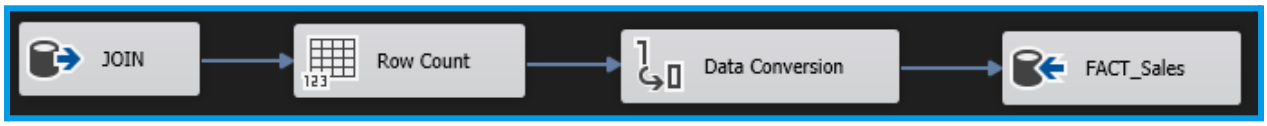
Joined Tables: MRR_Employees, InvoicesHelper

- **Products:**



Joined Tables: MRR_Part, MRR_FamilyTypes, MRR_Family

- **Sales:**



Joined Tables: MRR_Invoices, MRR_InvoiceItems

3.6. DataMart Stage

The Data Mart is a specialized subset of the overall data warehouse, tailored to meet the specific analytical needs of the sales department. In this stage, data is further transformed, aggregated, and organized into dimensional models that align with the business requirements of the sales team.

Upsert (Merge) Approach

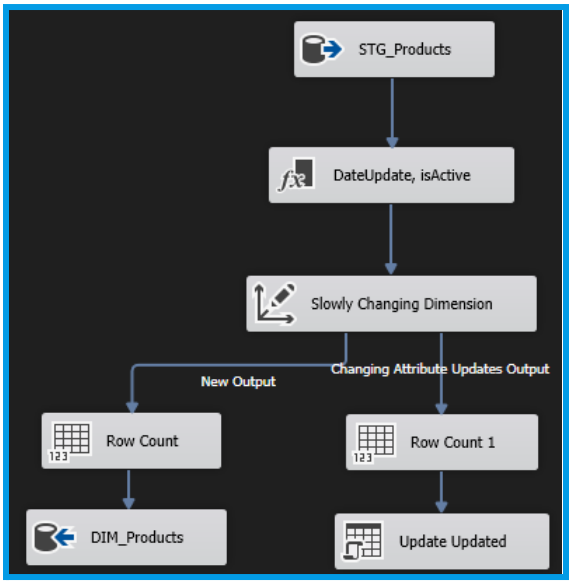
When applied to the transition from the staging table to the dimension table, this method involves comparing the data in the staging table with the existing records in the dimension table. Rows that match based on certain criteria (e.g., primary keys or unique identifiers) are updated with the latest information from the staging table, ensuring that the dimension table reflects the most recent and accurate data.

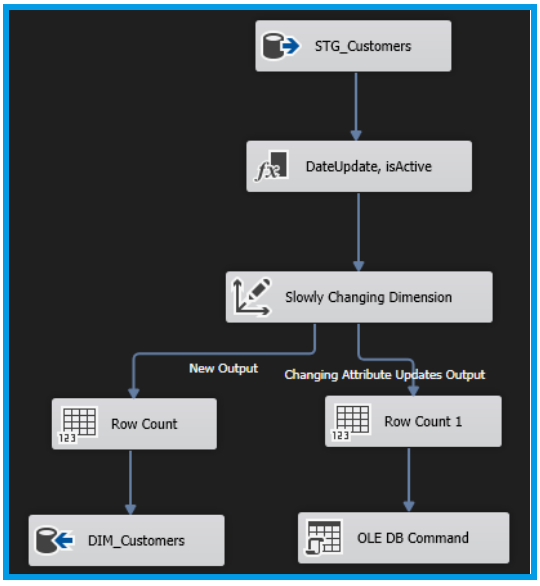
Stores Dimension & Employees Dimension

Slowly Changing Dimension Approach

The Slowly Changing Dimension approach ensures that changes in dimension attributes are handled appropriately, aligning the data mart with business requirements and enabling users to analyze data trends over time. This meticulous management of dimension changes enhances the data mart's ability to provide accurate and comprehensive insights for decision-making within the specific context of the sales department.

Products Dimension & Customers Dimension



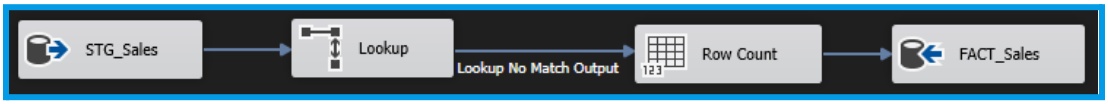


Fact Sales

The population of the fact table is performed by a simple lookup transformation that filters irrelevant data (e.g. Data that is already recorded in the fact table).

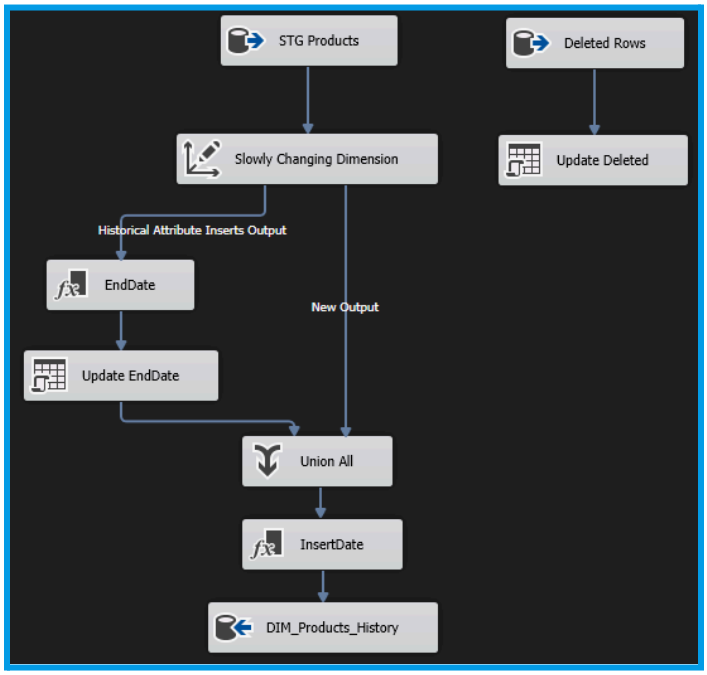
Control Flow:

Data Flow:



3.7. History Table for Products Dimension

Utilizing the SCD methodology, the History Table retains historical versions of product attributes, allowing for the tracking of changes such as updates, inserts, and deletions. Each row in the table represents a specific version of a product, complete with effective dates that delineate when the information became valid.



4.1. Jobs Overview

SQL Agent jobs consist of one or more steps, each representing a specific action or set of actions. These jobs can include tasks such as data extraction, transformation, loading, backups, and other database maintenance activities. The scheduling capabilities of SQL Agent enable users to plan and automate these tasks, reducing manual intervention and ensuring timely and consistent execution.

These SQL Agent jobs collectively form a robust and automated ETL process, enabling the Sales Department to harness the power of data for informed decision-making and strategic insights.

4.2. Jobs Definition

ETL_Mirror:

Step	Name
1	MRR_DIMS
2	MRR_FACTS

ETL_Staging:

Step	Name
1	STG_Stores
2	STG_Customers
3	STG_Employees
4	STG_Products
5	STG_Sales

ETL_DataMart:

Step	Name
1	DIM_Stores
2	DIM_Customers
3	DIM_Employees
4	DIM_Products
5	FACT_Sales

ETL_Run_All:

Step	Name
1	MRR_DIMS
2	MRR_FACTS
3	STG Stores
4	DIM Stores
5	STG_Employees
6	DIM_Employees
7	STG_Products
8	DIM_Products
9	STG_Customers
10	DIM_Customers
11	STG_Sales
12	FACT_Sales

4.3. Jobs Scheduling

The scheduling capabilities of SQL Agent enable users to plan and automate these tasks, reducing manual intervention and ensuring timely and consistent execution.

ID	Name	Enabl...	Description
1014	Morning Update	Yes	Occurs every day at 8:00:00 AM. Schedule will be used starting on 1/4/2024.

4.4. Development Environment

After confirming that the ETL process was functioning end-to-end in the production environment and that all the data was accurate and transferred as planned into the dimension and fact tables, I established a development environment. I copied all the tables, along with the production data, to this new environment. This allows for additional development work in the new environment without impacting the already deployed and operational production environment.

This is done with the following task in SSIS: