Dominick's Finer Foods Data Warehouse

REPORT BY :
MEGHNA PRADHAN, REKHA JACOB THOTTAN, TANAY MEHENDALE
DATE: 11/09/2024

TABLE OF CONTENTS

1. Introduction

Dominick's Finer Foods at the height of business operated over 100 stores and held 25% of the retail grocery market in Chicago. As a leading grocery chain, Dominicks offered a wide range of inventory to cater to the taste and requirements of their customers. During their heyday they were known for their high quality meat and produce, however since their acquisition by Safeway in 1998, the chain has seen a steep drop in market share.

The goal of this project is to design a data warehouse solution that can address the critical business questions enabling data-driven decision making. The data warehouse will provide insights into the sales patterns, personalized customer preferences, inventory management, this solution aims to help Dominick's regain and potentially surpass the market share they once held in the 1980s, positioning them for future growth.

By leveraging advanced analytics and various data visualization tools, the data warehouse will empower Dominick's Finer Foods and help the company identify key trends and business opportunities in real time. This includes understanding regional customer behavior, optimizing promotional strategies, and reducing operational inefficiencies. With this new data architecture, Dominick's will be able to not only enhance customer satisfaction but also be able to streamline supply chain processes ensuring that there is availability of high-demand products. This comprehensive approach and data driven decisions will position the company for success.

1.1 Understanding Dominick's Finer Foods Issues and Leveraging Data Warehousing for Growth

1. **Optimizing Shelf Management for Sales Growth.**
   A key challenge for DFF is analyzing consumer demographics such as age distribution, sales during holidays etc. A data warehouse solution will help DFF gain further insights into this.
   Effective shelf management plays a very important role in increasing sales and can
   Can be categorized into two primary strategies:
   a. Out of Store tactics - This focuses on attracting new consumers and retaining the customers that already frequently visit the store, this helps to counter external competition.
   b. In Store tactics - These tactics aim to increase sales among customers that are physically present in the store.

2. **Understanding Consumer Demographics and Sales Trends.**
   A fundamental setback for DFF has been the lack of analysis on their target consumers and demographics. Shifting their operations to California wouldn't have been such a disaster if they had an effective targeted analysis strategy to inform their logistic and

business decisions. Being able to better understand demographic data such as age, distribution, economic conditions, and household sizes would enable DFF to implement targeted marketing strategies.

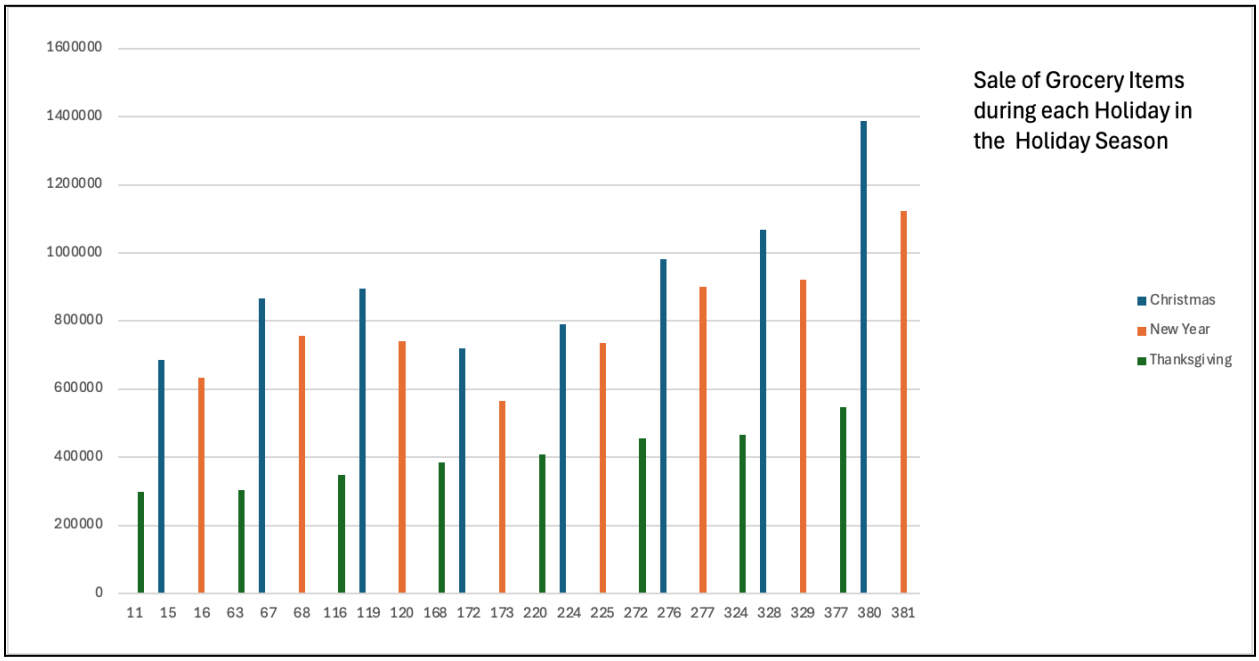
3. **Analyzing the Effectiveness of Price Promotions**
   DFF frequently implements promotional strategies, such as issuing coupons across various product categories, to boost sales. However the company should evaluate the effectiveness of these price promotions. By performing such analysis DFF can ensure that such promotions are effective.

1.2 Understanding the Data

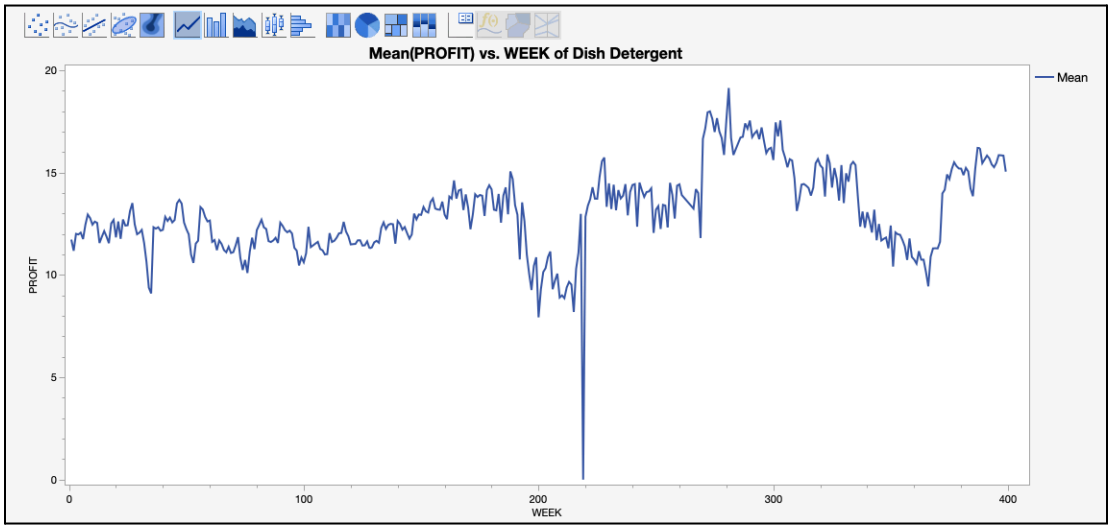
The data described revolves around shelf management, pricing, sales, and customer behavior for Dominick's Finer Foods. The dataset covers multiple aspects of retail management and also contains data of the demographic that the retail targets.

To gain a better understanding of the data we combined and aggregated the data across the different source files, for example for the table below “Sale of Grocery Items during each Holiday in the Holiday Season” we used the week code table to find the weeks which were tagged as a particular holiday and compared those weeks against the sale of grocery items in a week in the CCount table.

Working with data across tables helped us gain insight into the relationship between the different source files and what possible business questions could arise from the data set that we currently obtained.



We utilized tools like JMP and Excel Pivot Tables to uncover relationships within the data.



1.3 MetaData

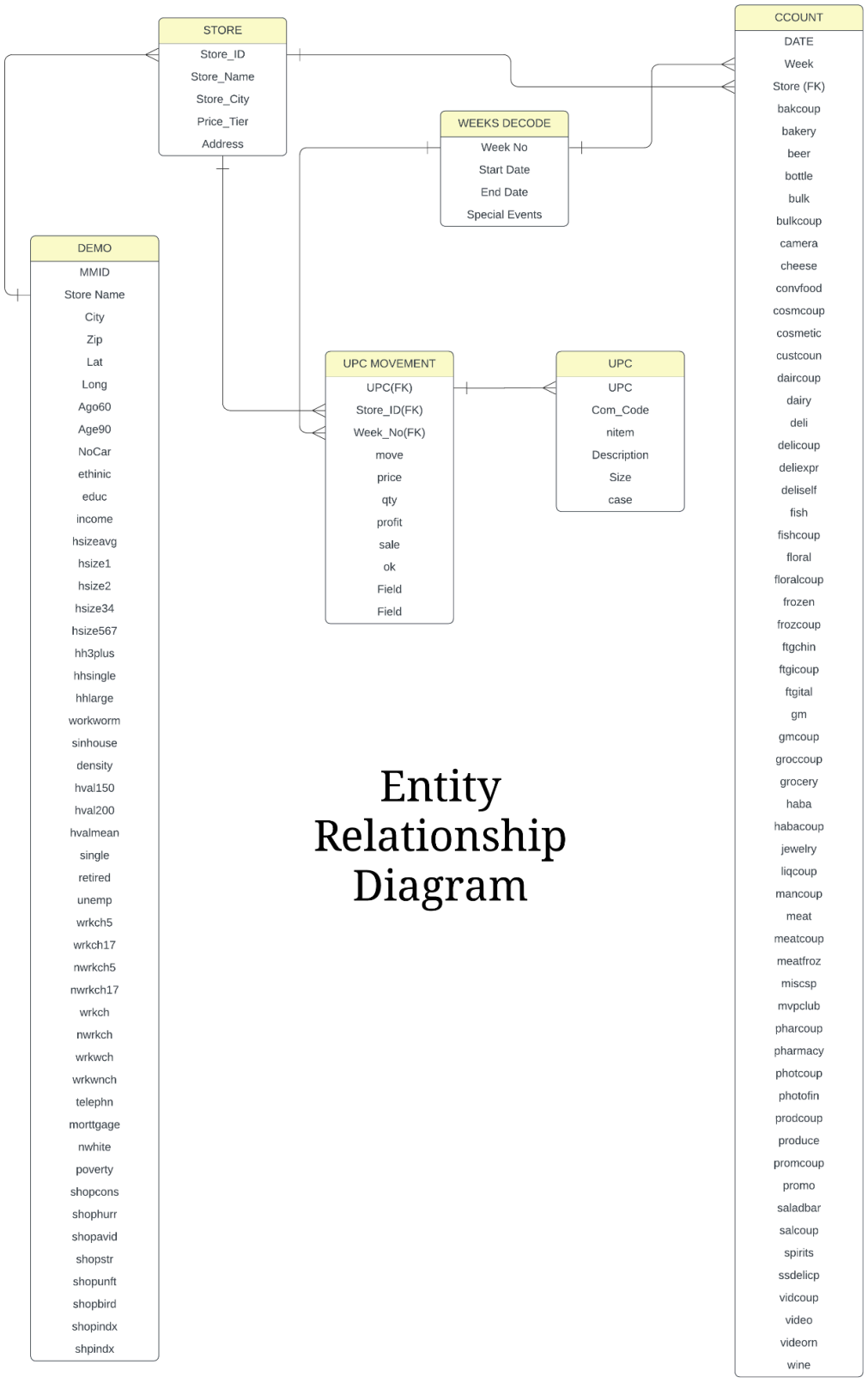
The Dominick's Finer Foods dataset consists of the following metadata, the dataset contains multiple files which are at different levels of granularity.

Source Files	Variables	Description
Customer Count File	The Customer Count File contains data that describes the traffic in each store, it includes data like customer count, daily sales figure and coupon usage of particular categories per week. Granularity - WEEK The file contains the sale of products and use of coupons by week marked by the WEEK field, the DATE field marks the point in time when the row value was recorded.	DATE : Date of Observation WEEK: Week number STORE: Store Code DAIRY: Dairy Sales FROZEN: Frozen Product Sales MEAT: Meat Product Sales SALADBAR: SaladBar Sales FLORAL: Floral Product Sales DELI: Deli Items Sale BAKCOUP: Bakery Coupon amounts redeemed LIQCOUP: Liquor Coupon amounts redeemed … Rest of the variables

		included daily sales figures of particular grocery items, coupon redemption values.
Store Level Demographics	The Store Level Demographic contains data that is based on the U.S government census for the Chicago metropolitan area, it provides demographic profiles for every store Granularity -STATIC, YEAR The demographic file has data with different granularity per field, most data is static example NAME, CITY, the rest of the data is by YEAR. The data originally comes from the U.S government census data for the Chicago metropolitan area for the year 1990.	MMID : Unique Identifier for each store NAME: Name of Store CITY: City the Store is located in ZIP: Zip Code of Store LAT: Latitude the store is located in LONG: Longitude the store is located in WEEKVOL: Volume of customers received AGE9: Population under the age of 9 AGE60: Population above the age of 60 ETHNIC: % of Black and Hispanic People EDUC: % of College Educated population NOCAR: % of population without a car INCOME: Log of median income WORKWORM: % of working women fulltime … included many other information of the population in the particular demographic
UPC Files	Contains a singular record each UPC (Universal Product Code) in a category. The information includes product name, size,	UPC : UPC number COM_CODE: Dominick's Commodity Code NITEM: Dominick's Item

	commodity code and item codes Granularity - STATIC The files contain static data regarding Product item details hence have a static granularity.	Code DESCRIP: Product Name SIZE: Product Size CASE: Number of items in a case
Movement Files	Contains data of weekly sales for each UPC in every store over a period of five years. Granularity - WEEK The data in movement files have granularity of a WEEK marked by the field with the same name,	PRICE: Unit Price UNITS_SOLD: Number of units sold PROFIT_MARGIN: Profit margin per unit DEAL_CODE: Promotion codes
Dominick's Stores	Contains data regarding all of the Dominick's stores within the retail chain. Granularity - STATIC The files have data regarding a particular store in the grocery chain and are static values.	STORE: Store unique Id CITY: City the store is located in PRICE_TIER: Price tier the store falls under, varies between high, low and medium ZONE: Zoning are ZIP CODE: Zip Code of store ADDRESS: Address of store

1.4 Entity Relationship Diagram



1.5 Retail Domain Understanding

In the retail sector, marketing and sales are impacted by consumer behavior, product pricing, promotions, and the efficiency of store layouts. In this case study of Dominick's Fine Foods (DFF), the data collected from 100 stores from 1989 to 1994 provides valuable insights into these factors. Such data forms a rich resource to identify market strategies that drive sales and customer engagement.

The research on retail store management and pricing strategies reveals several key areas that can be relevant to DFF's case. We have used the information from a combination of papers, https://www.company-histories.com/Dominicks-Finer-Foods-Inc-Company-History.html, https://www.chicagobooth.edu/research/kilts/datasets/dominicks and web search to give us an understanding of the marketing and sales within the retail domain.

From our research, we have learned the following about the retail domain,

1. **Shelf Management and Product Placement** - In the retail industry, the placement and organization of products place a crucial role in sales by manipulating the consumer. This is often referred to as shelf management. For example, products placed at eye level tend to sell more in comparison to products at lower or higher levels. A study shows that arranging cereals by type instead of by manufacturer decreased sales by 5%. This is because it reduced the chances of the customer coming across and buying additional products from the manufacturer which is referred to as “opportunistic sales”. [Strauss, 1998, “A Marketing Professor’s Shopping List”][1].
2. **Pricing Strategies** - Sales in the retail industry is greatly influenced by how consumers respond to price changes and promotions. Temporary price changes can cause sharp increase in sales but the effect depends on local competition. The sales are also varied for different product categories and stores. Research shows that “Everyday Low Pricing (ELDP)” is a difficult strategy for traditional retailers to maintain consistently. [Strauss, 1998, “A Marketing Professor’s Shopping List”][1]. This combined with the data about DFF sales shows that DFF should evaluate the pricing approach.
3. **Consumer Demographics** - To achieve highly targeted marketing, it is crucial to understand different demographics and their purchasing behavior. Factors like age, household size, income levels, family size can be used to bring more targeted promotions and product varieties. [Carpenter, Moore, 2006, “Consumer demographics, store attributes, and retail format choice in the US grocery market”][2]
4. **Impact of Store Location & Micro Marketing** - Research shows that accessibility and convenience of store locations are two major factors that influence the sales. Further, being in the proximity of a complementing business can also increase the pedestrian traffic to the stores. Studies have shown that retailers should not only look at the current state of the location but the potential to grow as well. It is seen that modern retailers rely on location intelligence and data analytics to make informed decisions about store

placement. While pricing based on store location seems logical, research also suggests that it might not always be profitable. [Jaravaza, Chitando, 2013, The Role of Store Location in Influencing Customers' Store Choice][3]. Using this information, we can find strategies for DFF for local market adaptation.

5. **Real Time Analytics** - Retailers can offer product recommendations, promotions, and even optimize inventory using real time analytics. This allows retailers to make data-driven decisions. Studies show that while setting up of real-time analytics requires a robust data infrastructure, the advantages it brings in terms of customer satisfaction, increased sales and improved operational efficiency makes it a key tool in the modern retail domain. [Tan, 2024, "Data Analytics: Challenges in Retail Basic Data Analytics"][4]

2. Business Questions, their Substantiations and Explanations

Out of the following 10 Business Questions initially presented, DFF has chosen the BQs 1, 2, 4, 5, 6. These chosen BQs and their justification is presented first. Later this section also explains the BQs that were not chosen by DFF.

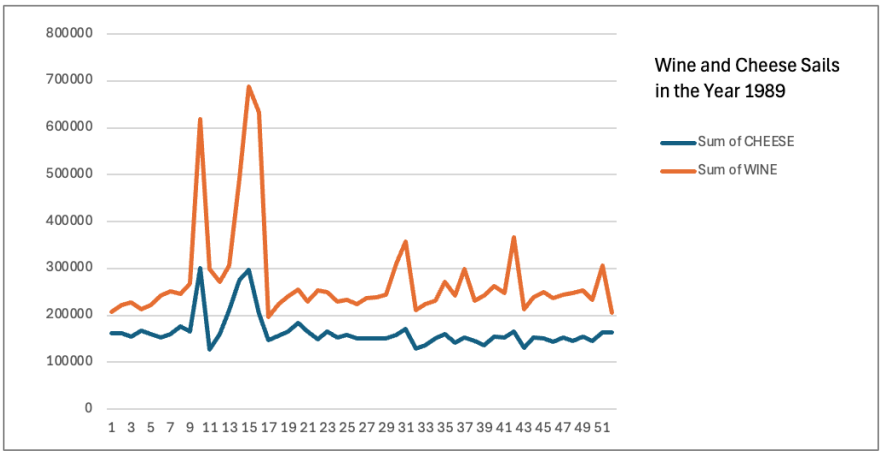
BQ1 What is the trend of wine sales during the holiday seasons - Christmas, Thanksgiving and New Year? How does the cheese sales trend during the same period and is there a correlation between wine and cheese?

Justification

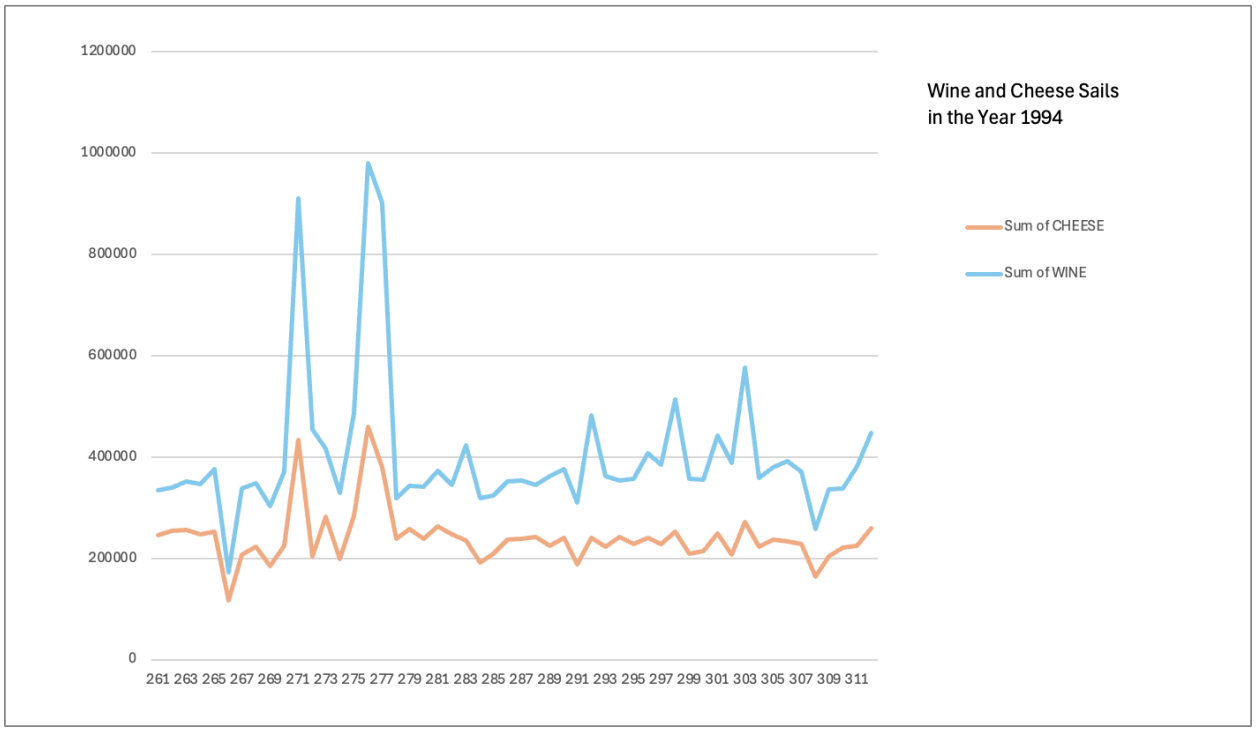
Products like wine are a huge part of holidays like Christmas, Thanksgiving and New Year. The analysis of sales with respect to seasonal demand will help DFF understand how to better manage these products during this period. In the previous section, we have seen how proper shelf management is crucial for sales in the retail domain. DFF can study similar surges in products during seasons to ensure proper shelf management, increase inventory and also run targeted promotions.

Further, it is common knowledge that wine and cheese are globally a good pair. But comparing the sales of cheese with wine during the holiday season, DFF can identify trends to decide if cheese should be placed near the wine aisle during holiday season so as to encourage opportunistic sales.

Pivot Table and Chart



In the above chart, a spike is seen during the weeks 11, 15,16 which are the thanksgiving, christmas and new year weeks for the year 1989. From the graph, it is seen that both cheese and wine sales show a spike during the holiday season. Further, the sales trends of cheese and wine are similar. This information will help DFF to stock cheese next to wine aisles and also ensure sufficient stock is available during the holiday season.



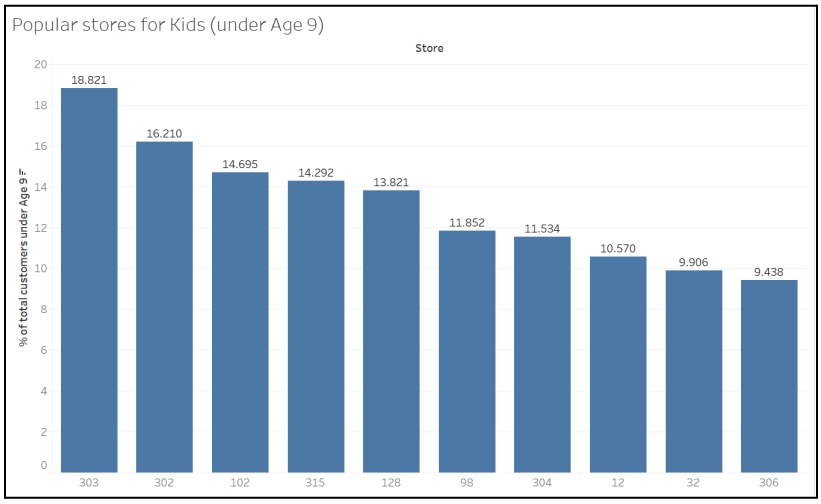
Similar to the previous chart, for the year 1994 as well the trend in wine and cheese sales are related to each other. Here the weeks 272,276 and 277 are the holiday weeks.

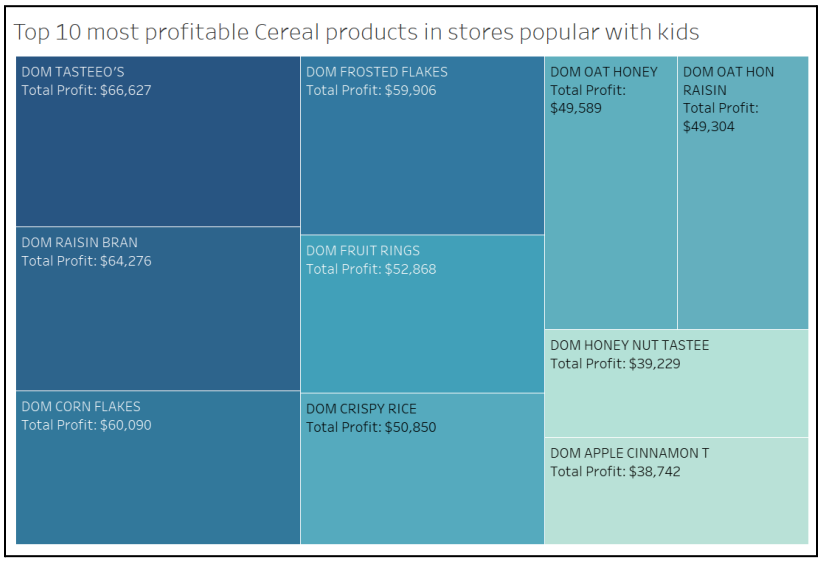
BQ2 Find the top 5 most profitable products in the Cereal category in 1994 in the store that is most popular to kids?

Justification

Identifying which of these cereal products are more popular among kids will help DFF understand their demographic better. DFF can focus on these products by placing them on easily visible shelves to increase the sales of this category. This information can also be used to stock other similar stores with these cereal products. Further, new launches of products in non-competing categories for kids can be bundled with these popular cereal products for pilot sale. Similarly, products from low selling categories can also be bundled with these popular products to increase sales in those categories. Promotional offers like Buy One Get One can also be done using this information.

Pivot Table and Chart





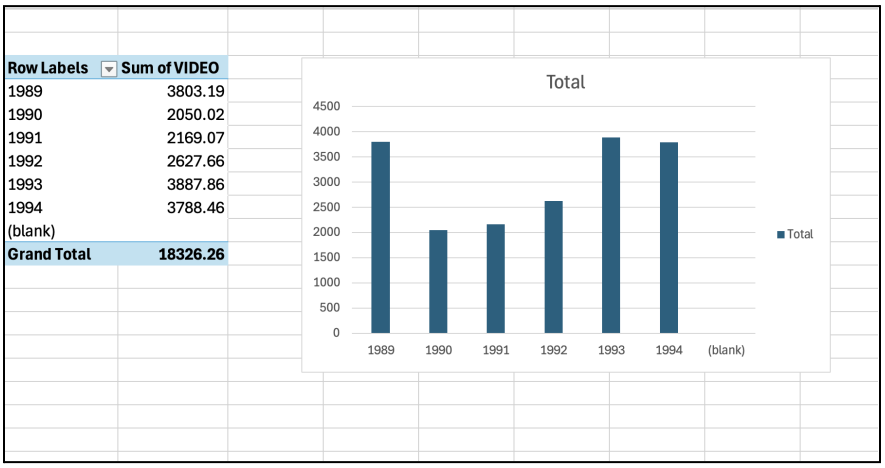
BQ4 What is the trend in sales of products like video and meat from the year 1989 to 1994?

Justification

This analysis helps identify how different products behaved across years and can be used for tracking the performance of different products. Based on this analysis, DFF can decide if there is a decline or incline in sales of videos and accordingly reduce their inventory for the following years. These types of questions are good for understanding the changes in consumer behavior. DFF can also choose to remove items from the stores, depending on the decline seen in these graphs.

Pivot Table and Chart

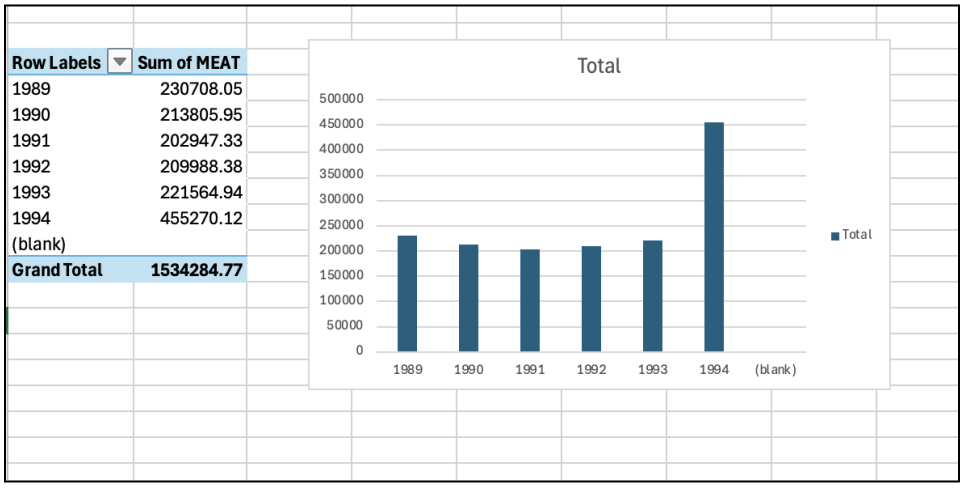
Sales of Video



Row Labels	Sum of VIDEO
1989	3803.19
1990	2050.02
1991	2169.07
1992	2627.66
1993	3887.86
1994	3788.46
(blank)	
Grand Total	**18326.26**

Sales of video showed a decline in years 1990 and 1991, however it rose in 1993 and is steady in 1994. This shows that DFF can continue to have similar stocks for the following year.

Sales of Meat



Row Labels	Sum of MEAT
1989	230708.05
1990	213805.95
1991	202947.33
1992	209988.38
1993	221564.94
1994	455270.12
(blank)	
Grand Total	**1534284.77**

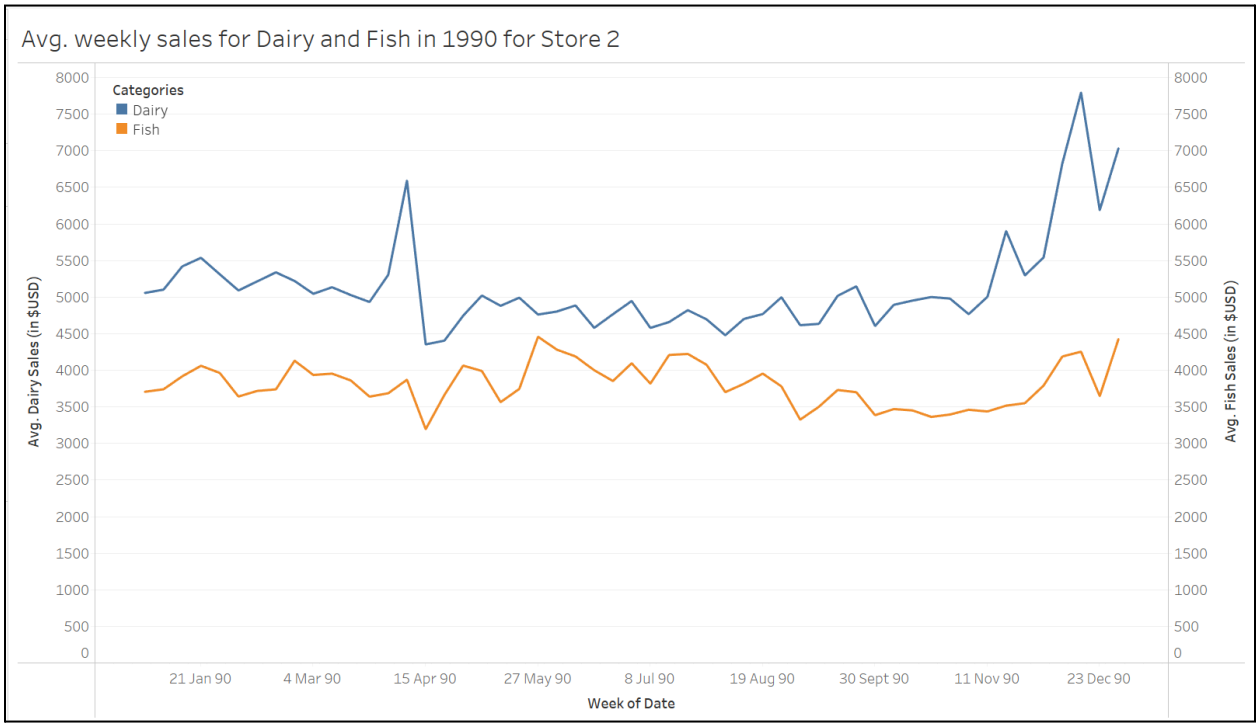
Here a sudden spike is seen in the year 1994 which shows that DFF should further look into the different reasons for this spike and accordingly manipulate their inventory.

BQ 5 What are the weekly sale trends for perishable goods like dairy and fish per store in the year 1990?

Justification

Sales of perishable products like dairy and fish have a strong relationship with the location of the store. The previous section talks about the impact of store location on sales. Through this analysis, DFF can understand the trend in sales of such items. This helps DFF plan the stock accordingly in these stores. Micro patterns like which weeks in a year show peaks and dips can be used to further understand the seasonal changes in the sale of these perishable items. This information can be further used to ensure no wastage of these products occurs in other weeks thereby increasing the profit margin in these categories. This information can also be used to increase the procurement. DFF can also plan for bundle offers to push sales during the low sales periods.

Pivot Table and Chart

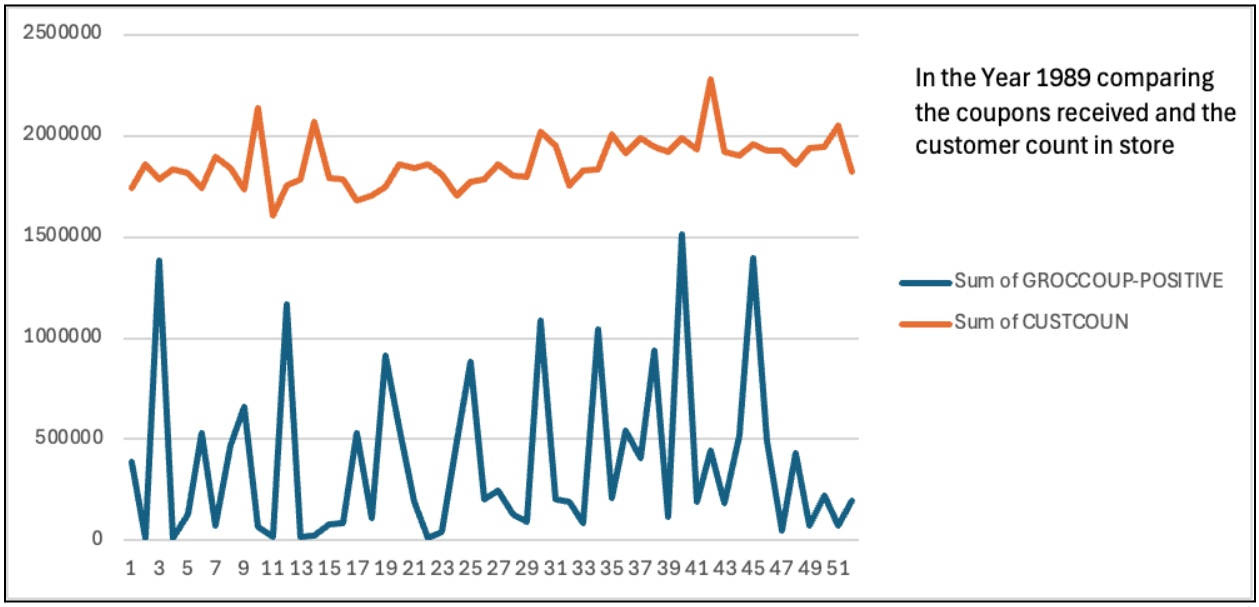


BQ 6 What are the weekly trend lines of total number of customers in a store and the total number of coupons redeemed? Is there a correlation between the two?

Justification

This analysis helps identify how different offers using coupons affect the store traffic. Coupons are a common strategy used by retail stores to increase store traffic. By comparing the redemption of different coupons with the store traffic, DFF can understand the success of these offers and promotions. This analysis can be used further to decide if more such offers should be made and in which stores can they start similar offers. DFF can also use this information to understand if offers should be made at specific times.

Pivot Table and Chart



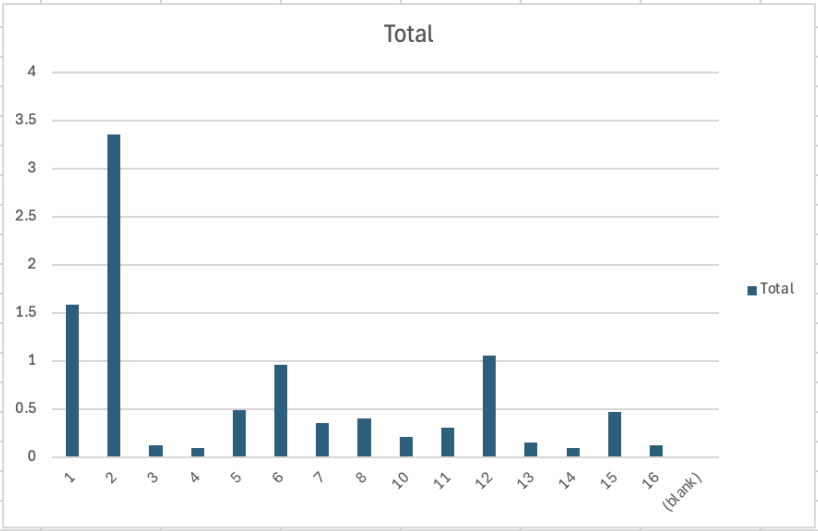
BQ3 Which zones are visited more by large sized households?

Justification

By identifying the zones that have a larger sized household population, DFF can identify the stores where they have to stock for jumbo packs of household items. This type of analysis is crucial for the retail domain as it allows for targeted marketing and sales. This information further allows DFF to stock other items that are popular among large households. For example, stores in large households will require jumbo packs of toilet paper more than small packs of toilet paper. They can also run promotional offers like discounts during weekdays to increase store traffic on otherwise non-peak hours.

Pivot Table and Chart

Row Labels	Sum of HHLARGE
1	1.591346759
2	3.356386925
3	0.130340815
4	0.101871469
5	0.493139468
6	0.969568846
7	0.362541495
8	0.40347715
10	0.214461158
11	0.316397841
12	1.062164572
13	0.154877575
14	0.096660967
15	0.473832139
16	0.125831821
(blank)	
Grand Total	9.852898999



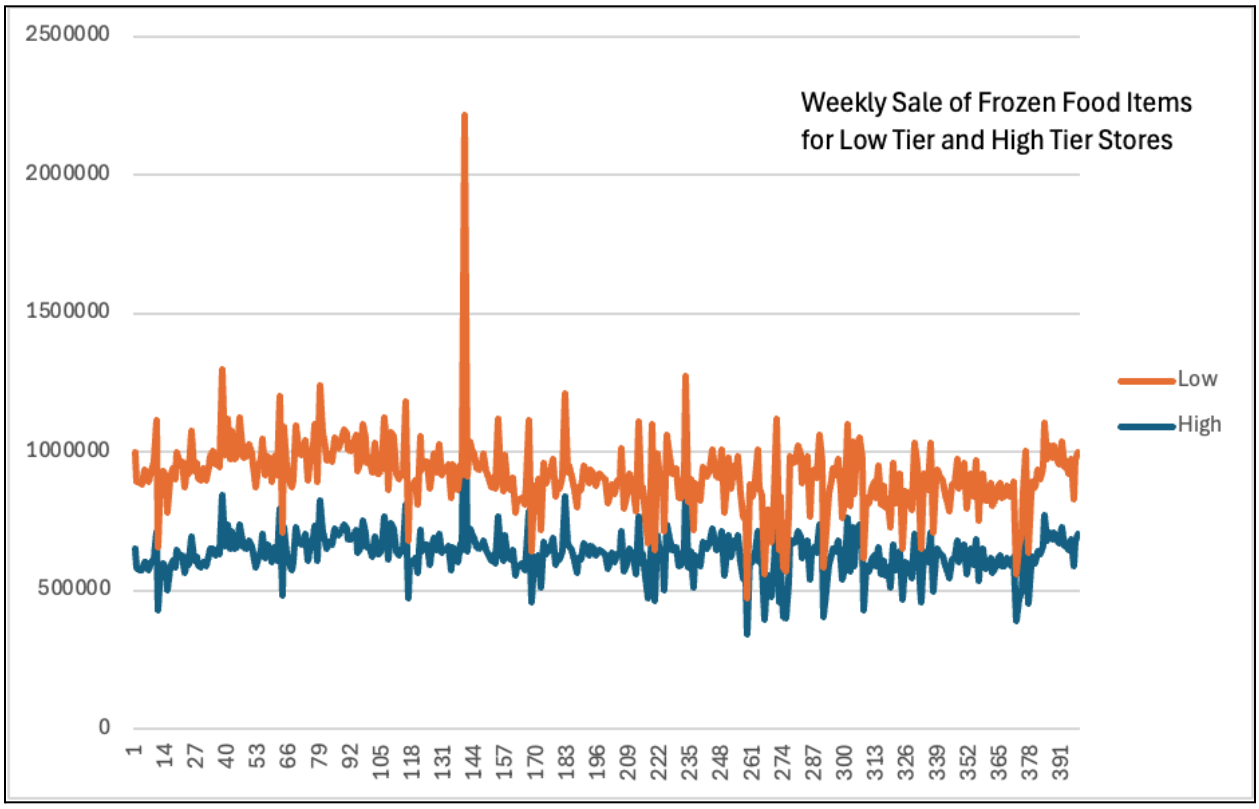
In the above graph, it is seen that zone 2 has a high population of large households. The stores in this zone can focus on selling products that are more lucrative to large households.

BQ 7 Compare the sales of ready to eat foods like canned soup and frozen dinners in the low tier stores and high tier stores. Does this trend tell us something about how to take advantage of these sales?

Justification

Ready to eat and canned food are generally popular in places where the consumers either lack time or cannot afford expensive groceries. The analysis of the sales of such products in the low and high tier stores will help DFF understand the behavioral differences in their consumers in these locations. This information can be then used to drive up sales by making promotional offers. For eg, in high tier stores, DFF can offer more healthier ready to eat options which can drive up sales. In the low tier, similarly DFF can bring in more cost effective products. This way DFF can maximize their profit margin

Pivot Table and Chart



BQ 8 What are the top 5 best selling categories in high-tier stores in 1991?

Justification

DFF can use this information to introduce more products into these categories in the high tier stores. This can drive up sales. Further, DFF can assign more shelf space to such categories thereby improving the shelf management mentioned in the previous section. Further, this information not only helps these stores to manage their inventory but also take advantage of the consumer behavior to run offers and promotions to drive up sales. This type of analysis is required for DFF to understand customer preferences.

BQ 9 Is there a relationship between price and sales volume for fish in the high tier stores?

Justification

Understanding the relationship between price and volume of sales of any product is required to make informed decisions around price sensitivity. Even though high tier stores have richer consumers, the sensitivity towards price can depend on quality of product, type etc. This can be seen especially in a product like fish, where consumers give high importance to quality in the

high tier community. A strong negative relationship between price and sales volume of fish indicates that occasional offers are required in this tier. A weaker correlation means the consumers prioritize quality over price and DFF can continue selling for higher prices without affecting the quality.

BQ 10 What is the yearly growth rate of paper towels in the 3 different tiers?

Justification
Paper towels are a staple household product. However analyzing the trends in different tiers is important to better forecast inventory. If lower tiers show growth in economy type labels, then DFF should stock more of these in the lower tiers. Like this, the higher tiers may opt for more eco friendly labels, then DFF should do more of this in the higher tiers.

3. Independent Data Marts Design using Kimball's approach

3.1 Steps in Kimball's Methodology

The Dimensional Modeling approach, or Kimball's Methodology, is one of the cornerstones in data warehouse design. It supports an architecture which focuses on creating a series of star schemas or data marts that are business-friendly, performance-optimized, and easier to understand. The whole process initiates with the recognition of a particular business process to be modeled, for example - sales or store information. This allows granularities in approaches towards data warehouse development.

Once a process has been selected, the next important decision is to determine the dimension's 'grain'. The grain is the level at which information is recorded in the fact table; it defines the level of detail where it is held and employed for analysis. The methodology helps designers to identify, next, the dimensions and facts: the former are the contextual attributes that outline the facts, whereas the latter are the numerical metrics generated by business process events. For example, for the sales fact table, one line item represents the total sales per store per week per product and for the StoreInfo fact table, one line item represents the total customer footfall per store per week.

Once these elements are defined, the star schema can be created. That is, one central fact table that is connected with surrounded dimension tables to make proper relationships between them. For a sales data mart, the sales fact table will be at the center connected with surrounding dimensions (store, time, and product). The star schema for this project has been explained in detail further in the report. The Kimball methodology goes further in emphasizing conformed

dimensions, meaning uniform dimensions that could be used across many fact tables for harmonization in the data warehouse.

The final steps involve determining the life expectancy of the database, planning for the storage of historical data and for future growth, and determining how to handle slowly changing dimensions. This latter consideration is crucial in deciding how to maintain accurate historical records as the attributes of dimensions change.

3.2 Importance of Kimball's Methodology

Following Kimball's approach for creating independent data marts is vital for several reasons -

1. The process-driven approach of Kimball's methodology ensures that every data mart addresses certain well-defined business needs, it ensures better understanding and greater acceptance by the business users.

2. Each data mart can be independently developed and deployed incrementally, which will give an organization the ability to start small and expand. Moreover, it allows adding new data marts without disturbing already existing ones, making the process agile.

3. Star schema designs implicitly help support higher query performance for doing analysis, which reduces complex joins and speeds up response times. This becomes important when the volume of data is huge and business users want speedier insights.

4. The use of conformed dimensions ensures consistency across different business processes, promoting a single version of the truth throughout the organization. This consistency is crucial for reliable decision-making and reporting across departments.

5. Because of the intuitiveness of the dimensional model, business users will find querying easier, which then simplifies report generation and dashboards. This could further reduce the learning curve for new users to a great extent and increase the adoption rate across the organization..

6. Eventually, the structured approach will result in shorter development times and efficient resource allocation. Because the structured approach provides a well-defined and replicable process for the development of data marts, development effort is more straightforward and resources can be deployed in focused fashion on core business areas.

3.2 Data Mart Matrix

The ***first step*** in the Kimball Modelling is to create the Data Mart Matrix table that lists the different data marts and the dimensions. For DFF, the five chosen Business Questions focus on 5 most profitable products popular to kids, total coupons redeemed per store and weekly sales of items like cheese, wine, video, meat, diary and fish. The source data to answer these questions are obtained from CCOUNT.csv, DEMO.csv, MOVEMENT and UPC files that are available on the Dominick's FF data website and Manual.

The four dimensions that will hold the data are Product, Store, Time and Coupon.
The two data marts will contain overlapping dimensions and are,

- Sales that will contain all sales related information that includes profit and margin per product per week per store. This data mart also aims to answer all future business questions associated with sales as the data grows.
- Store Information that includes store related data like the type of demographic that visits the store, coupons that are redeemed etc. This data mart not only answers the current BQ of coupons redeemed per store but also aims to answer futuristic questions like the type of customers that visit a store. This data mart can also be used along with the sales data mart using conformed dimensions to answer questions that require both types of data.

DATA MART	**DIMENSION**			
	dimProduct	**dimStore**	**dimTime**	**dimCoupon**
Sales	X	X	X	
Store Information		X	X	X

Table 1: Kimball's Data Mart Matrix

3.4 STAR Schema

The ***second step*** is the creation of fact tables. The fact table for the Sales Data Mart will contain information obtained from the source files - Movement and CCount.The grain of this fact table is that each line item in the fact table represents the sales of a particular product in a specific store during a week.
The fact table for the Sales Information Data Mart will be obtained from the source file CCount. The grain of this fact table is that each line item in the fact table represents the customer count of a particular store during a week.

The ***third step*** is the creation of dimension tables. The dimension Store stores the DFF store details like Store number, Price Tier, Zone, City and Zip Code obtained from Dominick's research project manual. It stores a field called AgeBelow9 obtained from the demographic csv file to show the store level demographics. In the future more columns can be added like AgeAbove60 etc to include more store level demographic details. The dimension Product stores details of the different products of each category obtained from the UPC.csv files. The name of the upc file is used to fill the category column to identify various categories like cheese, beer etc. The dimension Time stores the data related to the weeks and special events. The columns Month and Year are calculated from the Start Date and End Date. The dimension Coupon stores coupon redeemed information related to each different category and is obtained from the ccount file.

Each of the dimension tables also includes an auto generated surrogate primary key respectively like Store_ID, Product_ID, Time_ID and Coupon_ID to uniquely identify the stores. This is done so as to ensure that a consistent format is maintained in the future if the data is obtained from different sources.

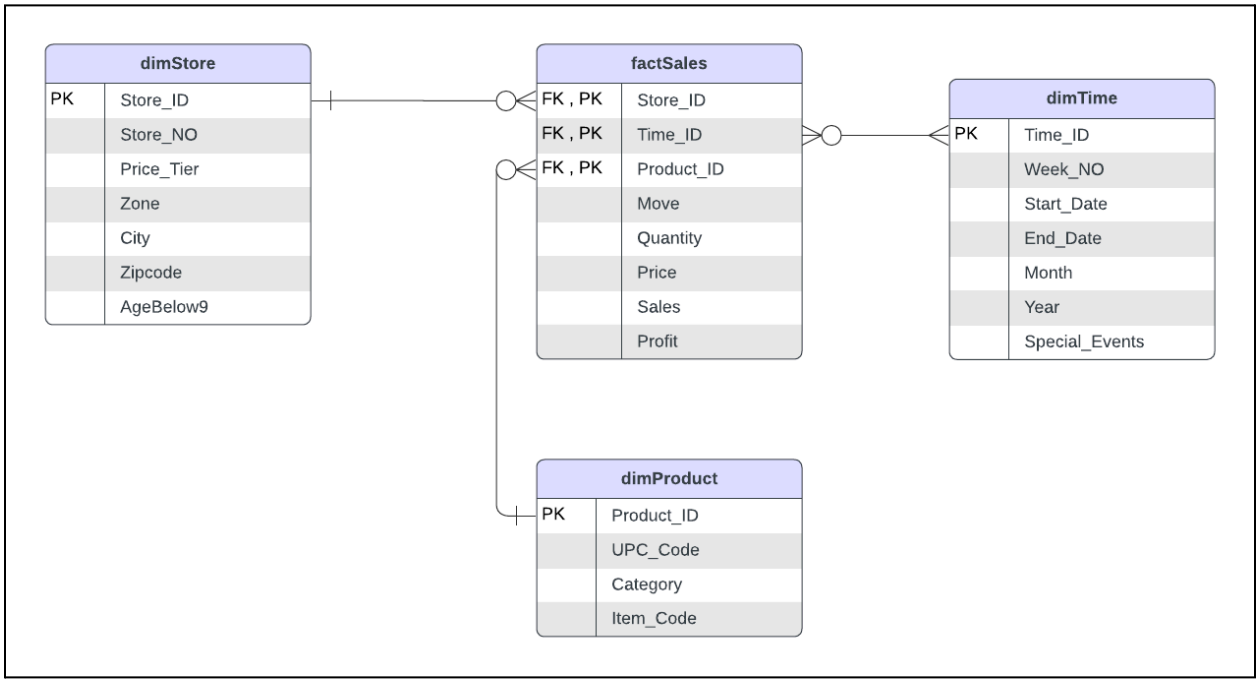


Figure 1: Sales Data Mart

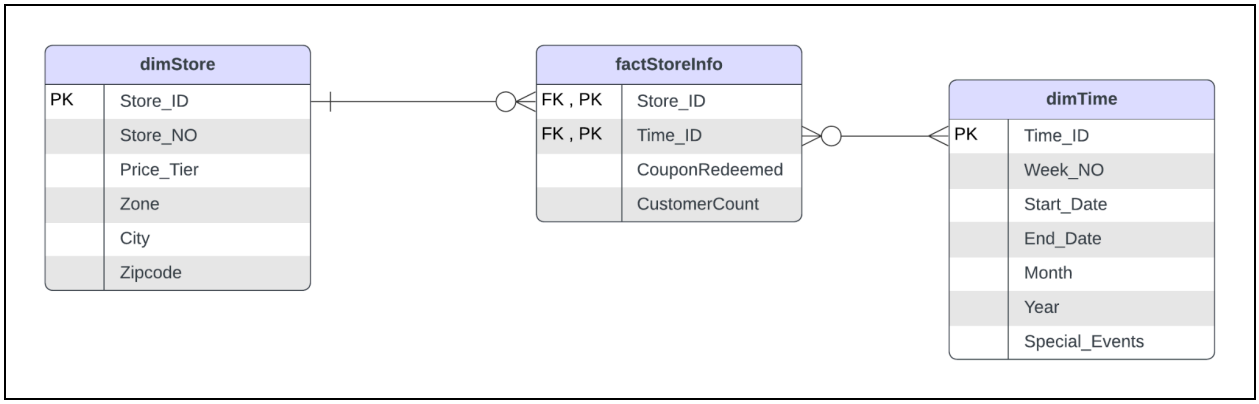


Figure 2: Store Information Data Mart

3.5 Justification of BQ and mapping to Schema

The ***fourth step*** is to obtain feedback for the logical design - the star schemas. This is done by analyzing if the schemas will help answer the chosen Business Questions and all future reports that the users expect to have.

Sales Data Mart

This data mart helps to answer BQ1, BQ2, BQ4, BQ5.

BQ 1 What is the trend of wine sales during the holiday seasons - Christmas, Thanksgiving and New Year? How does the cheese sales trend during the same period and is there a correlation between wine and cheese?

Justification - In Report 1, this question was answered using pivot tables on ccount and week encode files. The same can be achieved using the *Sales Data Mart*. We use the *Time Dimension* to obtain data from the *Week_No* attribute that has special events like holidays. This information combined with the product information on cheese and wine from the *Product Dimension* will be used to find the *Sales* values from the *Sales Fact* table. This is then plotted to show the relationship between the sales of the two products.

BQ 2 Find the top 5 most profitable products in the Cereal category in 1994 in the store that is most popular to kids?

Justification - In Report 1, this question was answered using pivot tables on demographic, movement and upc files. The same can be achieved using the *Sales Data Mar*t. We use the attribute *AgeBelow9* in the *Store Dimension* to find the stores popular with kids. This combined with the product information about the *Category Cereal* obtained from the *Product Dimension* will be used to look for the items with maximum *Profit* from the *Sales Fact* Table. This is then plotted to see the most profitable items popular among kids. In the future the same data mart can be used to answer other sales related questions for other demographics by adding more attributes to the dimensions.

BQ 4 What is the trend in sales of products like video and meat from the year 1989 to 1994?
Justification - In Report 1, this question was answered using pivot tables on ccount file. The same can be achieved using the *Sales Data Mart*. We use the product information on video and meat from the *Product Dimension* to identify the *Sales* value of the Products from the *Sales Fact* table. The *Year* attribute acts as a filter criteria for the *Time Dimension* to filter based on the years 1989 to 1994.

BQ 5 What are the weekly sale trends for perishable goods like dairy and fish per store in the year 1990?
Justification - In Report 1, this question was answered using pivot tables on ccount file. The same can be achieved using the *Sales Data Mar*t. We use the product information on dairy and fish from the *Product Dimension* to identify the *Sales* value of the Products from the *Sales Fact* table. The *Year* attribute acts as a filter criteria for the *Time Dimension* to filter based on the year 1990.

Store Information Data Mart
This data mart helps to answer BQ6

BQ 6 What are the weekly trend lines of total number of customers in a store and the total number of coupons redeemed? Is there a correlation between the two?
Justification - In Report 1, this question was answered using pivot tables on ccount file. The same can be achieved using the *Sales Information Data Mar*t. We use the total *custcount* information from the *Sales Fact* table, together with the sum of all coupons redeemed (attribute *CouponRedeemed*) to plot the graph that shows the relationship between the two.

This Data Mart can also be used in the future to answer Business Questions around store demographics. This can be achieved by including a new dimension *Demographic* that can store the number of customers at different age, income groups etc which can be obtained from the demographic source file.

3.6 Sources-to-Staging

Source Data	Source Data Field	Mapping	Staging Table Type	Staging Table Name	Attribute
Week_Decode_Table.csv	Week#	Copy	Relation	Time_staging	Week
	Start	Copy	Relation		Start_Date

	End	Copy	Relation		End_Date
	Special_Events	Copy	Relation		Special_Events
Dominicks_Stores.csv	Store	Copy	Relation	Store_staging	Store_NO
	City	Copy	Relation		City
	Price Tier	Copy	Relation		Price_Tier
	Zip Code	Copy	Relation		Zipcode
	Zone	Copy	Relation		Zone
Demo.csv	age9	Copy	Relation	Demo_staging	AgeBelow9
	Store	Copy	Relation		Store_NO
	MMID	Copy	Relation		MMID
	Zone	Copy	Relation		Zone
	SCluster	Copy	Relation		SCluster
	City	Copy	Relation		City
	Zip	Copy	Relation		Zip
upcxxx.csv	upc	Copy	Relation	UPCXXX_staging	upc_no
	nitem	Copy	Relation		item_code
	descrip	Copy	Relation		descrip
	size	Copy	Relation		size
	case	Copy	Relation		case
	com_code	Copy	Relation		com_code
wxxx.csv	upc	Copy	Relation	WXXX_movement_staging	upc_no
	store	Copy	Relation		Store_NO
	week	Copy	Relation		Week

	move	Copy	Relation		Move
	qty	Copy	Relation		Quantity
	profit	Copy	Relation		Profit
ccount.csv	Week	Copy	Relation	CCount_staging	Week
	Store	Copy	Relation		Store
	CUSTCOUN	Copy	Relation		CCount
	BAKCOUP	Copy	Relation		BAKCOUP
	BULKCOUP	Copy	Relation		BULKCOUP
	DAIRCOUP	Copy	Relation		DAIRCOUP
	DELICOUP	Copy	Relation		DELICOUP
	FISHCOUP	Copy	Relation		FISHCOUP
	FLORCOUP	Copy	Relation		FLORCOUP
	FROZCOUP	Copy	Relation		FROZCOUP
	FTGCCOUP	Copy	Relation		FTGCCOUP
	GMCOUP	Copy	Relation		GMCOUP
	GROCCOUP	Copy	Relation		GROCCOUP
	HABACOUP	Copy	Relation		HABACOUP
	LIQCOUP	Copy	Relation		LIQCOUP
	MANCOUP	Copy	Relation		MANCOUP
	MEATCOUP	Copy	Relation		MEATCOUP
	MISCSCP	Copy	Relation		MISCSCP
	PHARCOUP	Copy	Relation		PHARCOUP
	PHOTCOUP	Copy	Relation		PHOTCOUP

	PRODCOUP	Copy	Relation		PRODCOUP
	PROMCOUP	Copy	Relation		PROMCOUP
	SALCOUP	Copy	Relation		SALCOUP
	VIDCOUP	Copy	Relation		VIDCOUP

3.7 Staging-to-DataMart

Source Data in staging	**Staging table data field**	**Mapping**	**Data Mart Table Type**	**Table Name**	**Attribute**
Time_staging table	-derived column–	Transform (Autogenerated as surrogate key)	Dimension Table	dimTime	Time_ID
	Week	Copy	Dimension Table		Week_NO
	Start_Date	Copy	Dimension Table		Start_Date
	End_Date	Copy	Dimension Table		End_Date
	-derived column-	Transform (derived from the range of start_date and end_date)	Dimension Table		Month
	-derived column-	Transform (derived from the range of start_date and end_date)	Dimension Table		Year
Store_staging	-derived	Transform	Dimension	dimStore	Store_ID

table	column–	(Autogenerated as surrogate key)	Table		
	Store_NO	Copy	Dimension Table		Store_NO
	City	Copy	Dimension Table		City
	Price_Tier	Copy	Dimension Table		Price_Tier
	Zipcode	Copy	Dimension Table		Zipcode
	Zone	Copy	Dimension Table		Zone
Demo_staging table	AgeBelow9	Transform	Dimension Table		AgeBelow9 (This is needed only for the Sales Data Mart. The Store Information data mart will not have this column)
UPCXXX_staging table	Upc_code	Copy	Dimension Table	dimProduct	UPC_Code
	Item_Code	Copy	Dimension Table		Item_Code
	Category	Transform (based on the filename of UPC tables, the category will be determined)	Dimension Table		Category
WXXX_movement_staging table	Move	Copy	Fact Table	factSales	Move
	Quantity	Copy			Quantity
	Profit	Copy			Profit

CCount_staging table	CCount	Copy	Fact Table	factStoreInfo	CCount
	BAKCOUP	Transform (Aggregation of all Coupon fields from the CCount_staging table to the Coupon Redeemed)			CouponRedeemed
	BULKCOUP				
	DAIRCOUP				
	DELICOUP				
	FISHCOUP				
	FLORCOUP				
	FROZCOUP				
	FTGCCOUP				
	GMCOUP				
	GROCCOUP				
	HABACOUP				
	LIQCOUP				
	MANCOUP				
	MEATCOUP				
	MISCSCP				
	PHARCOUP				
	PHOTCOUP				
	PRODCOUP				
	PROMCOUP				
	SALCOUP				
	VIDCOUP				

3.8 Physical Design

The physical design of DFF's data marts will focus on ensuring scalability, performance, and efficient storage as the data grows. The **data aggregate plan** includes creation of summary tables. The data in the tables are stored at the week level. For future, this can also be rolled up to year and month level by making use of the year and month attributes in the *Time* Dimension. For the BQ coupons redeemed in stores, we have created the *Coupon* dimension where each row will have the sum of all coupons redeemed per store per week. This ensures the granularity of this dimension is inline with the others. In the future, we can also create summary tables for high-level sales trends and coupon usage to optimize performance for commonly requested reports like sales per month etc. The **indexing plan** includes creation of indexes for making retrieval faster and for making the tables query-centric. Auto-generated surrogate keys (_ID) are created for all dimensions like *Product*, *Store*, *Coupon* and Time. These are non clustered and help enhance query performance particularly for time-based and promotional analysis. This also ensures consistency in the event of obtaining data from different sources in the future. Clustered indexes are created in the *Sales* and *Sales Information* Fact tables based on the surrogate keys like *Time_ID, Store_ID, Product ID* and T*ime_ID, Store_ID, Coupon_ID* respectively. This is used to speed up joins with dimension tables.

For data **data standardization**, we will ensure consistent data formats across dimensions. This includes using the same date format (YYYY-MM-DD) for all date fields, standardized product categories in cases where data of products are present in the *ccount* files but not the *upc* files (eg:wine). A clear and concise naming convention is used for the staging and presentation tables and attributes. For the **storage plan**, we anticipate future growth and will use partitioning strategies for the *factSales* and *factStoreInfo* tables by *Time_ID* to manage large datasets efficiently. Further if required, we can choose to archive historical data older than five years into a lower-cost storage solution, ensuring that frequently accessed data remains in high-performance storage.

4. Data Cleaning and Integration

4.1 ETL Plan

4.1.1 Target Data

For our data warehouse solution we have modeled the following two Datamarts:

- Sales Data Mart - 3 dimension tables, 1 fact table
- Store Information Data Mart - 2 dimension tables, 1 fact table

The details and definitions for the dimension and fact tables used in each of the Datamarts defined above are presented in the following tables:

Data Mart : Sales Data Mart		
Dimension : dimStore		
DW Target Table	DW Target Column	DW Target Datatype
601_Group3_SalesData Mart.dimProduct	Store_ID [PRIMARY KEY]	INT
	Store_NO	INT
	City	VARCHAR(50)
	Price_Tier	VARCHAR(50)
	Zone	INT
	Zipcode	VARCHAR(50)
	AgeBelow9	FLOAT

Data Mart : Sales Data Mart		
Dimension : dimTime		
DW Target Table	DW Target Column	DW Target Datatype
601_Group3_SalesData Mart.dimTime	Time_ID [PRIMARY KEY]	INT
	Week_NO	INT
	Start_Date	DATE
	End_Date	DATE
	Month	INT
	Year	INT
	Special_Events	NVARHCAR(50)

Data Mart : Sales Data Mart

Dimension : dimProduct		
DW Target Table	DW Target Column	DW Target Datatype
601_Group3_SalesDataMart.dimProduct	Product_ID [PRIMARY KEY]	INT
	UPC_Code	VARCHAR(50)
	Item_Code	VARCHAR(50)
	Category	VARCHAR(50)

Data Mart : Sales Data Mart		
Dimension : factSales		
DW Target Table	DW Target Column	DW Target Datatype
601_Group3_SalesDataMart.factSales	Store_ID [FK , PK]	INT
	Time_ID [FK , PK]	INT
	Product_ID [FK , PK]	INT
	MOVE	INT
	QTY	INT
	PRICE	NUMERIC(18,2)
	PROFIT	NUMERIC(18,2)
	Sales	NUMERIC(18,2)

Data Mart : Store Information Data Mart		
Dimension : dimStore		
DW Target Table	DW Target Column	DW Target Datatype
601_Group3_StoreInformationDataMart.dimStore	Store_ID [PRIMARY KEY]	INT
	Store_NO	INT

	City	VARCHAR(50)
	Price_Tier	VARCHAR(50)
	Zone	VARCHAR(50)
	Zipcode	VARCHAR(50)

Data Mart : Store Information Data Mart		
Dimension : dimTime		
DW Target Table	DW Target Column	DW Target Datatype
601_Group3_StoreInformationDataMart.dimTime	Time_ID [PRIMARY KEY]	INT
	Week_NO	INT
	Start_Date	DATE
	End_Date	DATE
	Month	INT
	Year	INT
	Special_Events	NVARHCAR(50)

Data Mart : Sales Data Mart		
Dimension : factStoreInfo		
DW Target Table	DW Target Column	DW Target Datatype
601_Group3_StoreInformationDataMart.factStoreInfo	Store_ID [FK , PK]	INT
	Time_ID [FK , PK]	INT
	CouponRedeemed	NUMERIC(18,2)
	CustomerCount	INT

4.1.2 Data Sources

The following data that we use in our Data Warehouse solution is provided by James M. Kilts Center, University of Chicago Booth School of Business. The following table lists all of the source files and includes information of their origin.

Data File	Source
Week_Decode_Table.csv	Chicago Booth
Dominicks_Stores.csv	Chicago Booth
Demo.csv	The U.S. Government (1990) Census Data for the Chicago Metropolitan Area
upcxxx.csv	Chicago Booth
wxxx.csv	Chicago Booth
ccount.csv	Chicago Booth

4.1.3 Data Mappings from source to staging and staging to data warehouse

Data Mapping from Source Files to Staging Tables

For the transfer of data from the Source files to the staging tables, no transformation or data cleaning practices have been implemented expect for selection of required attributes.

Staging Area : Store_staging			
Source File	Source File Attributes	Staging Area	Staging Area Table Attributes
Dominicks_Stores.csv	Store	601_Group3_staging-area.Store_staging	Store_NO
	City		City
	Price Tier		Price_Tier
	Zip Code		Zipcode
	Zone		Zone

Staging Area : Time_staging			
Source File	Source File Attributes	Staging Area	Staging Area Attributes
Week_Decode_Table.csv	Week#	601_Group3_staging-area.Time_staging	Week
	Start		Start_Date
	End		End_Date
	Special_Events		Special_Events

Staging Area : Demo_staging			
Source File	Source File Attributes	Staging Area	Staging Area Table Attributes
Demo.csv	age9	601_Group3_staging-area.Demo_staging	AgeBelow9
	Store		Store_NO
	MMID		MMID
	Zone		Zone
	SCluster		SCluster
	City		City
	Zip		Zip

Staging Area : UPCCER_staging			
Source File	Source File Attributes	Staging Area	Staging Area Table Attributes
upccer.csv	upc	601_Group3_staging-area.UPCCER_staging	upc
	nitem		nitem

	descrip		descrip
	size		size
	case		case
	com_code		com_code

Staging Area : WCER_staging			
Source File	Source File Attributes	Staging Area	Staging Area Table Attribute
wcer.csv	upc	601_Group3_staging-area.WCER_staging	upc_no
	store		Store_NO
	week		Week
	move		Move
	qty		Quantity
	profit		Profit

Staging Area : CCOUNT_staging			
Source File	Source File Attributes	Staging Area	Staging Area Table Attribute
ccount.csv	Week	601_Group3_staging-area.CCOUNT_staging	Week
	Store		Store
	CUSTCOUN		CCount
	BAKCOUP		BAKCOUP
	BULKCOUP		BULKCOUP
	DAIRCOUP		DAIRCOUP

	DELICOUP		DELICOUP
	FISHCOUP		FISHCOUP
	FLORCOUP		FLORCOUP
	FROZCOUP		FROZCOUP
	FTGCCOUP		FTGCCOUP
	GMCOUP		GMCOUP
	GROCCOUP		GROCCOUP
	HABACOUP		HABACOUP
	LIQCOUP		LIQCOUP
	MANCOUP		MANCOUP
	MEATCOUP		MEATCOUP
	MISCSCP		MISCSCP
	PHARCOUP		PHARCOUP
	PHOTCOUP		PHOTCOUP
	PRODCOUP		PRODCOUP
	PROMCOUP		PROMCOUP
	SALCOUP		SALCOUP
	VIDCOUP		VIDCOUP

Data Mapping from Staging Tables to Data Warehouse

Dimension : dimStore				
Staging Table	Staging Area Columns	Transformation	DW Table	DW Attributes

601_Group 3_staging-area.Store_staging	-derived column–	Derived Column Transformation (Autogenerated as surrogate key)	601_Group3_SalesDataMart.dimStore	Store_ID
	Store_NO			Store_NO
	City			City
	Price_Tier			Price_Tier
	Zipcode			Zipcode
	Zone			Zone

Dimension : dimProduct				
Staging Table	Staging Area Columns	Transformation	DW Table	DW Attributes
601_Group 3_staging-area.UPCXXXX_staging table	Upc_code	Copy	601_Group3_SalesDataMart.dimProduct	UPC_Code
	Item_Code	Copy		Item_Code
	Category	Derived Column Transform for Category (based on the filename of UPC tables, the Category will be determined)		Category
	Price_Tier			
	Zipcode			
	Zone			

Dimension : dimTime				
Staging Table	Staging Area Columns	Transformation	DW Table	DW Attributes
601_Group 3_staging-area.Time_staging	-derived column–	Transform (Autogenerated as surrogate key)	601_Group3_SalesDataMart.dimTime	Time_ID
	Week	Copy		Week_NO
	Start_Date	Copy		Start_Date

	End_Date	Copy		End_Date
	-derived column-	Transform (derived from the range of start_date and end_date)		Month
	-derived column-	Transform (derived from the range of start_date and end_date)		Year

Fact : factSales				
Staging/Dimension Table	Staging Area/Dimension Columns	Transformation	DW Table	DW Attributes
601_Group3_SalesDataMart.dimStore	Store_ID	Lookup	601_Group3_SalesDataMart.factSales	Store_ID
601_Group3_SalesDataMart.dimTime	Time_ID	Lookup		Time_ID
601_Group3_SalesDataMart.dimProduct	Product_ID	Lookup		Product_ID
601_Group3_staging-area.WXXX_movement_staging table	Move	Copy		Move
	Quantity	Copy		Quantity
	Profit	Copy		Profit
601_Group3_staging-area.CCount_staging table	Price	Copy		Price
	Sales	Copy		Sales

Fact : factStoreInfo

Staging Table	Staging Area Columns	Transformation	DW Table	DW Attributes
CCOUNT_ staging table	*all COUP fields in the CCOUNT staging	Aggregate (Aggregation of all Coupon fields from the CCOUNT Table to the Coupon Redeemed Attributes)	601_Group3_ StoreInfoDataMart.factStoreInfo	CouponRedeemed
	CCount	Copy		CustomerCount
601_Group3_StoreInfoDataMart.dimStore	Store_ID	Lookup		Store_ID
601_Group3_StoreInfoDataMart.dimProduct	Time_ID	Lookup		Time_ID

4.1.4 Data Extraction Rules

Data Extraction is the first step in the ETL process. It is the process of extracting data from the source systems/files and loading into the staging area. The staging area is where all the transformations take place before the data is loaded into the presentation layer (Data Warehouse). For this report, the data is loaded into Microsoft SQL Server Studio tables to ensure consistency between the data obtained from various different sources (Eg: csv files and the data from the DFF data manual).

In this report, the data from the sources are extracted and loaded as in into the staging tables without performing any transformations. Hence, there are no explicit Data Extraction Rules. The data in these staging tables are then subjected to cleansing and transformation rules to load them into the Data Warehouse. This transformation can be seen in the following section.

4.1.5 Data Transformation Rules

The data present in the staging tables are subjected to transformation and cleansing rules to ensure the loading of clean data into the Data Warehouse tables. Further, only the relevant data required to answer the Business Questions have been chosen to be loaded into the Data Warehouse. This allows the project to contain clean, consistent and complete data that can be used for reporting purposes.

The transformation rules for this project are designed to directly address specific categories of data quality issues mentioned in the beginning of this section.

Rules to Ensure Referential Integrity

- The data provided by the University of Chicago Booth while comprehensive has missing entries that is disrupting the referential integrity of the data. Specifically, the CCOUNT staging table contains quantitative measures, such as total daily sales of MEAT products. However, the corresponding UPC files, which should detail individual MEAT items, are missing. This absence prevents the factSales table from referencing the Product table, creating gaps in the data structure. In order to maintain referential integrity we have devised the following rules:
    - **Identity Missing Data** - Determine the entities or data that is missing for example in this case Product data for MEAT.
    - **Define and Verify the Placeholder Entities** - Create placeholder records for missing entities and make sure that references to generate the data is from the original and verified data source
    - **Automate Placeholder Creation using Script Transformation** - In the Script Transformation configure the Transformation as a Source and use the CreateNewOutputRows method to generate the placeholder data.
- **Surrogate Key generation** - All dimension tables should be populated with surrogate keys that can be referenced in the fact tables. This is done so as to ensure consistency when data is obtained from different data sources.
- Removed records with invalid STORE numbers (0s, blanks) in CCOUNT staging table to maintain referential integrity with dimStore
- Ensured WEEK numbers in CCOUNT_staging table exist in valid range to maintain integrity with dimTime

Rules to Ensure Format

- **Data Type Conversion** - Data extracted from source were stored in the staging tables as type varchar. These have to be converted to other data types like int, float, date etc respective to what the data originally intents to convey.
- **Date Format** - The format of date data should be maintained as data containing dates should be ensured as “yyyy-mm-dd” consistency across all tables. Additionally, Month and Year data can be extracted from the dates and stored for better reporting.
- **Week Number -** Ensure consistent format for Week numbers (cumulative week numbers from the time of inception instead of week numbers per year)

Rules to Ensure Data Completion

- **Removal of Null Values** - The csv files contained null values for different rows. These should be removed.

- **Removal of Dirty Data** - Some rows contained blank data or "." in different column names in the CCOUNT file. These records should be removed to ensure data completion
- **Verification of Coupon-related Columns:** Verification of coupon-related columns is needed in order to get the aggregation of all coupon-related values accurately.

Rules to Ensure Data Accuracy

- **Week Numbers:** Verify if WEEK numbers correspond to actual weeks
- **Dates:** Ensure DATE values match with corresponding WEEK numbers.

Rules to Ensure Data Validity

- **Business Timeframe:** Check if DATE values are within valid business timeframe
- **Week Numbers:** Validate whether WEEK numbers are within acceptable range i.e. starting from 1 for cumulative week numbers
- **Customer Count:** Verify whether CUSTCOUN values are non-negative.

Rules to Ensure Fit For Purpose

- **Loading only necessary records/columns** - Only data that is relevant to answering the 5 BQs chosen by DFF should be loaded to the dimension tables.
- **Derived Columns** - Derived columns should be generated for both dimension and fact tables to ensure the availability of valuable business data. The SSIS "Derived Column" transformation can be used to achieve this.
    - Time Dimension - Time and Year (stored as int) are derived from the Start Data attribute.
    - Product Dimension - Category (stored as varchar) is derived from the title of the UPC files. In the case of products obtained from CCOUNT file, category was derived from the attribute name.

4.1.6 Plan for Aggregate Tables

Aggregation is required for storing the total coupons redeemed in the Store Information Data Mart. Further, aggregation is also required for all fact tables to ensure that the data is stored in the level of the granularity of the linked dimension tables.

Sales Fact Table

The Movement, Quantity and Sales Amount obtained from the Movement files are stored in the fact table and aggregated on the basis of Store_ID, Time_ID and Product_ID. This ensures that the granularity of the underlying dimension tables is considered. Store_ID is at the store level granularity. Time_ID is at the weekly granularity and Product_ID is at the per product granularity. Since it is stored at a weekly granular level, we can roll up and down to Month and Year at a later point in the project.

Store Information Fact Table

CCOUNT file is the source data for the Store Information Fact Table. Lookup is done on the Time Dimension and store Dimension to store data into the Store Information Fact Table. Thus the data is aggregated here at the Store and Week level. Store_ID is at the store level granularity. Time_ID is at the weekly granularity. Since it is stored at a weekly granular level, we can roll up and down to Month and Year at a later point in the project.

Further, the data for the coupons redeemed are aggregated from the different columns of the CCOUNT table to store into the fact table.

4.1.6 Procedure for Data Extractions and Loadings

The data from the sources (csv files and DFF Data Manual) were first extracted into the staging tables. The cleaning and transformations were performed using the SSIS functions like,

- **Data Conversion** - used to convert one data type to another.
- **Derived Column -** to create new columns valid for BQs. These are created by applying expressions to the original columns.
- **Aggregate** - In cases like calculating the sum of all coupons redeemed, aggregation was required.
- **Lookup** - To create the fact tables and reference them with the dimension tables, the lookup transformation was used.

Through extensive cleaning and transformations, the data quality issues were removed and the records were loaded into the fact and dimension tables accordingly as seen in the next section - Implementation of ETL plan.

4.2 ETL Implementation

4.2.1 Time Dimension

The data in the Time Dimension is populated from the Week Decode source file.

Data Granularity: One row in the time dimension represents one week i.e. one row per week is the granularity of the time dimension identified by Time_ID

ETL of Week Decode Table

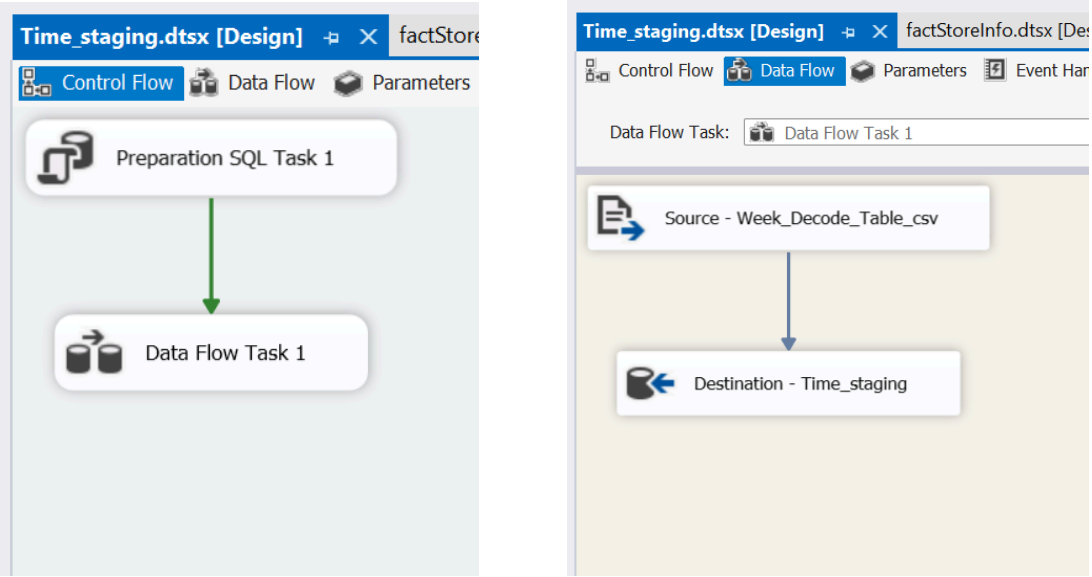


Figure 4.3.1.1 Week Decode Load to Staging Tables

```
SELECT *  FROM [601_Group3_staging-area].[dbo].[Time_staging]
```

100 %

Results | Messages

	Week #	Start	End	Special Events
1	1	09-14-89	09-20-89	
2	2	09-21-89	09-27-89	
3	3	09-28-89	10-04-89	
4	4	10-05-89	10-11-89	
5	5	10-12-89	10-18-89	
6	6	10-19-89	10-25-89	
7	7	10-26-89	11-01-89	Halloween
8	8	11-02-89	11-08-89	
9	9	11-09-89	11-15-89	
10	10	11-16-89	11-22-89	
11	11	11-23-89	11-29-89	Thanksgiving
12	12	11-30-89	12-06-89	
13	13	12-07-89	12-13-89	
14	14	12-14-89	12-20-89	
15	15	12-21-89	12-27-89	Christmas
16				
17	16	12-28-89	01-03-90	New-Year
18	17	01-04-90	01-10-90	
19	18	01-11-90	01-17-90	
20	19	01-18-90	01-24-90	
21	20	01-25-90	01-31-90	
22	21	02-01-90	02-07-90	

Figure 4.3.1.2 Week Decode Data in Staging Table before transformation

Figure 4.3.1.3 Data Flow Transformation to Time Dimension

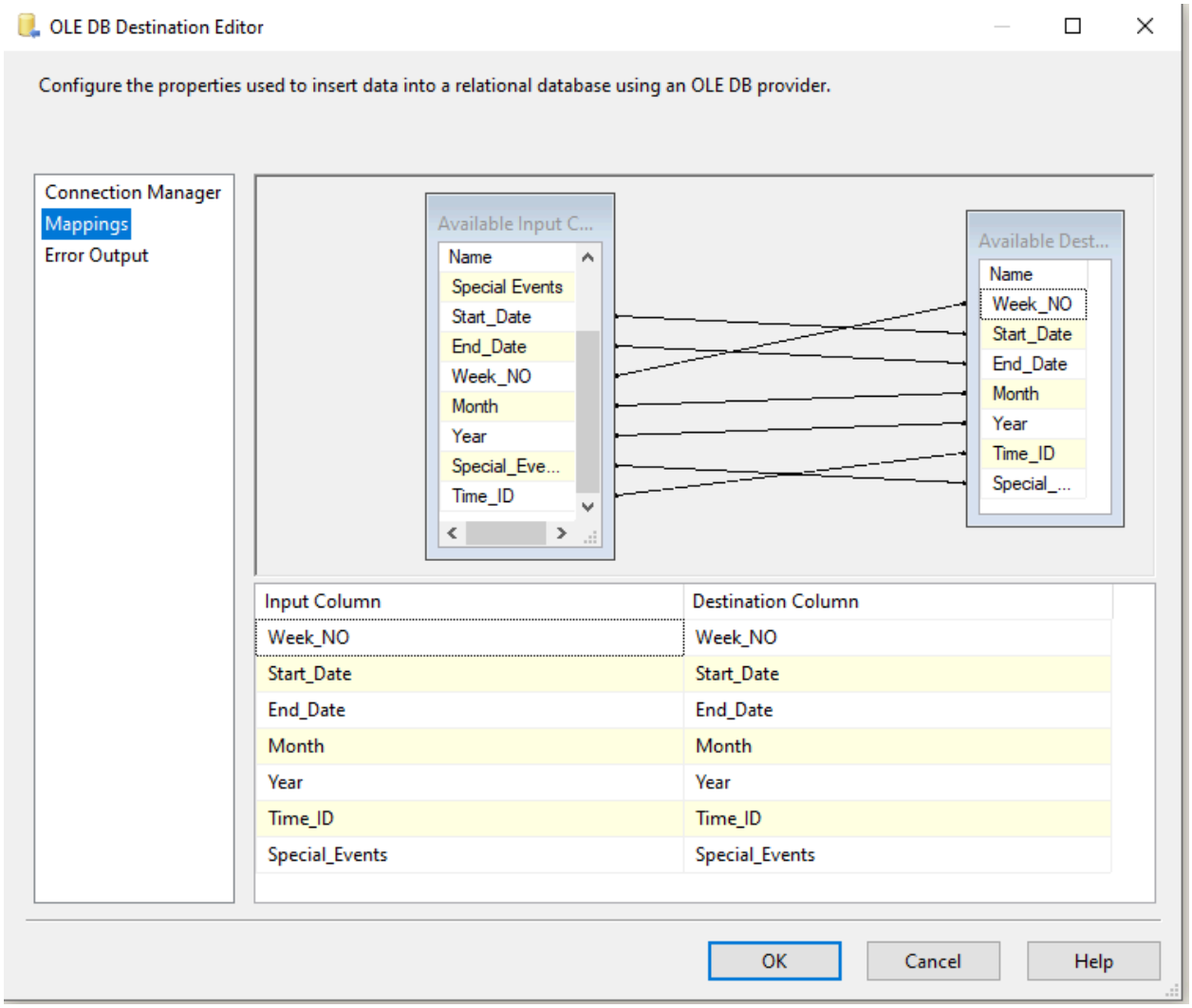


Figure 4.3.1.4 Time Dimension Mappings in Data Flow

Snapshot of Time Dimension

```
SELECT * FROM [601_Group3_SalesDataMart].[dbo].[dimTime]
```

100 %

Results | Messages

	Time_ID	Week_NO	Start_Date	End_Date	Month	Year	Special_Events
1	0	1	1989-09-14	1989-09-20	9	1989	
2	1	2	1989-09-21	1989-09-27	9	1989	
3	2	3	1989-09-28	1989-10-04	9	1989	
4	3	4	1989-10-05	1989-10-11	10	1989	
5	4	5	1989-10-12	1989-10-18	10	1989	
6	5	6	1989-10-19	1989-10-25	10	1989	
7	6	7	1989-10-26	1989-11-01	10	1989	Halloween
8	7	8	1989-11-02	1989-11-08	11	1989	
9	8	9	1989-11-09	1989-11-15	11	1989	
10	9	10	1989-11-16	1989-11-22	11	1989	
11	10	11	1989-11-23	1989-11-29	11	1989	Thanksgiving
12	11	12	1989-11-30	1989-12-06	11	1989	
13	12	13	1989-12-07	1989-12-13	12	1989	
14	13	14	1989-12-14	1989-12-20	12	1989	
15	14	15	1989-12-21	1989-12-27	12	1989	Christmas
16	15	16	1989-12-28	1990-01-03	12	1989	New-Year
17	16	17	1990-01-04	1990-01-10	1	1990	
18	17	18	1990-01-11	1990-01-17	1	1990	
19	18	19	1990-01-18	1990-01-24	1	1990	

Figure 4.3.1.5 Time Dimension Data Snapshot in Sale Data Mart after ETL

```
SELECT * FROM [601_Group3_StoreInformationDataMart].[dbo].[dimTime]
```

100 %

Results | Messages

	Time_ID	Week_NO	Start_Date	End_Date	Month	Year	Special_Events
1	0	1	1989-09-14	1989-09-20	9	1989	
2	1	2	1989-09-21	1989-09-27	9	1989	
3	2	3	1989-09-28	1989-10-04	9	1989	
4	3	4	1989-10-05	1989-10-11	10	1989	
5	4	5	1989-10-12	1989-10-18	10	1989	
6	5	6	1989-10-19	1989-10-25	10	1989	
7	6	7	1989-10-26	1989-11-01	10	1989	Halloween
8	7	8	1989-11-02	1989-11-08	11	1989	
9	8	9	1989-11-09	1989-11-15	11	1989	
10	9	10	1989-11-16	1989-11-22	11	1989	
11	10	11	1989-11-23	1989-11-29	11	1989	Thanksgiving
12	11	12	1989-11-30	1989-12-06	11	1989	
13	12	13	1989-12-07	1989-12-13	12	1989	
14	13	14	1989-12-14	1989-12-20	12	1989	
15	14	15	1989-12-21	1989-12-27	12	1989	Christmas
16	15	16	1989-12-28	1990-01-03	12	1989	New-Year
17	16	17	1990-01-04	1990-01-10	1	1990	
18	17	18	1990-01-11	1990-01-17	1	1990	
19	18	19	1990-01-18	1990-01-24	1	1990	
20	19	20	1990-01-25	1990-01-31	1	1990	

Figure 4.3.1.6 Time Dimension Data Snapshot in Store Information Data Mart after ETL

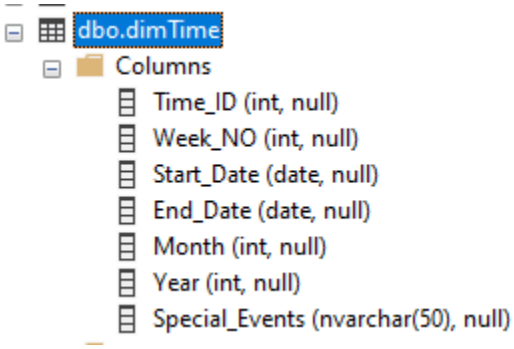


Figure 4.3.1.7 Time Dimension Table Data Types

4.2.2 Product Dimension

The Product Dimension is populated using a combination of the UPCCER and WCER table.

Data Granularity: One row in the product dimension represents one unique product i.e. one row per unique product is the granularity of the product dimension identified by Product_ID.

ETL of UPCCER & WCER file:

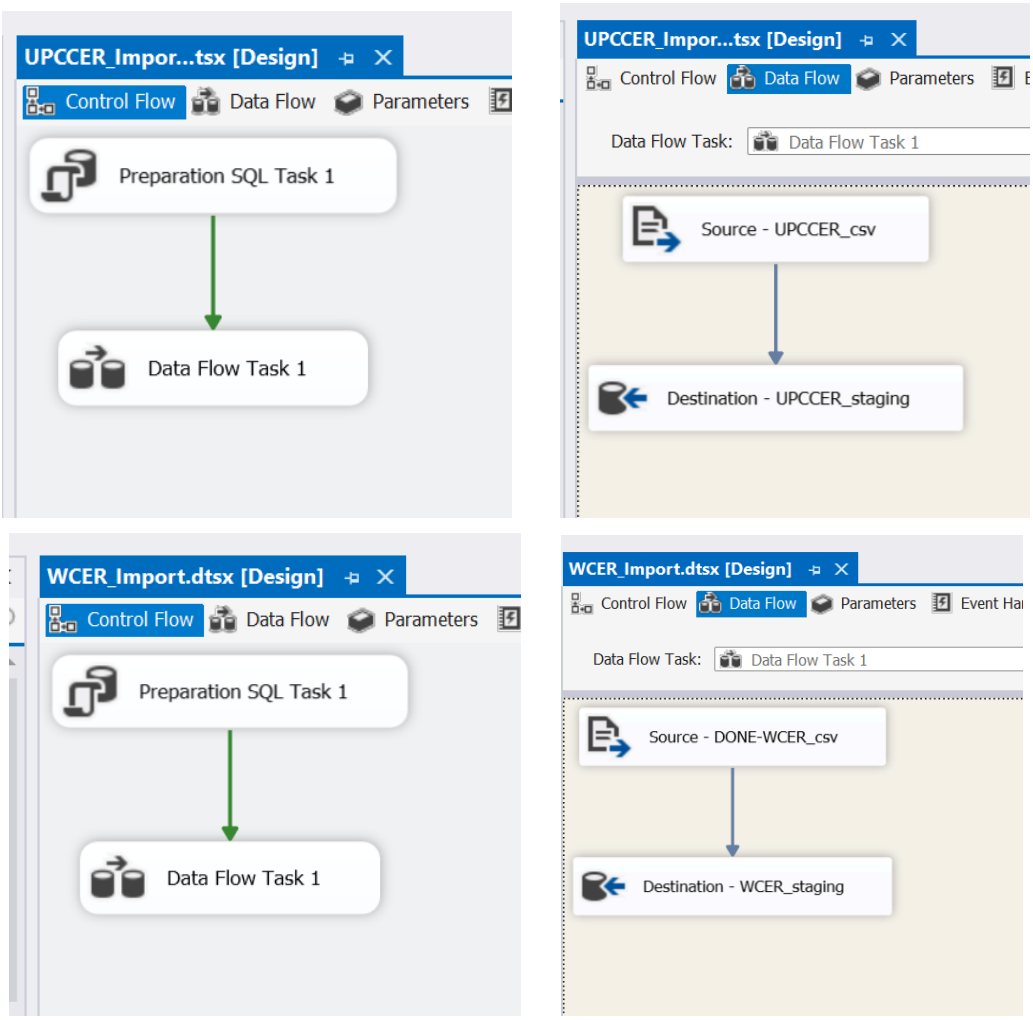
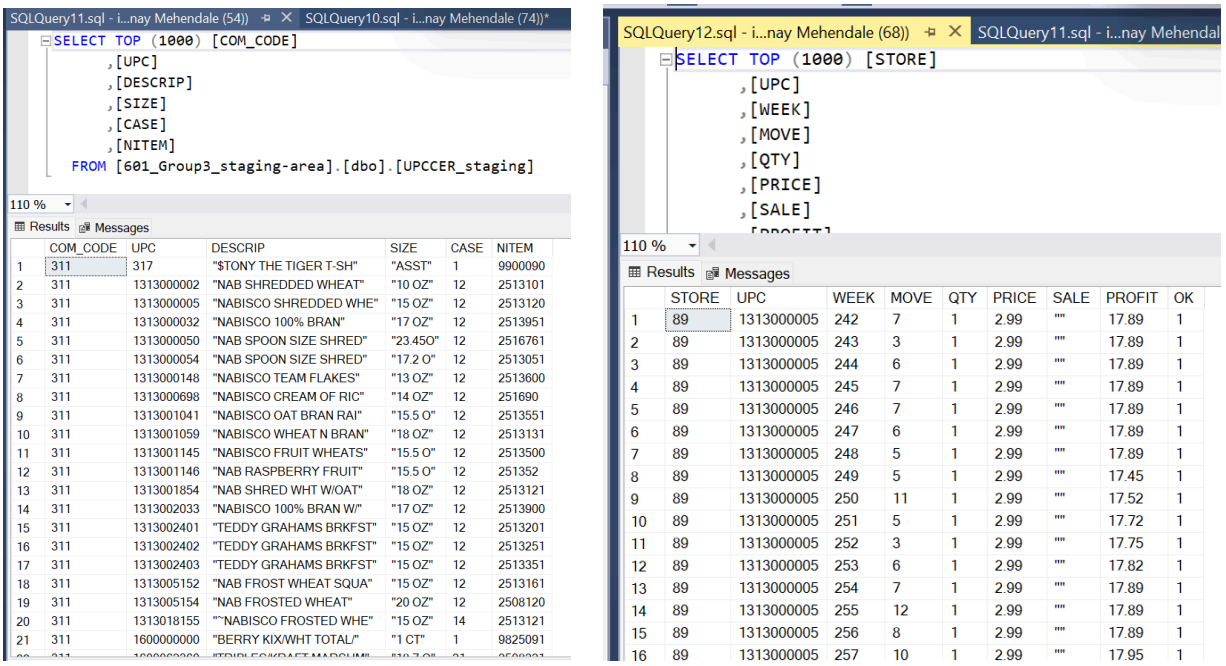


Figure 4.3.2.1 Loading UPCCER, WCER files to Staging Tables



SQLQuery11.sql - i...nay Mehendale (54)) | SQLQuery10.sql - i...nay Mehendale (74))*

```
SELECT TOP (1000) [COM_CODE]
      ,[UPC]
      ,[DESCRIP]
      ,[SIZE]
      ,[CASE]
      ,[NITEM]
  FROM [601_Group3_staging-area].[dbo].[UPCCER_staging]
```

110 % | Results | Messages

	COM_CODE	UPC	DESCRIP	SIZE	CASE	NITEM
1	311	317	"$TONY THE TIGER T-SH"	"ASST"	1	9900090
2	311	1313000002	"NAB SHREDDED WHEAT"	"10 OZ"	12	2513101
3	311	1313000005	"NABISCO SHREDDED WHE"	"15 OZ"	12	2513120
4	311	1313000032	"NABISCO 100% BRAN"	"17 OZ"	12	2513951
5	311	1313000050	"NAB SPOON SIZE SHRED"	"23.45O"	12	2516761
6	311	1313000054	"NAB SPOON SIZE SHRED"	"17.2 O"	12	2513051
7	311	1313000148	"NABISCO TEAM FLAKES"	"13 OZ"	12	2513600
8	311	1313000698	"NABISCO CREAM OF RIC"	"14 OZ"	12	251690
9	311	1313001041	"NABISCO OAT BRAN RAI"	"15.5 O"	12	2513551
10	311	1313001059	"NABISCO WHEAT N BRAN"	"18 OZ"	12	2513131
11	311	1313001145	"NABISCO FRUIT WHEATS"	"15.5 O"	12	2513500
12	311	1313001146	"NAB RASPBERRY FRUIT"	"15.5 O"	12	251352
13	311	1313001854	"NAB SHRED WHT W/OAT"	"18 OZ"	12	2513121
14	311	1313002033	"NABISCO 100% BRAN W/"	"17 OZ"	12	2513900
15	311	1313002401	"TEDDY GRAHAMS BRKFST"	"15 OZ"	12	2513201
16	311	1313002402	"TEDDY GRAHAMS BRKFST"	"15 OZ"	12	2513251
17	311	1313002403	"TEDDY GRAHAMS BRKFST"	"15 OZ"	12	2513351
18	311	1313005152	"NAB FROST WHEAT SQUA"	"15 OZ"	12	2513161
19	311	1313005154	"NAB FROSTED WHEAT"	"20 OZ"	12	2508120
20	311	1313018155	"~NABISCO FROSTED WHE"	"15 OZ"	14	2513121
21	311	1600000000	"BERRY KIX/WHT TOTAL/"	"1 CT"	1	9825091

SQLQuery12.sql - i...nay Mehendale (68)) | SQLQuery11.sql - i...nay Mehendal

```
SELECT TOP (1000) [STORE]
      ,[UPC]
      ,[WEEK]
      ,[MOVE]
      ,[QTY]
      ,[PRICE]
      ,[SALE]
      ,[PROFIT]
```

110 % | Results | Messages

	STORE	UPC	WEEK	MOVE	QTY	PRICE	SALE	PROFIT	OK
1	89	1313000005	242	7	1	2.99	""	17.89	1
2	89	1313000005	243	3	1	2.99	""	17.89	1
3	89	1313000005	244	6	1	2.99	""	17.89	1
4	89	1313000005	245	7	1	2.99	""	17.89	1
5	89	1313000005	246	7	1	2.99	""	17.89	1
6	89	1313000005	247	6	1	2.99	""	17.89	1
7	89	1313000005	248	5	1	2.99	""	17.89	1
8	89	1313000005	249	5	1	2.99	""	17.45	1
9	89	1313000005	250	11	1	2.99	""	17.52	1
10	89	1313000005	251	5	1	2.99	""	17.72	1
11	89	1313000005	252	3	1	2.99	""	17.75	1
12	89	1313000005	253	6	1	2.99	""	17.82	1
13	89	1313000005	254	7	1	2.99	""	17.89	1
14	89	1313000005	255	12	1	2.99	""	17.89	1
15	89	1313000005	256	8	1	2.99	""	17.89	1
16	89	1313000005	257	10	1	2.99	""	17.95	1

Figure 4.3.2.2 Data in UPCCER, WCER Staging Table, CCOUNT before transformation

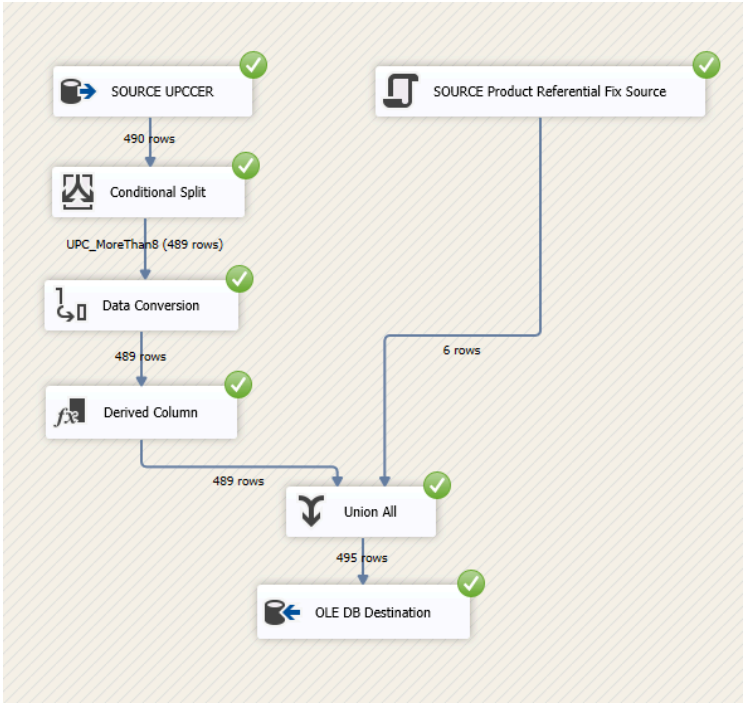


Figure 4.3.2.3 Data Flow Transformation to Product Dimension

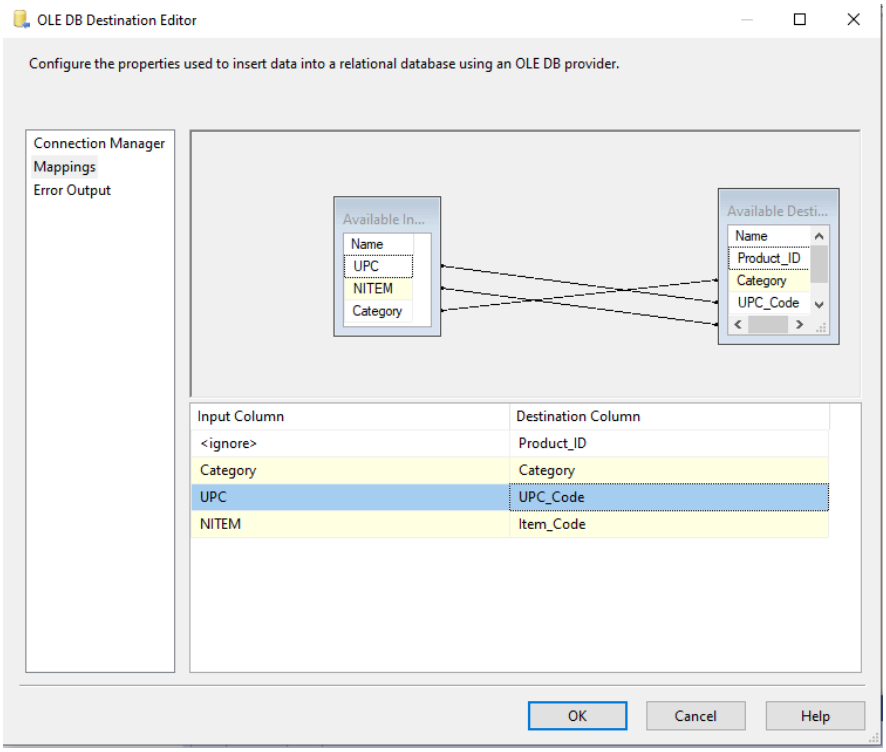
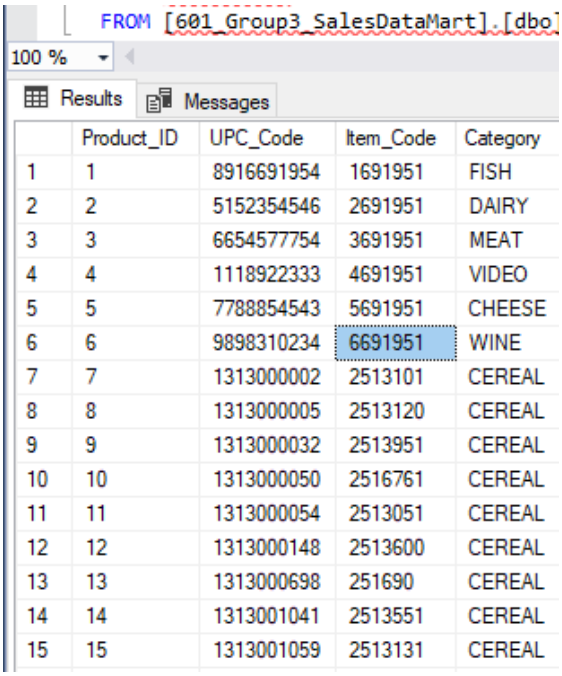


Figure 4.3.2.4 Product Dimension Mappings in Data Flow



	Product_ID	UPC_Code	Item_Code	Category
1	1	8916691954	1691951	FISH
2	2	5152354546	2691951	DAIRY
3	3	6654577754	3691951	MEAT
4	4	1118922333	4691951	VIDEO
5	5	7788854543	5691951	CHEESE
6	6	9898310234	6691951	WINE
7	7	1313000002	2513101	CEREAL
8	8	1313000005	2513120	CEREAL
9	9	1313000032	2513951	CEREAL
10	10	1313000050	2516761	CEREAL
11	11	1313000054	2513051	CEREAL
12	12	1313000148	2513600	CEREAL
13	13	1313000698	251690	CEREAL
14	14	1313001041	2513551	CEREAL
15	15	1313001059	2513131	CEREAL

Figure 4.3.2.5 Product Dimension Data Snapshot in Sale Data Mart after ETL

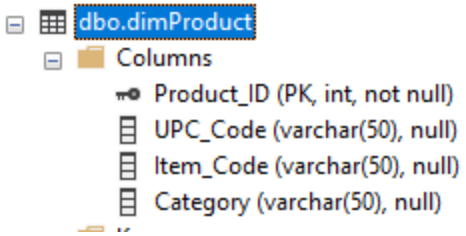


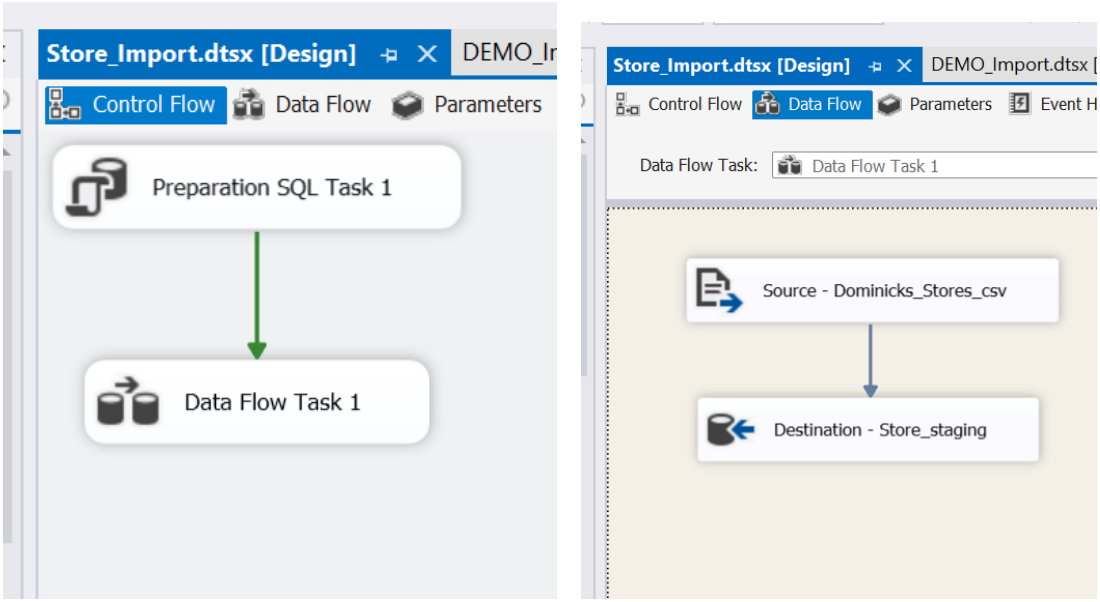
Figure 4.3.2.6 Product Dimension Table Data Types

4.2.3 Store Dimension

The Store Dimension table is populated from the Dominicks_Stores.csv flat file and DEMO.csv flat files

Data Granularity: One row in the store dimension represents one unique store i.e. one row per unique store is the granularity of the product dimension identified by Store_ID.

ETL of Dominicks_Stores & DEMO file:



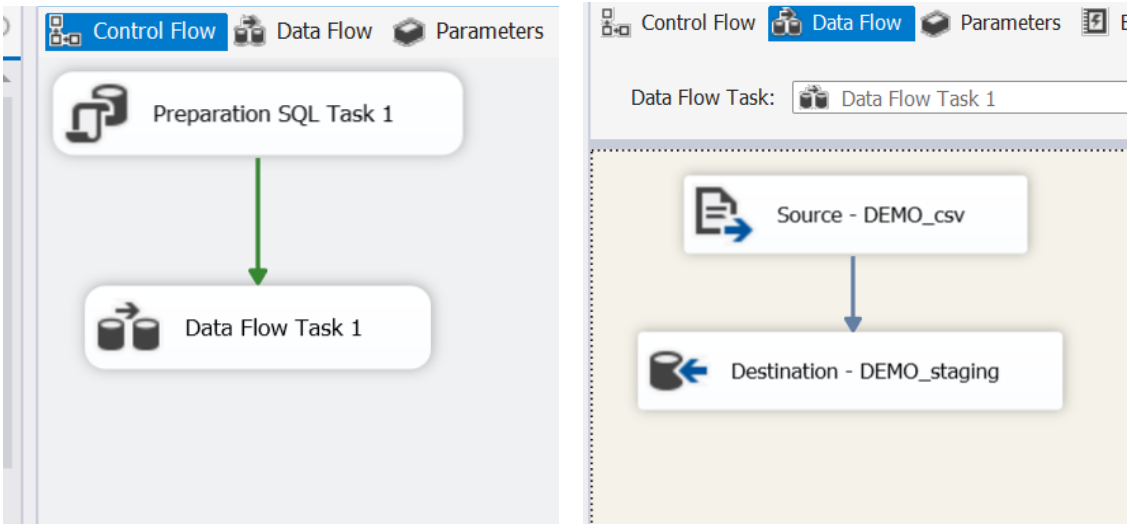


Figure 4.3.3.1 Loading Store and Demographic files to Staging Tables

SQLQuery13.sql - i...nay Mehendale (54)) | dummy.sql - infoda...ay Mehendale (58))

```
SELECT TOP (1000) [Store]
      ,[City]
      ,[Price Tier]
      ,[Zone]
      ,[Zip Code]
      ,[Address]
  FROM [601_Group3_staging-area].[dbo].[Store_staging]
```

110 %

Results | Messages

	Store	City	Price Tier	Zone	Zip Code	Address
1	2	River Forest	High	1	60305	7501 W. North Ave.
2	4	Park Ridge	Medium	2	60068	Closed
3	5	Palatine	Medium	2	60067	223 Northwest HWY.
4	8	Oak Lawn	Low	5	60435	8700 S. Cicero Ave.
5	9	Morton Grove	Medium	2	60053	6931 Dempster
6	12	Chicago	High	7	60660	6009 N. Broadway Ave.
7	14	Glenview	High	1	60025	1020 Waukegan Rd.
8	18	River Grove	Low	5	60171	8355 W. Belmont Ave.
9	19	Glen Ellyn			60137	Closed
10	21	Hanover Park	CubFighter	6	60103	1440 Irving Park Rd.
11	25	Chicago			60639	Closed
12	28	Mt. Prospect	Medium	2	60054	"1145-55 Mt Prospect Pz."
13	32	Park Ridge	High	1	60068	"1900 S. Cumberland Ave."
14	33	Chicago	High	7	60657	3012 N. Broadway Ave.
15	39	Waukegan			60085	Closed
16	40	Bridgeview	CubFighter	6	60455	8825 S. Harlem
17	44	Western Sp...	Medium	2	60558	14 Garden Market St.
18	45	Wheeling	Medium	2	60090	550 W. Dundee Rd.

SQLQuery14.sql - i...nay Mehendale (68)) | SQLQuery13.sql - i...nay Mehendale (54))

```
SELECT TOP (1000) [MMID]
      ,[CITY]
      ,[ZIP]
      ,[STORE]
      ,[SCLUSTER]
      ,[ZONE]
      ,[AGE9]
  FROM [601_Group3_staging-area].[dbo].[DEMO_staging]
```

110 %

Results | Messages

	MMID	CITY	ZIP	STORE	SCLUSTER	ZONE	AGE9
1	16892	"RIVER FOREST"	60305	2	"C"	1	0.117508576
2	16893	"PARK RIDGE"	60068	4	"A"	2	0.0950895057
3	16894	"PALATINE"	60067	5	"D"	2	0.1414334827
4	16895	"OAK LAWN"	60453	8	"C"	5	0.123155416
5	16896	"MORTON GROVE"	60053	9	"A"	2	0.1035030974
6	16898	"CHICAGO"	60660	12	"B"	7	0.1056967397
7	16899	"GLENVIEW"	60025	14	"A"	1	0.129589372
8	16901	"RIVER GROVE"	60171	18	"A"	5	0.1100949839
9	.	""	.	19	""	.	.
10	16903	"HANOVER PARK"	60103	21	"D"	6	0.1759263459
11	.	""	.	25	""	.	.
12	16905	"MOUNT PROSPECT"	60056	28	"A"	2	0.1288795371
13	16906	"PARK RIDGE"	60068	32	"C"	1	0.0990606319
14	16907	"CHICAGO"	60657	33	"B"	7	0.0460709172
15	.	""	.	39	""	.	.
16	16909	"BRIDGEVIEW"	60455	40	"D"	6	0.1336846485
17	16912	"WESTERN SPRIN	60558	44	"A"	2	0.1448834853

Figure 4.3.3.2 Data in Stores Staging table, Demo staging table before transformation

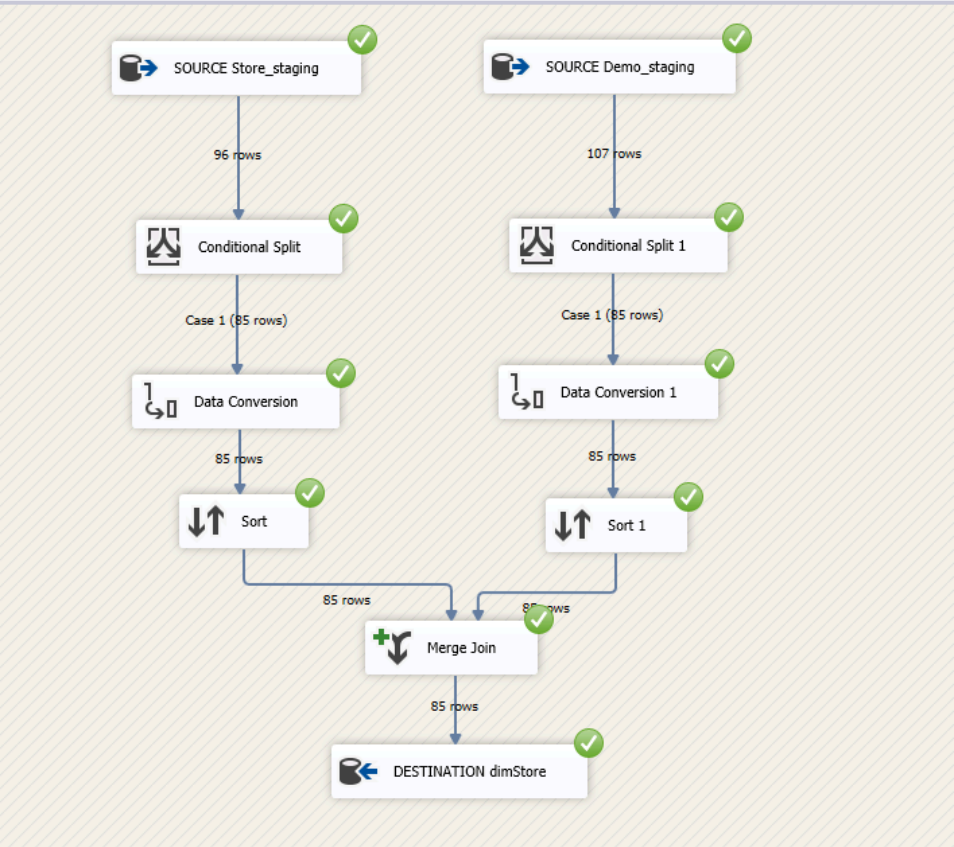


Figure 4.3.3.3 Data Flow Transformation to Store Dimension for Sales Data Mart

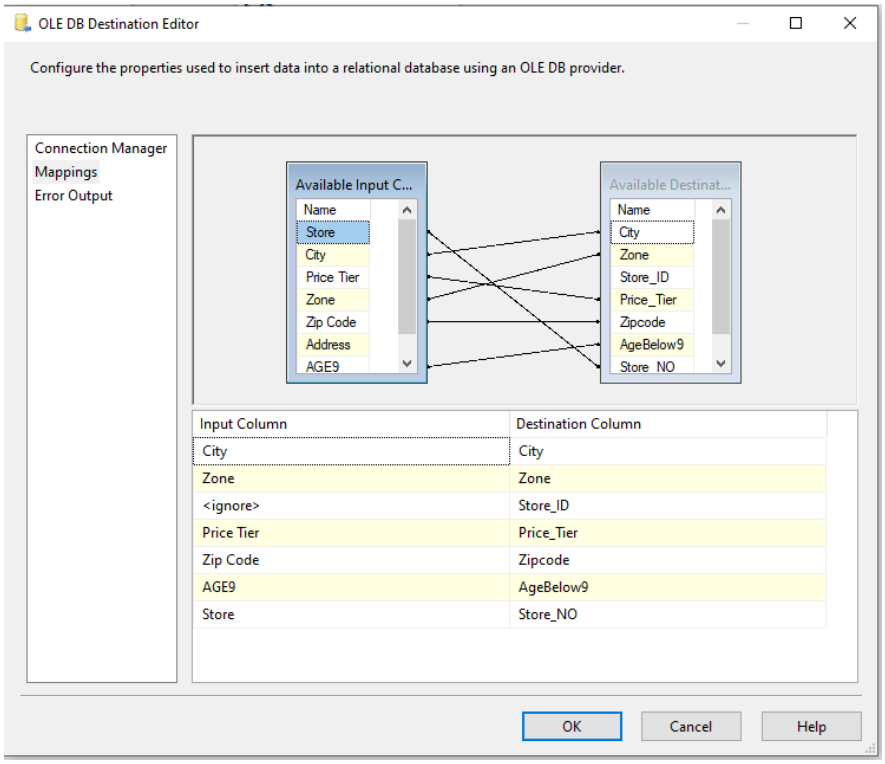


Figure 4.3.3.4 Store Dimension Mappings in Data Flow for Sales Data Mart

```
      ,[AgeBelow9]
  FROM [601_Group3_SalesDataMart].[dbo].[dimStore]
```

100 %

Results | Messages

	Store_ID	Store_NO	City	Price_Tier	Zone	Zipcode	AgeBelow9
1	1	100	Chicago	High	11	60698	0.1843609762
2	2	101	Des Plaines	Medium	12	60016	0.1169249346
3	3	102	Merrionette Park	Low	15	0655	0.1469459202
4	4	103	Bolingbrook	Low	15	60439	0.1853939837
5	5	104	St. Charles	High	8	60174	0.1629361422
6	6	105	Melrose Park	Medium	12	60160	0.1475648774
7	7	106	Montgomery	High	8	60538	0.1874031576
8	8	107	Westchester	Medium	2	60153	0.119089317
9	9	109	Bannockburn	High	7	60015	0.1475198042
10	10	110	East Dundee	Medium	2	60118	0.1755940555
11	11	111	Chicago	High	1	60620	0.1458742213
12	12	112	Buffalo Grove	Low	14	60090	0.1636118598
13	13	113	Chicago	Medium	2	60646	0.1030849308
14	14	114	Calumet City	Medium	12	60409	0.1468197601
15	15	115	Naperville	Medium	12	60540	0.1854935109
16	16	116	Elmhurst	Medium	2	60126	0.1398772238
17	17	117	Schaumburg	Medium	2	60193	0.1390937165
18	18	118	Morton Grove	Medium	10	60053	0.1037918216
19	19	119	Buffalo Grove	Medium	2	60090	0.1456695805

Figure 4.3.3.5 Store Dimension Data Snapshot in Sale Data Mart after ETL

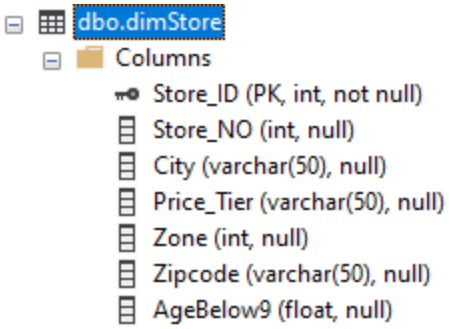


Figure 4.3.3.6 Product Dimension Table Data Types for Sales Data Mart

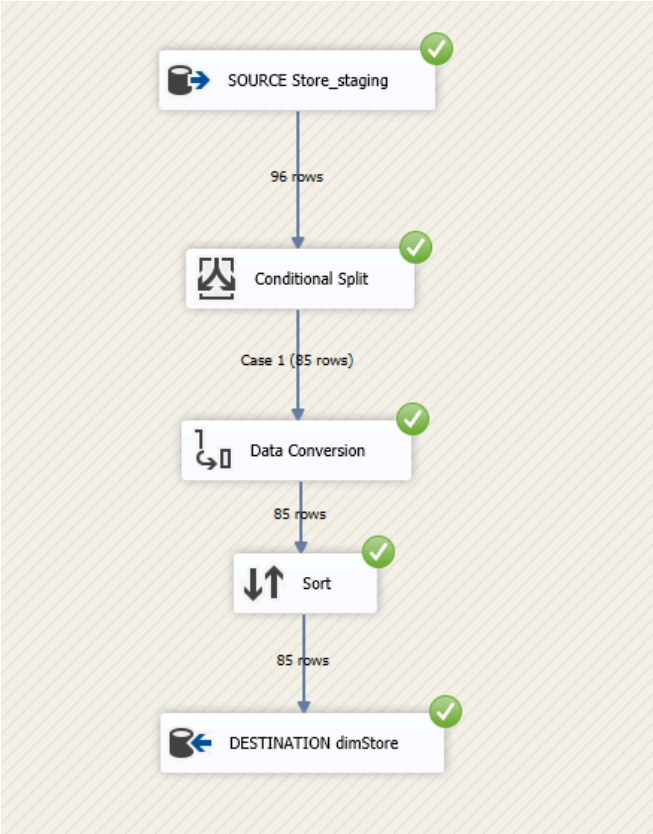


Figure 4.3.3.7 Data Flow Transformation to Store Dimension for Store Information Data Mart

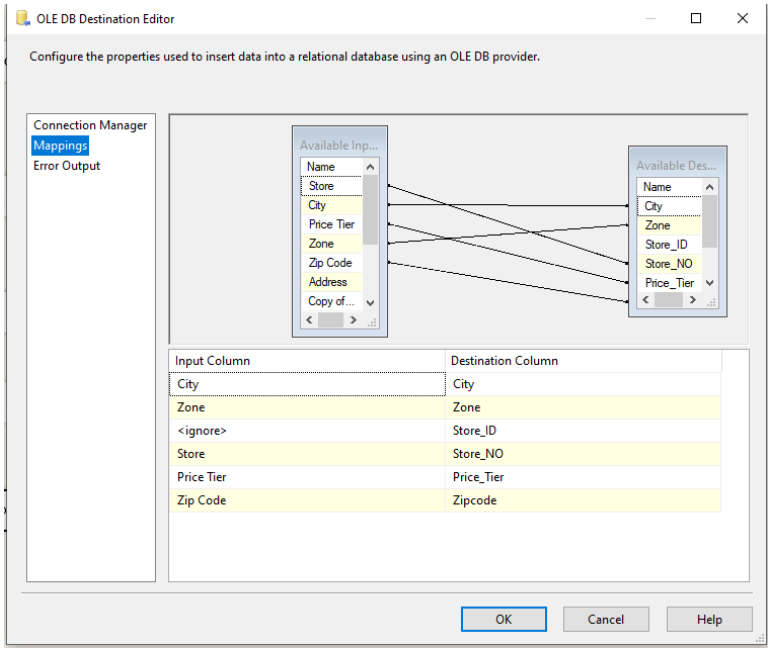


Figure 4.3.3.8 Store Dimension Mappings in Data Flow for Store Information Data Mart

```
FROM [601_Group3_StoreInformationDataMart].[dbo].[di
```

100 %

Results | Messages

	Store_ID	Store_NO	City	Price_Tier	Zone	Zipcode
1	1	100	Chicago	High	11	60698
2	2	101	Des Plaines	Medium	12	60016
3	3	102	Merrionette Park	Low	15	0655
4	4	103	Bolingbrook	Low	15	60439
5	5	104	St. Charles	High	8	60174
6	6	105	Melrose Park	Medium	12	60160
7	7	106	Montgomery	High	8	60538
8	8	107	Westchester	Medium	2	60153
9	9	109	Bannockburn	High	7	60015
10	10	110	East Dundee	Medium	2	60118
11	11	111	Chicago	High	1	60620
12	12	112	Buffalo Grove	Low	14	60090
13	13	113	Chicago	Medium	2	60646
14	14	114	Calumet City	Medium	12	60409
15	15	115	Naperville	Medium	12	60540
16	16	116	Elmhurst	Medium	2	60126
17	17	117	Schaumburg	Medium	2	60193
18	18	118	Morton Grove	Medium	10	60053
19	19	119	Buffalo Grove	Medium	2	60090

Figure 4.3.3.9 Store Dimension Data Snapshot in Store Information Data Mart after ETL

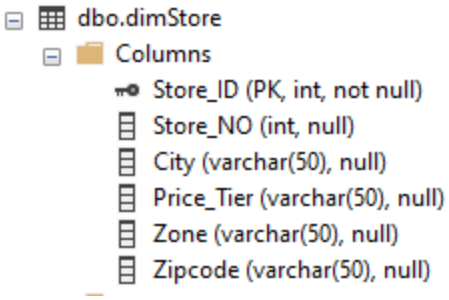


Figure 4.3.3.10 Product Dimension Table Data Types for Store Information Data Mart

4.2.4 Cleaning of CCOUNT_staging

To get accurate data into the Sales Fact table and Store Information fact table, extensive transformations were performed on the CCOUNT_staging table. Snapshots of these detailed transformations are given as follows,

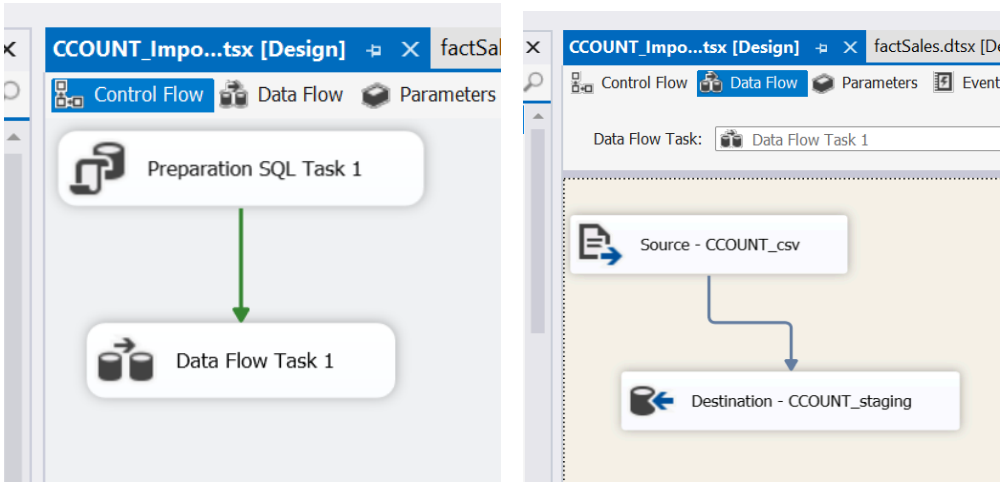


Figure 4.3.4.1 CCOUNT.csv Load to Staging Tables

Results Messages

	STORE	DATE	GROCERY	DAIRY	FROZEN	BOTTLE	MVPCLUB	GROCCOUP	MEAT	MEATFROZ	MEATCOUP	FISH	FISHCOUP	PROMO	PROMCOUP	PRODUCE	BUL
617	0	"/□%"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
618	0	"/'<"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
619	0	"/_"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
620	0	"/¢@"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
621	0	"/²Í"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
622	0	"/»*"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
623	0	"930...	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
624	0	"930...	45603	11896	8592	0	0	0	13806	0	0	710	0	0	0	11153	140
625	0	"930...	11452	2781	2127	0	0	0	3717	0	0	0	0	0	0	1955	160
626	0	"'-"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
627	0	"'Í"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
628	0	"'«□"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
629	0	"'¸□"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
630	0	"a□"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
631	0	"a□"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
632	0	"aa@"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
633	0	"ap"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
634	0	"a£@"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
635	0	"a£Œ"	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Figure 4.3.4.2 Data in CCOUNT_staging table before transformation

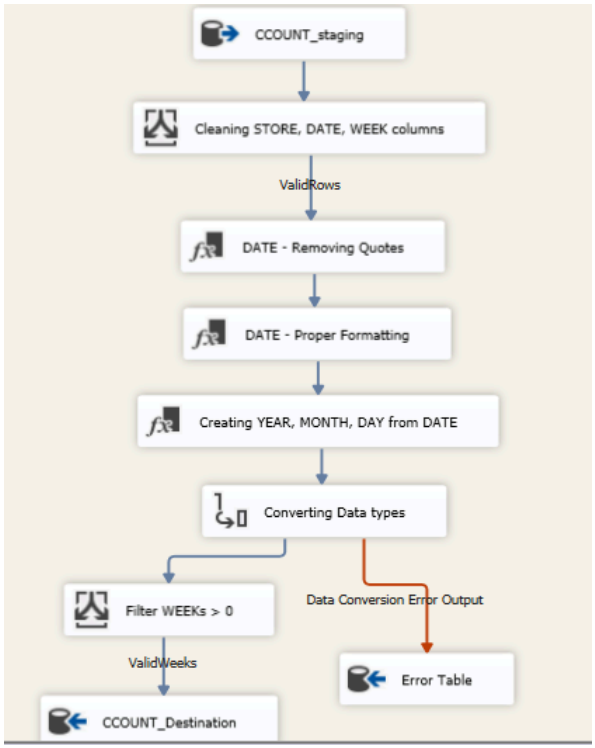
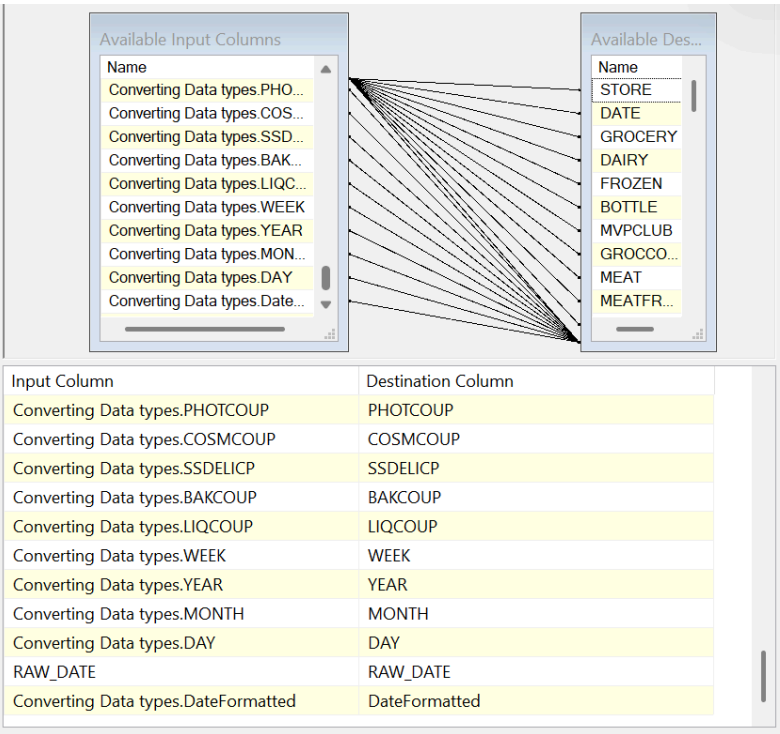


Figure 4.3.4.3 Data Flow Transformation



Input Column	Destination Column
Converting Data types.PHOTCOUP	PHOTCOUP
Converting Data types.COSMCOUP	COSMCOUP
Converting Data types.SSDELICP	SSDELICP
Converting Data types.BAKCOUP	BAKCOUP
Converting Data types.LIQCOUP	LIQCOUP
Converting Data types.WEEK	WEEK
Converting Data types.YEAR	YEAR
Converting Data types.MONTH	MONTH
Converting Data types.DAY	DAY
RAW_DATE	RAW_DATE
Converting Data types.DateFormatted	DateFormatted

Figure 4.3.4.4 CCOUNT (cleaned) table mappings in Data Flow

Results Messages

	STORE	RAW_DATE	DateFormatted	DATE	YEAR	MONTH	WEEK	DAY	CUSTCOUN	GROCERY	DAIRY	FROZEN	BOTTLE	MVPCLUB	GROCCOUP	MEAT	ME
1	100	"920421"	19920421	1992-04-21	1992	4	136	21	2992	17146.22	3978.52	2444.72	-0.80	0.00	-330.38	3295.45	34
2	100	"920422"	19920422	1992-04-22	1992	4	136	22	3158	16722.56	4074.01	2506.14	-12.90	0.00	-329.67	4121.89	47
3	100	"920423"	19920423	1992-04-23	1992	4	137	23	3158	18861.10	3955.89	2815.58	-3.20	0.00	-388.44	4698.59	44
4	100	"920424"	19920424	1992-04-24	1992	4	137	24	3488	22032.43	4366.69	3298.52	-4.80	0.00	-389.98	5627.32	71
5	100	"920425"	19920425	1992-04-25	1992	4	137	25	4119	29910.82	6299.30	4517.76	-8.40	0.00	-359.52	8636.05	79
6	100	"920426"	19920426	1992-04-26	1992	4	137	26	3689	28090.96	5701.28	3969.41	-5.10	0.00	-277.62	5963.31	73
7	100	"920427"	19920427	1992-04-27	1992	4	137	27	3115	18738.81	3948.29	2965.83	-3.30	0.00	-165.48	3904.32	60
8	100	"920428"	19920428	1992-04-28	1992	4	137	28	2937	15110.58	3076.53	2108.13	-0.80	0.00	-195.89	3776.64	34
9	100	"920429"	19920429	1992-04-29	1992	4	137	29	2923	15186.97	3117.95	2289.19	-1.60	0.00	-132.35	3584.27	36
10	100	"920430"	19920430	1992-04-30	1992	4	138	30	3270	20126.08	3689.69	2857.45	-6.40	0.00	-177.07	8125.67	38
11	100	"920501"	19920501	1992-05-01	1992	5	138	1	3575	21176.48	4276.28	3384.31	-33.60	0.00	-204.17	8356.40	52
12	100	"920502"	19920502	1992-05-02	1992	5	138	2	4416	34519.72	6764.99	5255.69	-10.00	0.00	-309.29	13778.50	10
13	100	"920503"	19920503	1992-05-03	1992	5	138	3	3756	27833.80	5938.79	4279.78	-0.90	0.00	-196.90	7796.50	85
14	100	"920504"	19920504	1992-05-04	1992	5	138	4	3257	20140.17	4162.61	3116.62	-4.80	0.00	-120.97	6945.51	41
15	100	"920505"	19920505	1992-05-05	1992	5	138	5	3272	16961.15	3694.96	2614.56	0.00	0.00	-127.90	6722.30	44
16	100	"920506"	19920506	1992-05-06	1992	5	138	6	3301	17093.87	3597.30	2713.00	-3.32	0.00	-144.91	6896.85	43
17	100	"920507"	19920507	1992-05-07	1992	5	139	7	3463	18901.67	3791.62	2868.71	-1.50	0.00	-157.23	5719.75	38
18	100	"920508"	19920508	1992-05-08	1992	5	139	8	3967	22157.57	4744.50	3408.85	-0.80	0.00	-201.19	7068.24	51

Query executed successfully. | infodata16.mbs.tamu.edu (13... | TANAY-PC\Tanay Mehenda... | 601_Group3_staging-area | 00:00:00 | 1,000 rows

Figure 4.3.4.5 Data Snapshot of CCOUNT table after ETL

Figure 4.3.4.6 New CCOUNT data Types after ETL

4.2.5 Sales Fact Table

The Sales Fact table (factSales) is populated primarily by the movement staging table (WCER_staging), looking up three dimension tables (dimStore, dimTime and dimProduct) and calculating the total sales value accordingly. There are also products for which the movement data is not available (i.e. there is no WXXX file corresponding to them). Such products are termed as 'Special Products' and are handled separately than the normal Data Flow.

Data Granularity: One row in the Sales fact table represents one store per week per product i.e. One row per Store-Time-Product combination is the granularity of fact sales table.

ETL Flow:

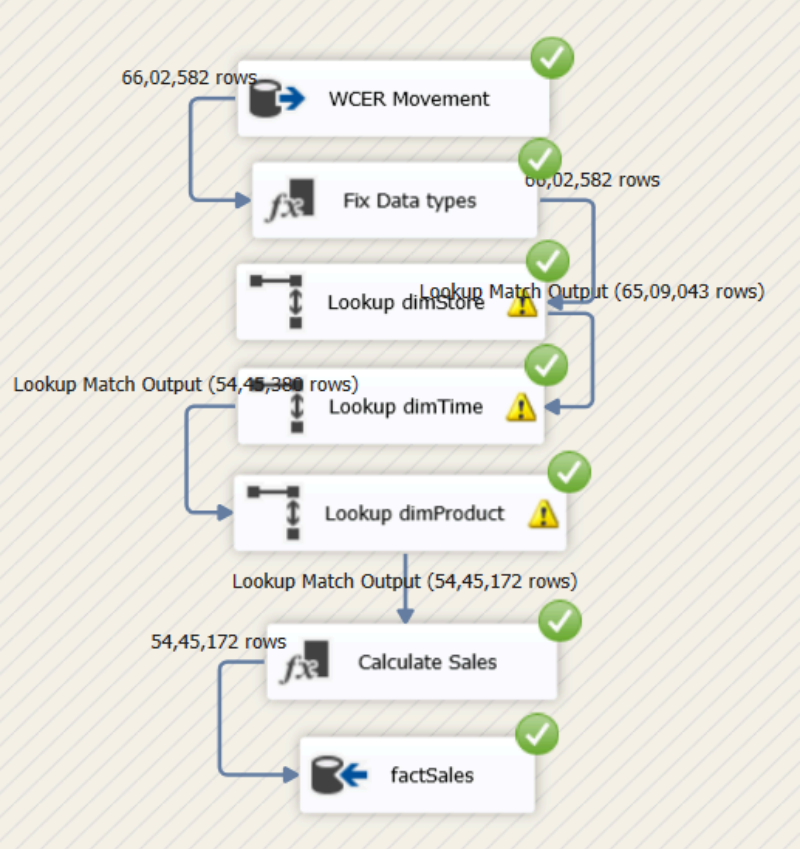
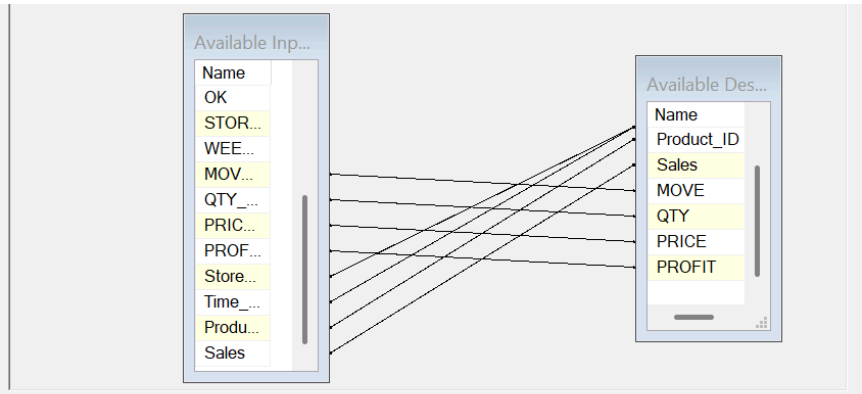


Figure 4.3.5.1 Data Flow Transformation to Sales fact table



Input Column	Destination Column
Store_ID	Store_ID
Time_ID	Time_ID
Product_ID	Product_ID
Sales	Sales
MOVE_Conv	MOVE
QTY_Conv	QTY
PRICE_Conv	PRICE
PROFIT_Conv	PROFIT

Figure 4.3.5.2 Fact Sales Mappings in Data Flow

Results | Messages

	Store_ID	Time_ID	Product_ID	MOVE	QTY	PRICE	PROFIT	Sales
1	81	262	69	19	1	3.39	19.91	64.41
2	81	263	69	14	1	3.39	19.91	47.46
3	81	264	69	15	1	3.39	19.91	50.85
4	81	265	69	14	1	3.39	19.91	47.46
5	81	266	69	11	1	3.39	19.91	37.29
6	81	267	69	14	1	3.39	19.91	47.46
7	81	268	69	16	1	3.39	19.88	54.24
8	81	269	69	20	1	3.39	19.45	67.80
9	81	270	69	22	1	3.39	19.37	74.58
10	81	271	69	12	1	3.39	19.35	40.68
11	81	272	69	8	1	3.39	19.32	27.12
12	81	273	69	12	1	3.39	19.32	40.68
13	81	274	69	5	1	3.39	19.32	16.95
14	81	275	69	5	1	3.39	19.32	16.95
15	81	276	69	10	1	3.39	19.32	33.90
16	81	277	69	9	1	3.39	19.32	30.51
17	81	278	69	7	1	3.39	19.32	23.73
18	81	279	69	8	1	3.39	19.32	27.12
19	81	280	69	15	1	3.39	19.32	50.85
20	81	281	69	9	1	3.39	19.32	30.51
21	81	282	69	15	1	3.39	19.32	50.85

Figure 4.3.5.3 Data Snapshot of Fact Sales table after ETL

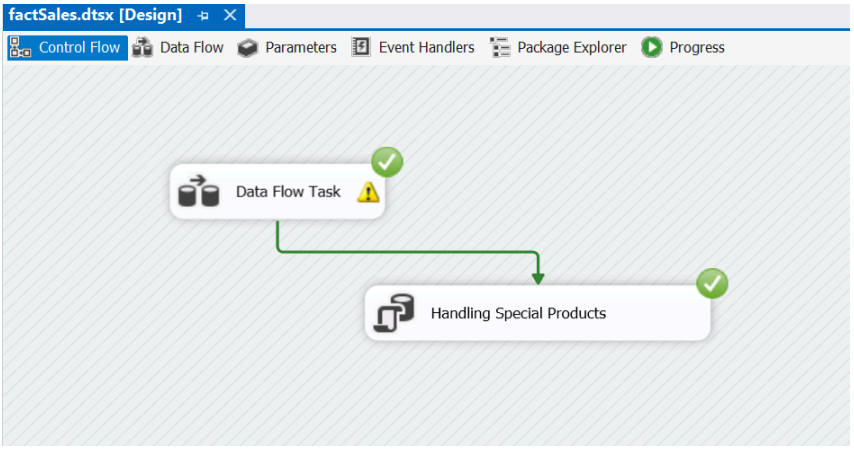


Figure 4.3.5.4 Control Flow - Handling special products

Results | Messages

	Store_ID	Time_ID	Product_ID	MOVE	QTY	PRICE	PROFIT	Sales
265	1	1	441	0	1	0.00	0.00	0.00
266	1	1	79	60	1	3.69	16.91	221.40
267	1	1	70	20	1	2.57	36.58	51.40
268	1	1	62	6	1	2.90	21.90	17.40
269	1	1	1	NULL	NULL	NULL	NULL	773.21
270	1	1	2	NULL	NULL	NULL	NULL	5748.27
271	1	1	3	NULL	NULL	NULL	NULL	10524.41
272	1	1	4	NULL	NULL	NULL	NULL	235.65
273	1	1	5	NULL	NULL	NULL	NULL	289.47
274	1	1	6	NULL	NULL	NULL	NULL	238.39
275	1	1	1	NULL	NULL	NULL	NULL	1330.84
276	1	1	2	NULL	NULL	NULL	NULL	6169.24
277	1	1	3	NULL	NULL	NULL	NULL	11845.97
278	1	1	4	NULL	NULL	NULL	NULL	481.46
279	1	1	5	NULL	NULL	NULL	NULL	526.74
280	1	1	6	NULL	NULL	NULL	NULL	391.17

Figure 4.3.5.5 Data Snapshot of Fact Sales table including special products

Note: Records with NULL values for MOVE, QTY, PRICE and PROFIT indicate that these are special products i.e. they do not have any movement data but their Sales are directly fetched from the CCOUNT.csv file.

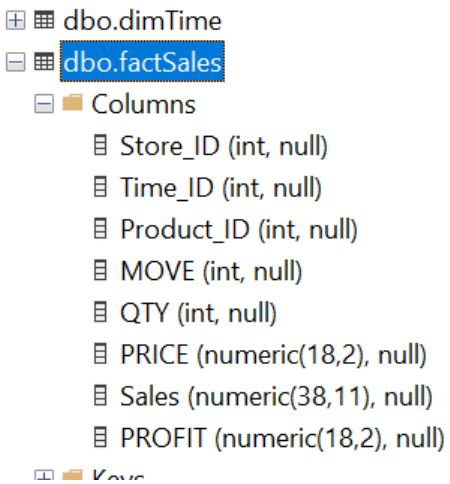


Figure 4.3.5.6 Fact Sales table new data Types

4.2.6 Store Information Fact Table

The Store Information Fact table (factStoreInfo) is populated primarily by looking up two dimension tables (dimStore and dimTime), calculating the CouponRedeemed values, as well as getting CustomerCount from the cleaned CCOUNT table.

Data Granularity: One row in the Store Information fact table represents one store per week i.e. One row per Store-Time combination is the granularity of the store information fact table.

ETL Flow:

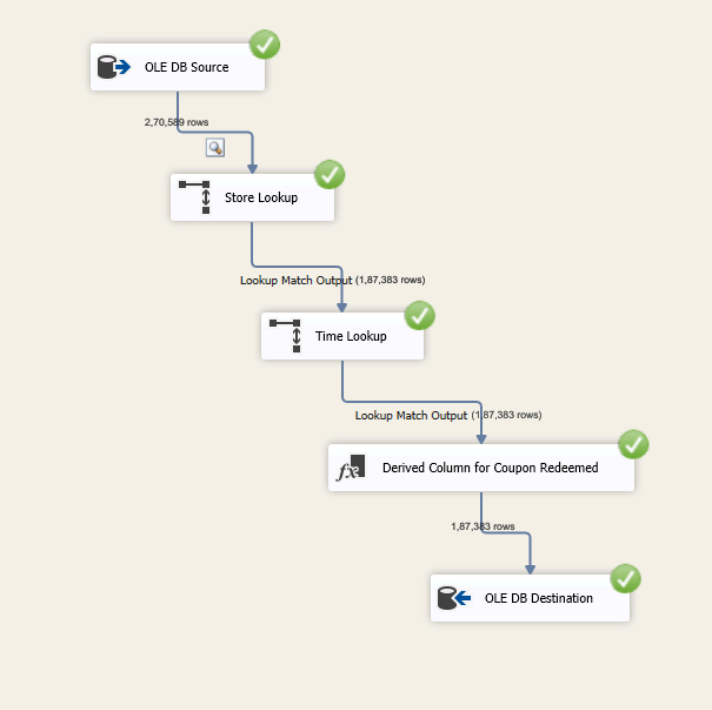


Figure 4.3.6.1 Data Flow Transformation to StoreInfo fact table

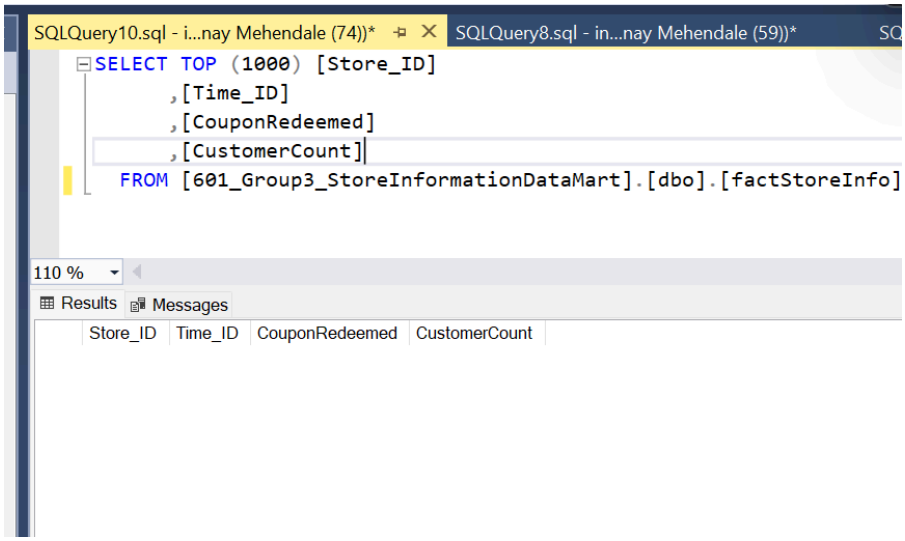
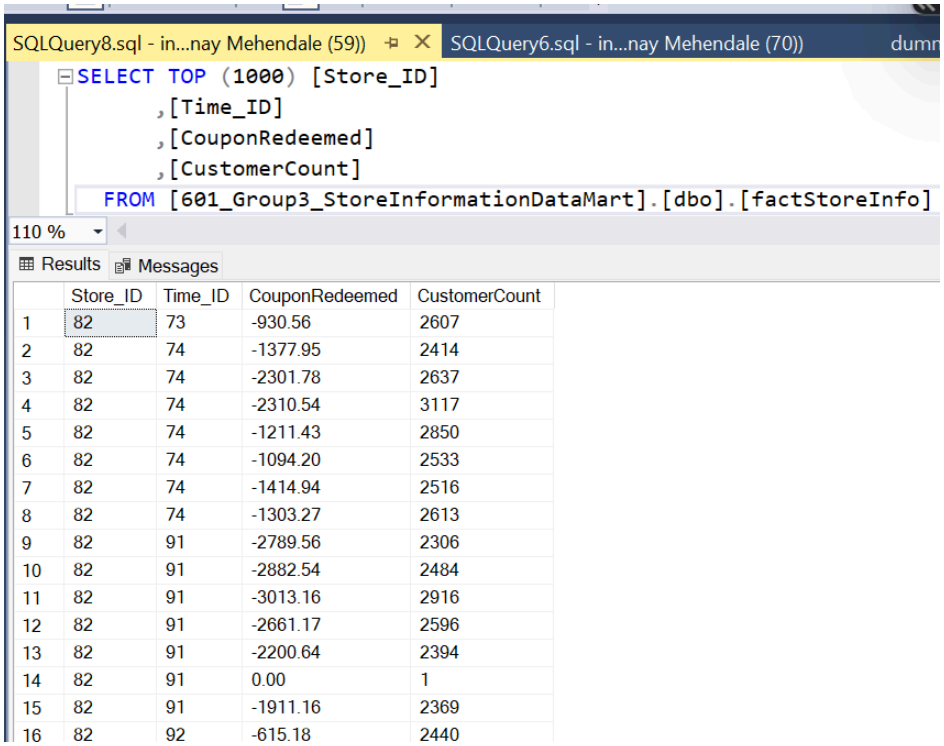


Figure 4.3.6.2 Data Snapshot of Store Info fact table before transformations



	Store_ID	Time_ID	CouponRedeemed	CustomerCount
1	82	73	-930.56	2607
2	82	74	-1377.95	2414
3	82	74	-2301.78	2637
4	82	74	-2310.54	3117
5	82	74	-1211.43	2850
6	82	74	-1094.20	2533
7	82	74	-1414.94	2516
8	82	74	-1303.27	2613
9	82	91	-2789.56	2306
10	82	91	-2882.54	2484
11	82	91	-3013.16	2916
12	82	91	-2661.17	2596
13	82	91	-2200.64	2394
14	82	91	0.00	1
15	82	91	-1911.16	2369
16	82	92	-615.18	2440

Figure 4.3.6.3 Data Snapshot of Fact StoreInfo table after ETL

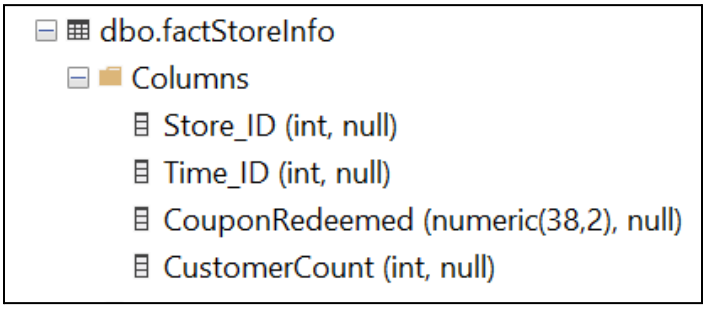


Figure 4.3.6.4 Data types of the Fact StoreInfo table

5. BI Reporting

5.1 Reporting Plan

For the BI reporting and analysis of the 5 chosen Business Questions, we have chosen the following tools to create the reports.

- **SSRS** - This is used for tabular and graphical reports that are directly linked to the data warehouse.
- **SSAS** - This is used for cube-based reporting and flexible analysis using OLAP.
- **Tableau** - This gives us interactive visual dashboards which can be created either directly on data marts or on cubes.
- **SSAS + Tableau** - This can be done for more advanced visualizations that are built on top of analytical cubes.

The reporting strategy used for each BQ is carefully designed to optimize and ensure high accuracy for decision making. Following are the tools used in this report and the BQs solved using each of these tools.

Reporting Tool	Business Question
SSRS	BQ1
SSAS	BQ4
Tableau	BQ2, BQ5
SSAS + Tableau	BQ6

5.1.1 Target Reports for BQs

BQ1 What is the trend of wine sales during the holiday seasons - Christmas, Thanksgiving and New Year? How does the cheese sales trend during the same period and is there a correlation between wine and cheese?

Report generated using only SSRS.

SSRS was chosen as the reporting tool for this BQ because SSRS is excellent for structured reports involving timeframes like holidays. It also has a good correlation analysis which is required for this particular business question. Using SSRS Charts we used line graphs to depict the fluctuations in sales of “Cheese” and “Wine” products of a particular year and compare them to each other. To create the SSRS report we used the dimTime, dimProduct, factSales tables which provided the Sales data, Product Category (for Wine and Cheese), the Week_NO and the Year. Since we were interested in seeing the sales trends during the holidays we used the Week_No as the x_axis and the Sales as the y_axis, and used the Year as a parameterized field which gave us the ability to generate a line chart for the sales of each year within the report. Therefore we could easily compare and see the trends of sales of the Wine and Cheese products of each year.

BQ2 Find the top 5 most profitable products in the Cereal category in 1994 in the store that is most popular to kids?

Report generated using Tableau.

Tableau was chosen because of its ability to provide interactive reports and advanced filtering, since we needed to find popular stores with kids as well as then use those stores as a filter to get Top 5 most profitable cereal products. To find most popular stores with kids, AgeBelow9 attribute of dimStore was used, which was converted to percentages by creating a calculated field (Age9Percent) for better interpretation. Once, we found popular stores with kids, we then used Store NO as a filter since we only wanted these stores. We also used Product ID as a filter to match with IDs of 'Cereals' (which were all except 1 to 7). We then chose a treemap since it was aligned with our business question. For better interpretation, we then listed the Product ID, Category, Item Code and Profit on each of the elements of treemap.

BQ4 What is the trend in sales of products like video and meat from the year 1989 to 1994?

Report generated using SSAS cube.

We leveraged SSAS Cube for this business question because of its ability to efficiently handle large historical datasets, along with complex hierarchies and better time-series analysis abilities. For this we used dimStore, dimProduct, dimTime and factSales, basically the entire Sales Data Mart to come up with the solution. We also created a Named Query 'YearlySalesByCategory' which had 'Year' from dimTime, 'Category' from dimProduct and 'Sales' from factSales. We created this query since we needed to specifically aggregate the Sales at year level instead of week level (how it is stored in the data mart)

BQ5 What are the weekly sale trends for perishable goods like dairy and fish per store in the year 1990?

Report generated using Tableau.

We utilized Tableau for this business question for its interactive and rich annotation features needed for detailed weekly analysis and trend comparison between categories. We utilized 'Week NO' from dimTime, Product ID from dimProduct and 'Sales' from factSales to answer this business question. A calculated field 'ActualWeek1990' was created from 'Week NO' to better understand the trend of sales during the actual weeks in 19990 instead of the cumulative week numbers from the inception of the business.

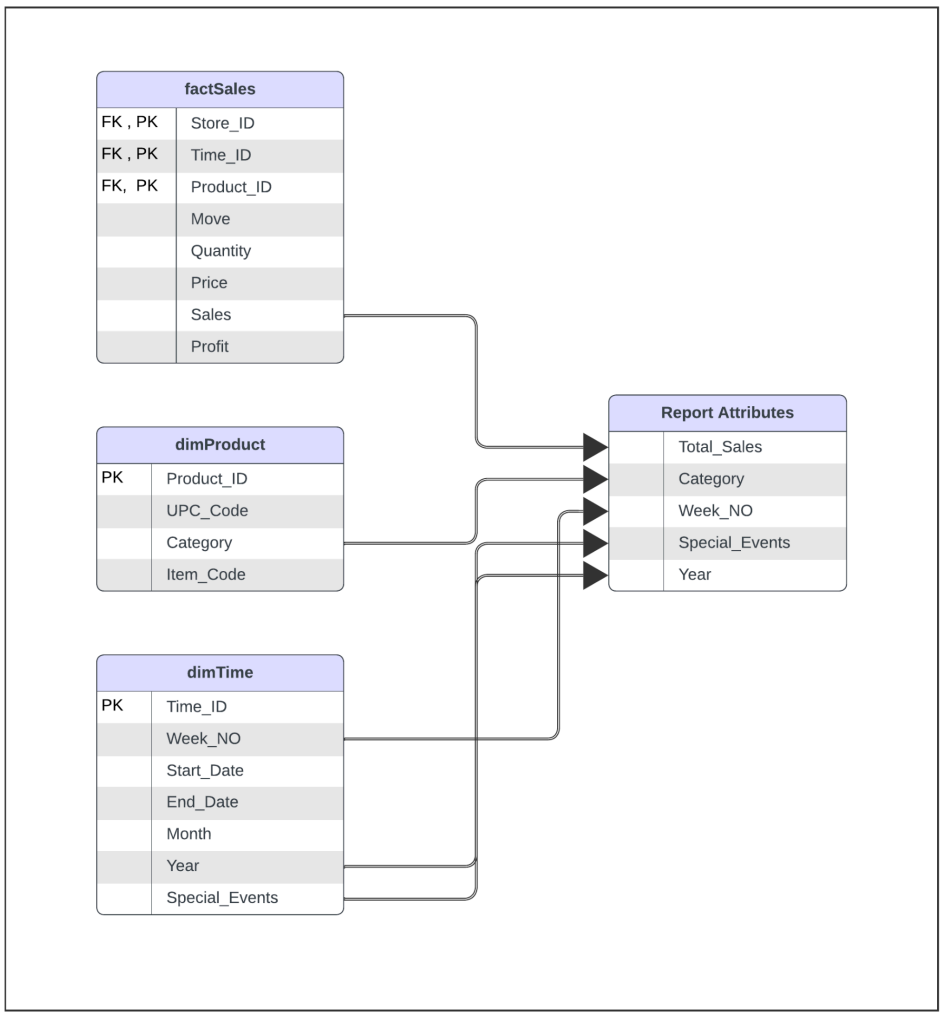
BQ6 What are the weekly trend lines of total number of customers in a store and the total number of coupons redeemed? Is there a correlation between the two?

Report generated using Tableau on top of SSAS cube.

We combined Tableau with SSAS to leverage both efficient data processing of large datasets through SSAS and Tableau's dual-axis charts and correlation analysis, which provided an accurate graph for comparison and pattern identification. For this business question, we used dimTime, dimStore and factStoreInfo tables from the Store Information Data Mart. We used Week NO field to analyze the trend at the week level, used 'Year' as a filter to look at a specific year (e.g. 1990). 'CouponRedeemed' from factStoreInfo fact table had an unusually high number of negative values, which represented the change in the actual sales for that week. To look at the correlation between the number of customers visiting the store and the usage of coupons, a calculated field was added which converted the CouponRedeemed values to its absolute value (i.e. AbsoluteCouponRedeemedValue). We then utilized this calculated field to visualize the graph and get the weekly correlation.

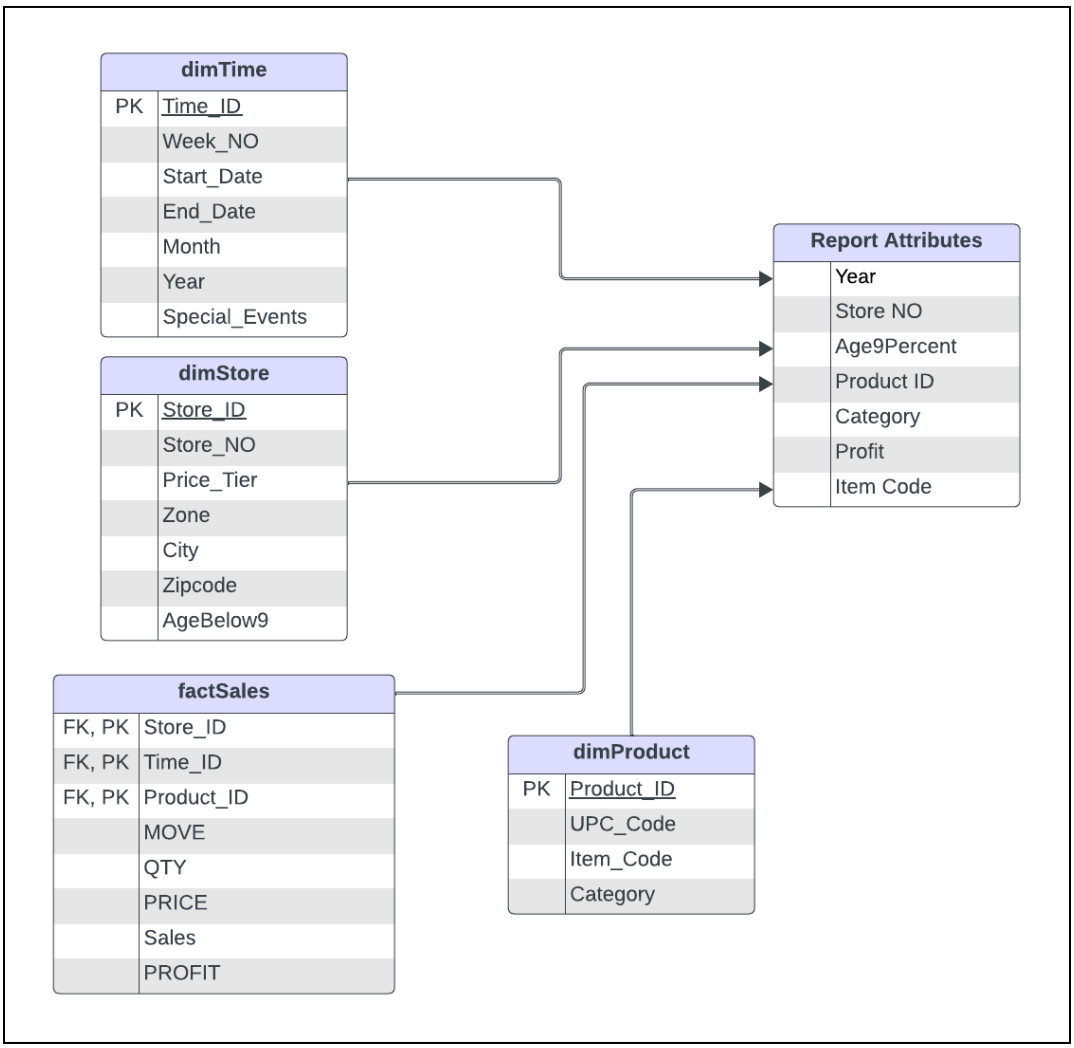
5.1.2 Mappings from Data Marts to Report Attributes

BQ1 What is the trend of wine sales during the holiday seasons - Christmas, Thanksgiving and New Year? How does the cheese sales trend during the same period and is there a correlation between wine and cheese?



Attribute Name	Dimension/Fact Table	Groupings/Filters	Report Attribute
Year	dimTime	Parameterized Filtering according to the Year attribute	Year
Sales	factSales		Total_Sales
Category	dimProduct	Filter by ‘WINE’ and ‘CHEESE’	Category
Week_NO	dimTime		Week_NO
Special_Events	dimTime		Special_Events

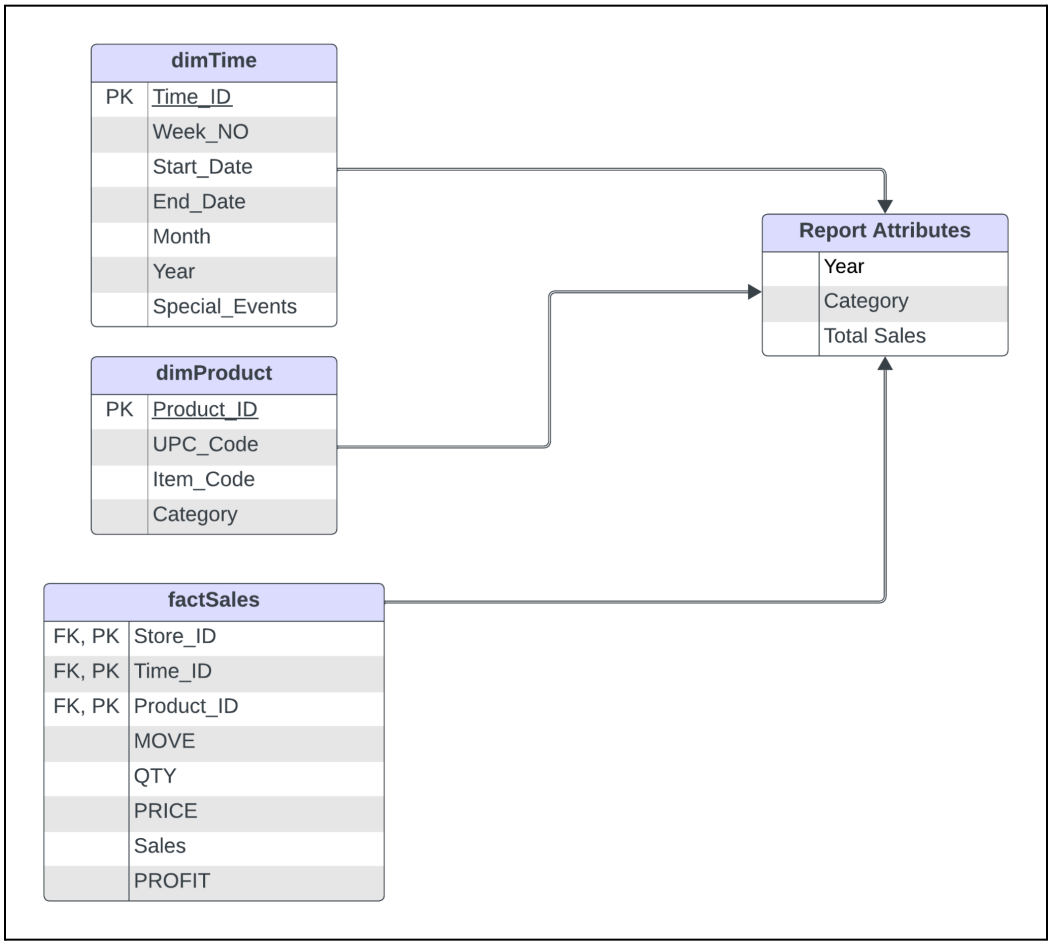
BQ2 Find the top 5 most profitable products in the Cereal category in 1994 in the store that is most popular to kids?



Attribute Name	Dimension/Fact Table	Groupings/Filters	Report Attribute
Year	dimTime	Filter on 1994	Year
Store_NO	dimStore	Top 10 Stores	Store NO
AgeBelow9	dimStore	Convert to percentage (multiply by 100)	Age9Percent
Product_ID	dimProduct	Filter on ‘Cereals’ Top 5 ‘Cereals’	Product ID

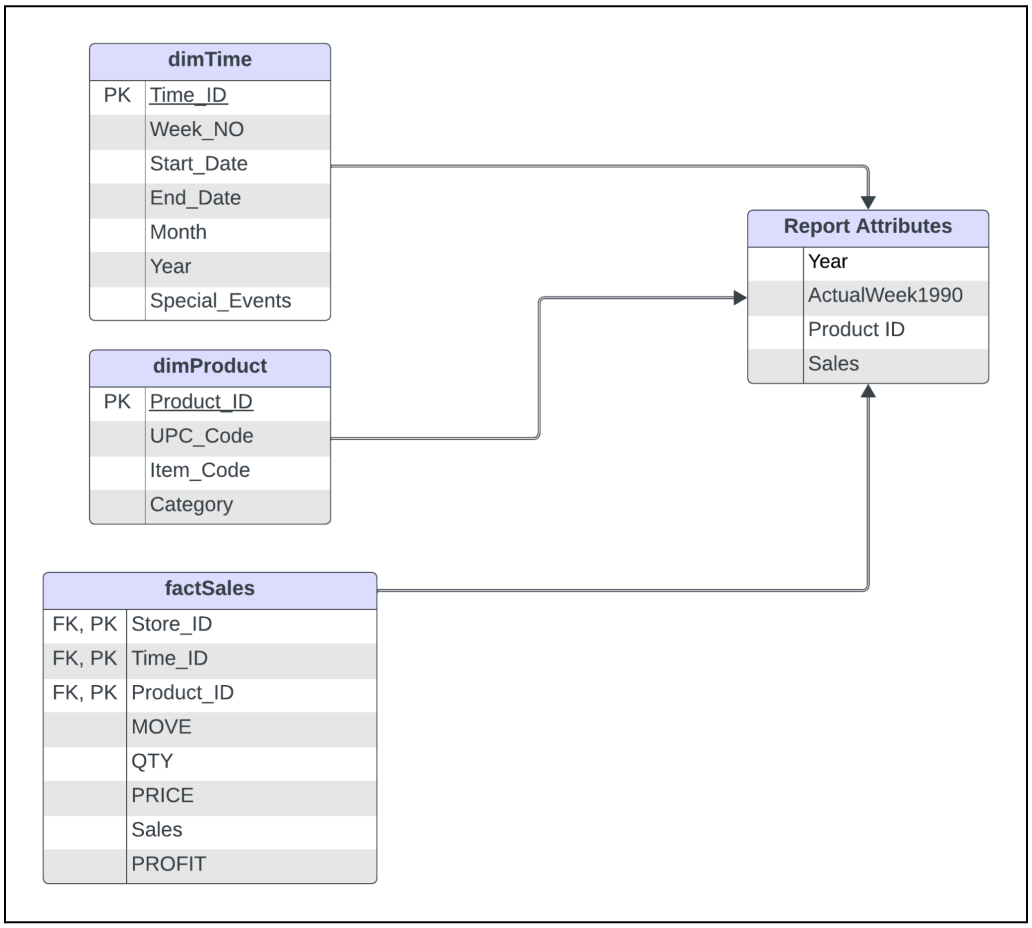
Category	dimProduct		Category
Profit	factSales		Profit
Item_Code	dimProduct		Item Code

BQ4 What is the trend in sales of products like video and meat from the year 1989 to 1994?



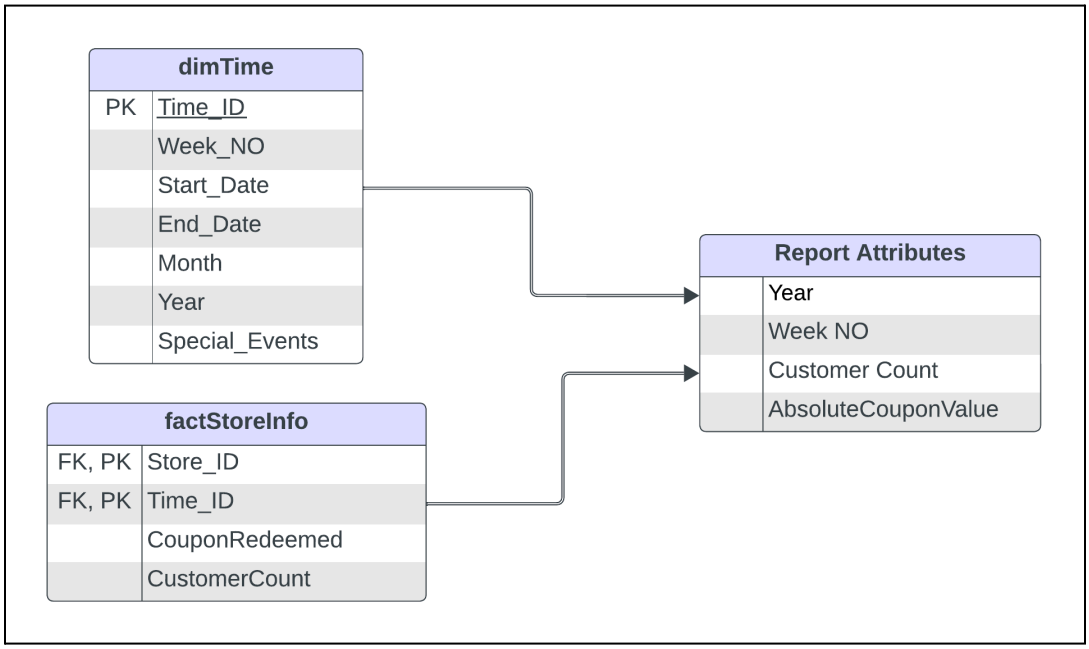
Attribute Name	Dimension/Fact Table	Groupings/Filters	Report Attribute
Year	dimTime	Filter between 1989 and 1994	Year
Category	dimProduct	Filter on ‘Video’ and ‘Meat’	Category
Sales	factSales	Group By Year and Category	Total Sales

BQ5 What are the weekly sale trends for perishable goods like dairy and fish per store in the year 1990?



Attribute Name	Dimension/Fact Table	Groupings/Filters	Report Attribute
Year	dimTime	Filter on 1990	Year
Week NO	dimTime	Calculate Week number for 1990	ActualWeek1990
Product_ID	dimProduct	Filter on ‘Dairy’ and ‘Fish’	Product ID
Sales	factSales		Sales

BQ6 What are the weekly trend lines of total number of customers in a store and the total number of coupons redeemed? Is there a correlation between the two?



Attribute Name	Dimension/Fact Table	Groupings/Filters	Report Attribute
Year	dimTime	Filter on 1990	Year
Week NO	dimTime		Week NO
CustomerCount	factStoreInfo		Customer Count
CouponRedeemed	factStoreInfo	Calculate absolute value of Coupons Redeemed	AbsoluteCouponValue

5.2 Report Implementation

5.2.1 SSRS Report for BQ1

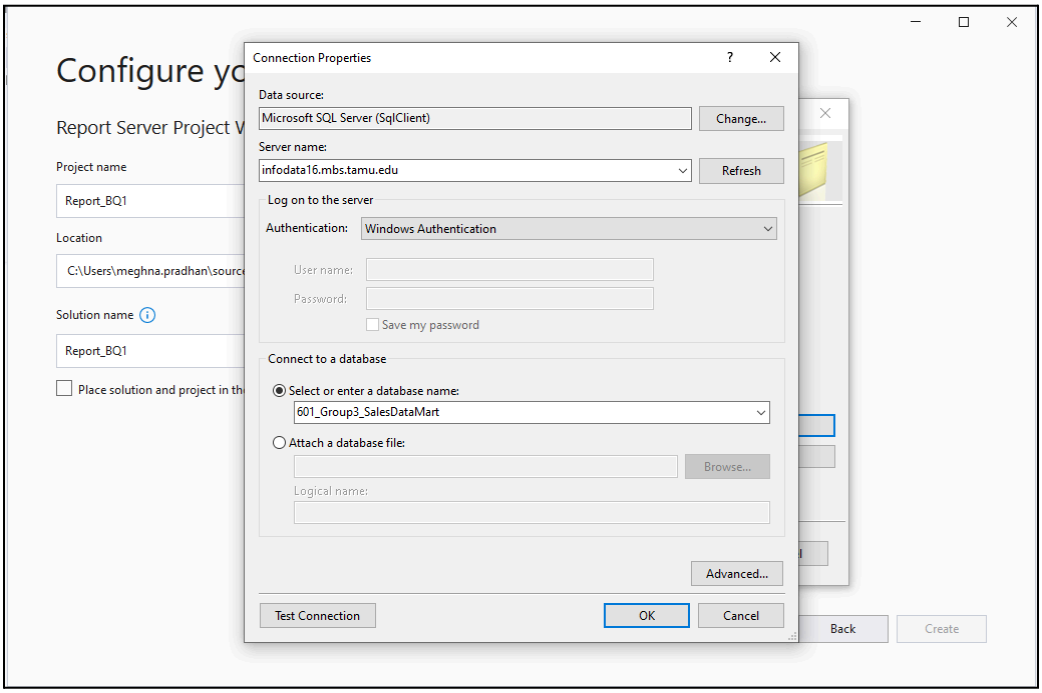


Figure 5.2.1.1 Connecting SSRS to the Data Mart on the server

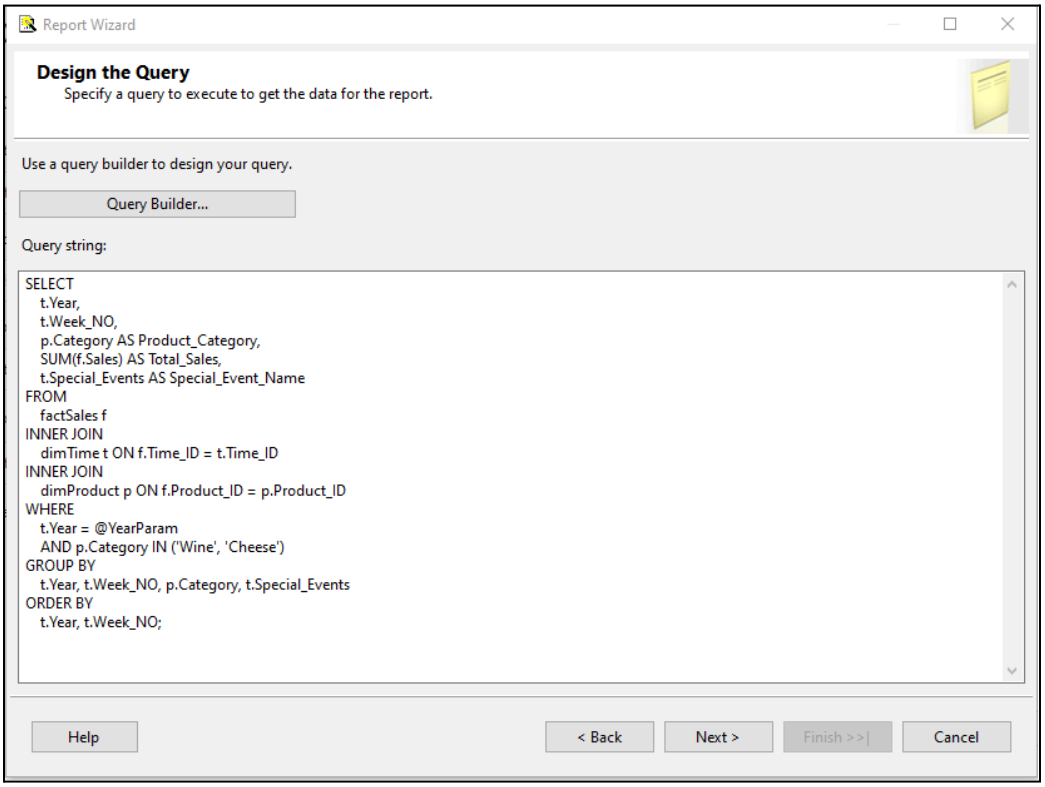


Figure 5.2.1.2 Designing Query for the report in SSRS

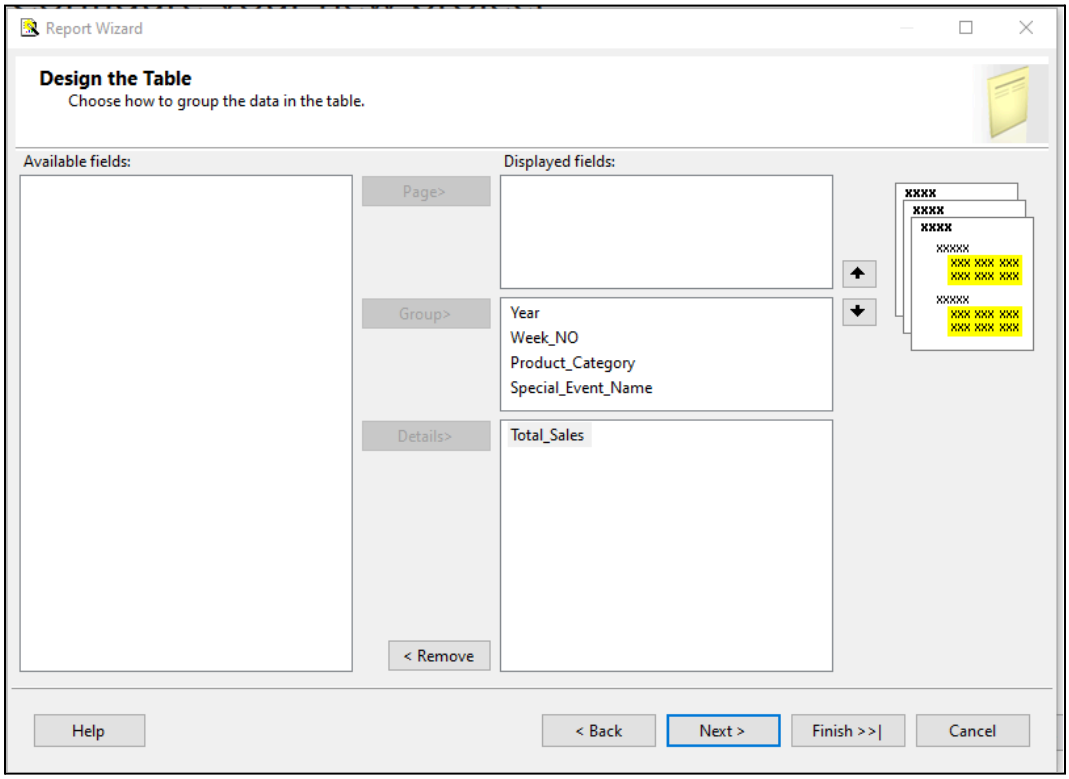


Figure 5.2.1.3 Designing Table for the report in SSRS

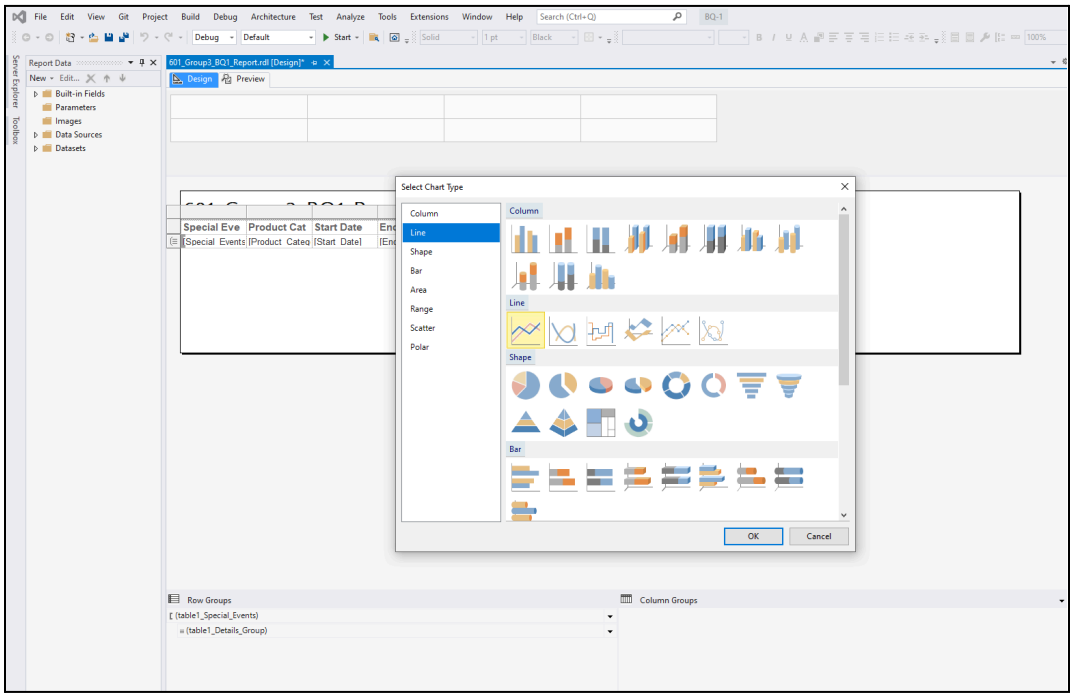
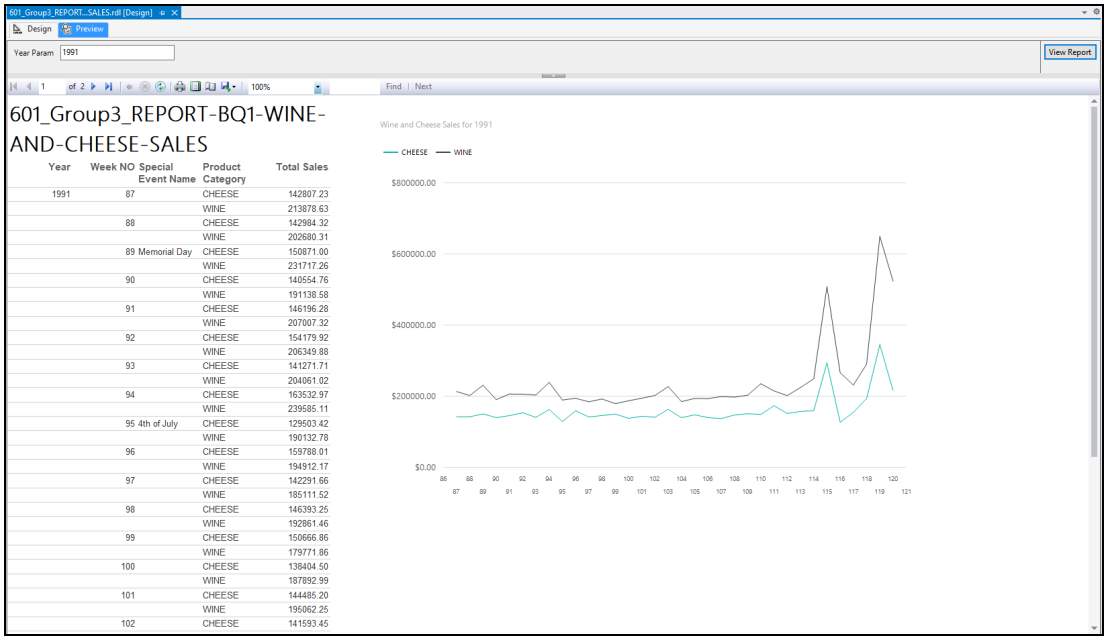
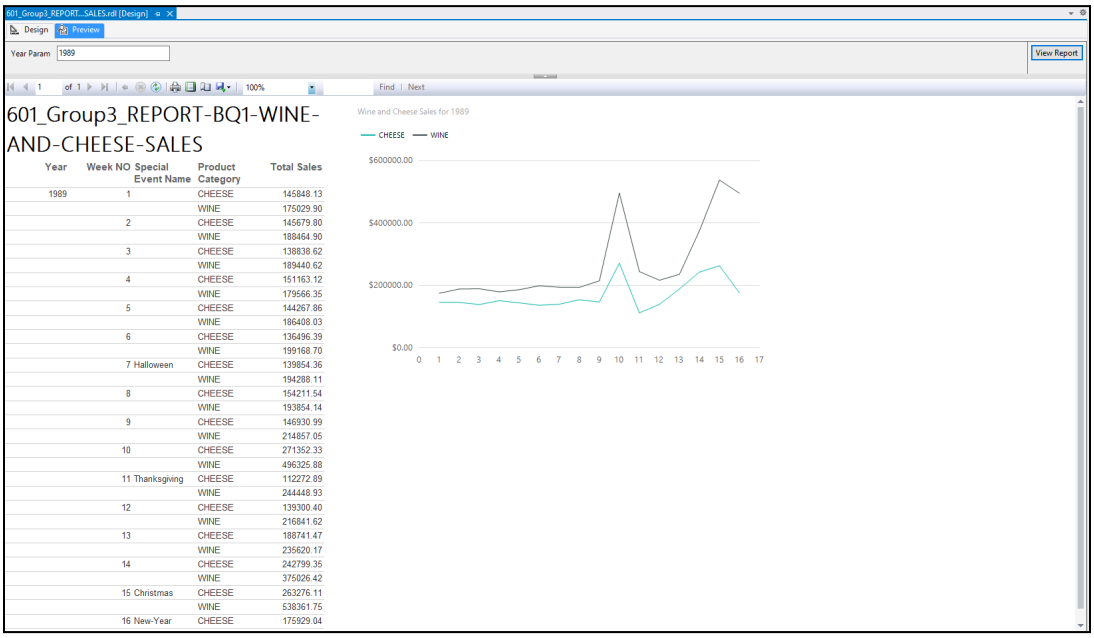


Figure 5.2.1.4 Building the report



Year	Week NO	Special Event Name	Product Category	Total Sales
1991	87		CHEESE	142807.23
			WINE	213878.63
	88		CHEESE	142984.32
			WINE	202680.31
	89	Memorial Day	CHEESE	150871.00
			WINE	231717.26
	90		CHEESE	140554.76
			WINE	191138.58
	91		CHEESE	146196.28
			WINE	207007.32
	92		CHEESE	154179.92
			WINE	206349.88
	93		CHEESE	141271.71
			WINE	204061.02
	94		CHEESE	163532.97
			WINE	239585.11
	95	4th of July	CHEESE	129503.42
			WINE	190132.78
	96		CHEESE	159788.01
			WINE	194912.17
	97		CHEESE	142291.66
			WINE	185111.52
	98		CHEESE	146393.25
			WINE	192861.46
	99		CHEESE	150666.86
			WINE	179771.86
	100		CHEESE	138404.50
			WINE	187892.99
	101		CHEESE	144485.20
			WINE	195062.25
	102		CHEESE	141593.45

Figure 5.2.1.5 Rise for Wine and Cheese Sales during Holiday Season for Year 1991



Year	Week NO	Special Event Name	Product Category	Total Sales
1989	1		CHEESE	145848.13
			WINE	175029.90
	2		CHEESE	145679.80
			WINE	188464.90
	3		CHEESE	138838.62
			WINE	189440.62
	4		CHEESE	151163.12
			WINE	179566.35
	5		CHEESE	144267.86
			WINE	186408.03
	6		CHEESE	136496.39
			WINE	199168.70
	7	Halloween	CHEESE	139854.36
			WINE	194288.11
	8		CHEESE	154211.54
			WINE	193854.14
	9		CHEESE	146930.99
			WINE	214857.05
	10		CHEESE	271352.33
			WINE	496325.88
	11	Thanksgiving	CHEESE	112272.89
			WINE	244448.93
	12		CHEESE	139300.40
			WINE	216841.62
	13		CHEESE	188741.47
			WINE	235620.17
	14		CHEESE	242799.35
			WINE	375026.42
	15	Christmas	CHEESE	263276.11
			WINE	538361.75
	16	New-Year	CHEESE	175929.04

Figure 5.2.1.6 Rise for Wine and Cheese Sales during Holiday Season for Year 1989

Conclusion

The above report plots the sale trends of Wine and Cheese products throughout a particular year, highlighting the peak of sales during specific time periods of the year i.e during Thanksgiving, Christmas and New Year weeks. This trend stays consistent throughout each year for both Wine and Cheese, with Wine sales uniformly being a higher value than Cheese

sales. The figure no 5.2.1.6 shows how the sales of Wine and Cheese peak at Week no 10 which is the week before the Thankgiving week and then again peak during Week 15 which is the Christmas week. This pattern can be seen in all years between 1989 and 1993. Therefore our report answers the question that there is indeed a correlation between Cheese and Wine sales which consistently peak during the holiday season.

5.2.2 Tableau Report for BQ2

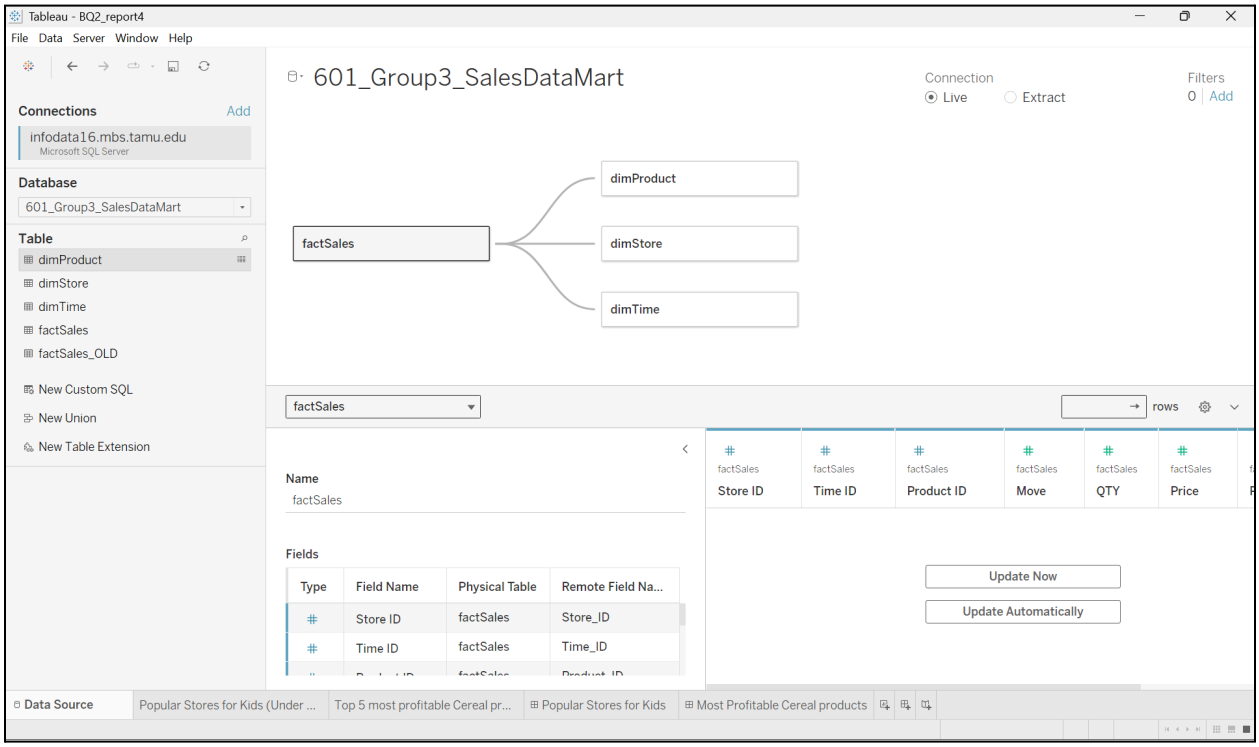


Figure 5.2.2.1 Connected Tableau to Sales Data Mart present on the server

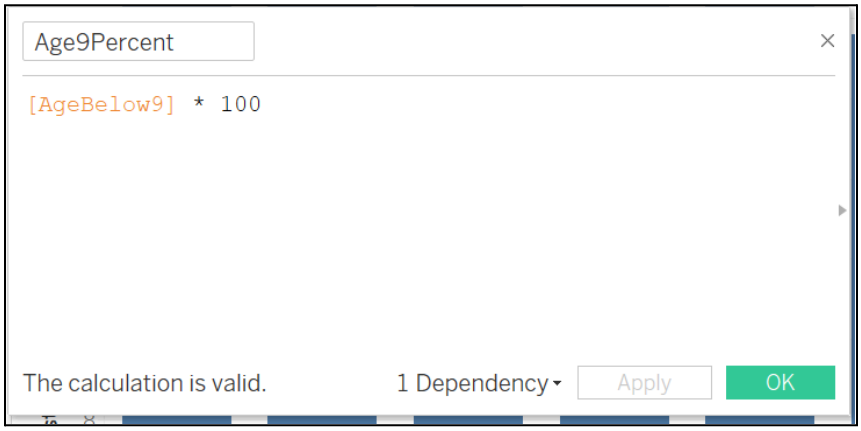


Figure 5.2.2.2 Creating a calculated field for better understanding

Figure 5.2.2.3 Building visualization - finding popular stores for kids under age 9

Figure 5.2.2.4 Final Report with popular stores used as a filter

Conclusion

From the Tableau bar graphs, we can see the top stores that kids under the age of 9 visit. This information can be used to figure out the top 10 or top 15 or top 5 stores that are popular among kids. The most profitable cereal products from these stores then help us understand which are the cereals most favoured by kids. Using this information, DFF management can understand which are the cereals that should be stocked in other stores with a high population of kids. Further analytics can be done to understand if any of these cereals can be bundled with other childrens items for increasing the overall sales of DFF.

5.2.3 SSAS Report for BQ4

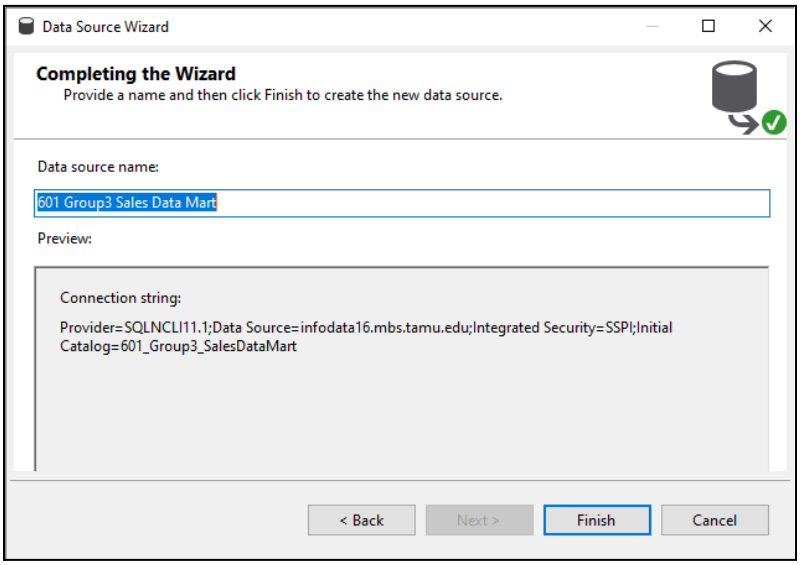


Figure 5.2.3.1 Creating Data Source

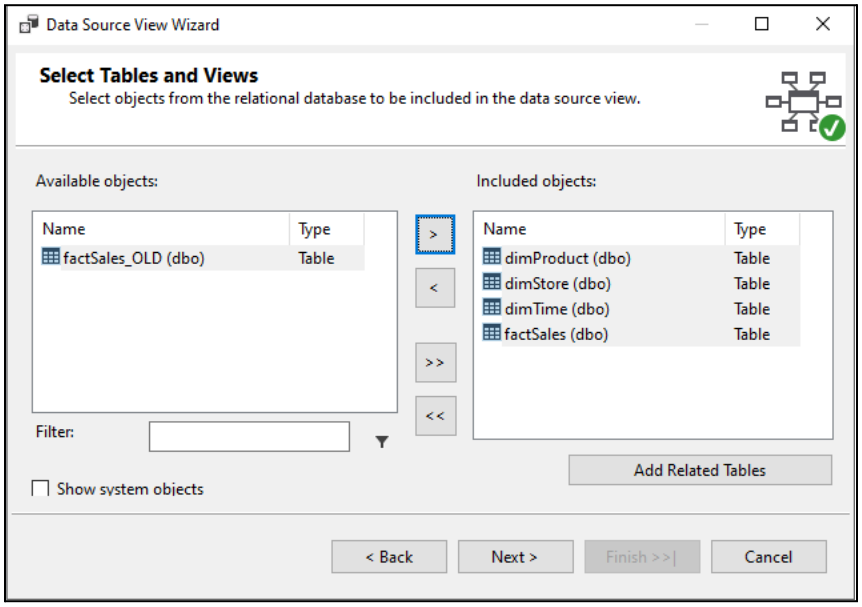


Figure 5.2.3.2 Creating Data Source View

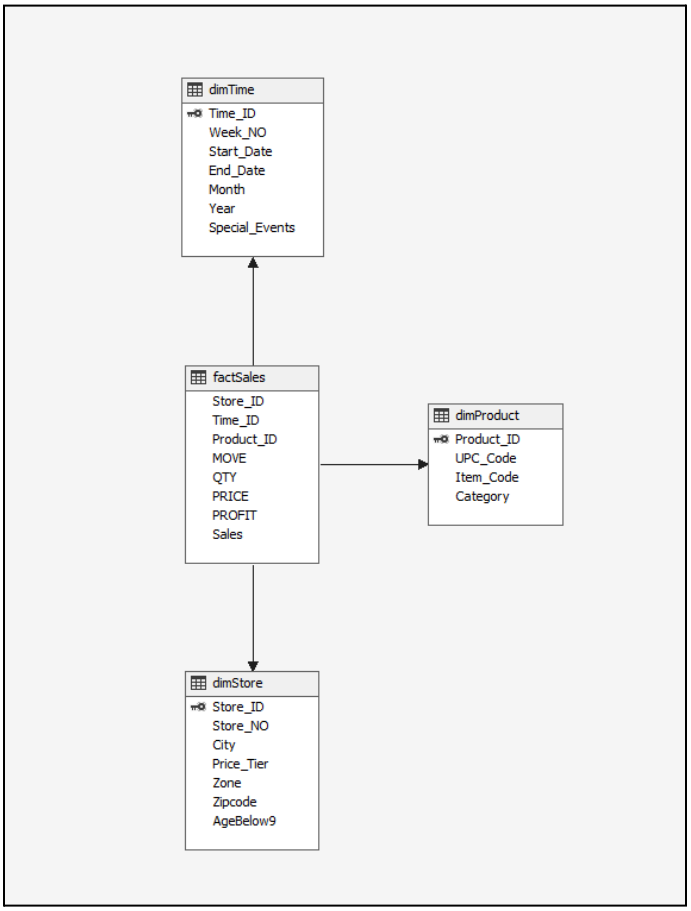


Figure 5.2.3.3 Data Source View - Sales Data Mart

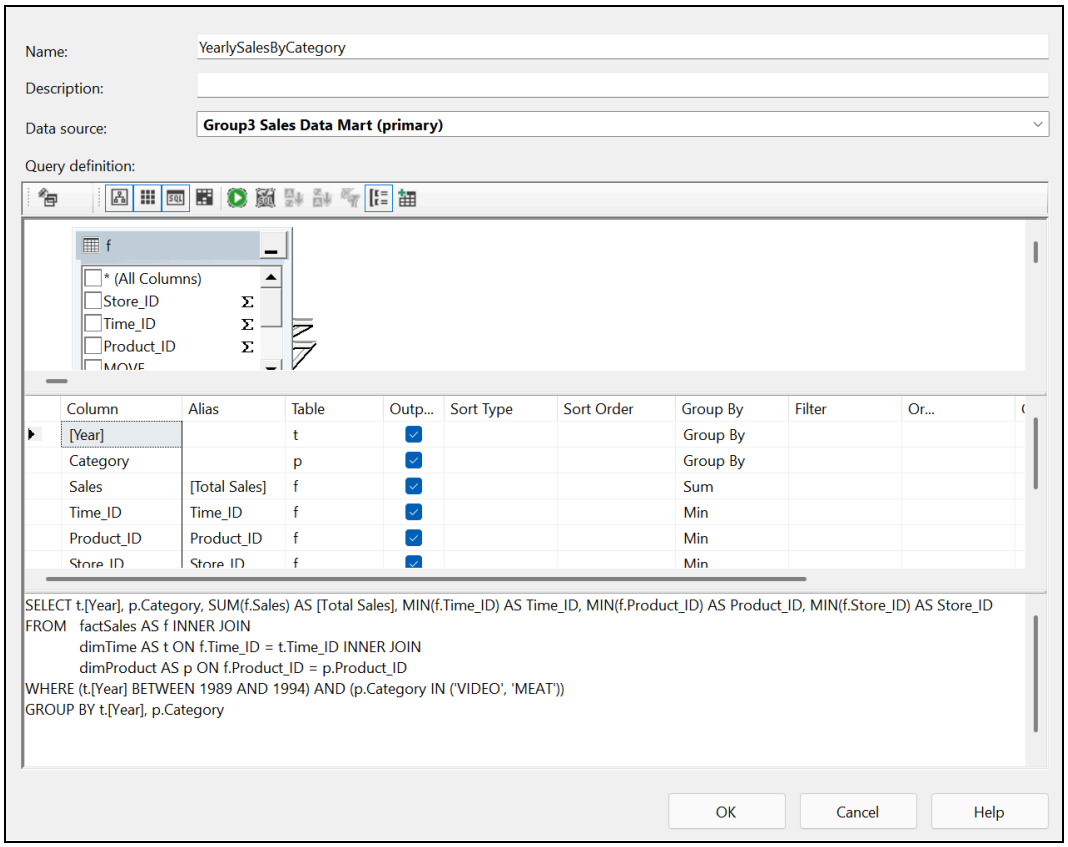


Figure 5.2.3.4 Creating Named Query for the report

Named Query created as follows -

```
SELECT t.[Year], p.Category, SUM(f.Sales) AS [Total Sales], MIN(f.Time_ID) AS Time_ID,
MIN(f.Product_ID) AS Product_ID, MIN(f.Store_ID) AS Store_ID
FROM   factSales AS f INNER JOIN
          dimTime AS t ON f.Time_ID = t.Time_ID INNER JOIN
          dimProduct AS p ON f.Product_ID = p.Product_ID
WHERE (t.[Year] BETWEEN 1989 AND 1994) AND (p.Category IN ('VIDEO', 'MEAT'))
GROUP BY t.[Year], p.Category
```

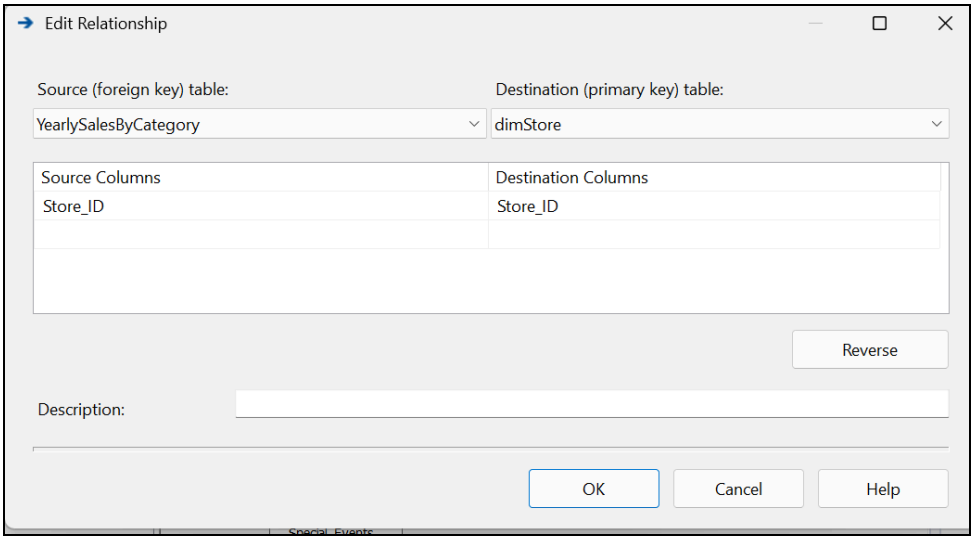


Figure 5.2.3.5 Mapping named query to dimension tables

A similar mapping was done with other dimensions dimTime (YearlySalesByCategory.Time_ID = dimTime.Time_ID) and dimProduct (YearlySalesByCategory.Product_ID = dimProduct.Product_ID)

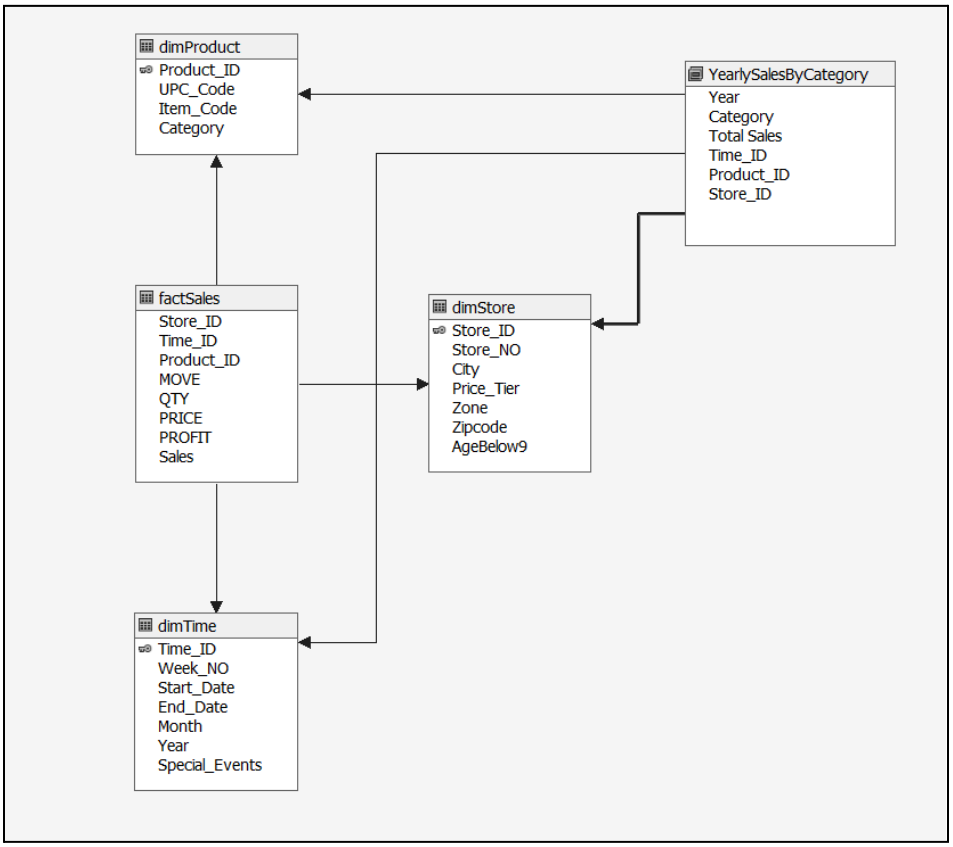


Figure 5.2.3.6 Updated Data Source View after adding named query

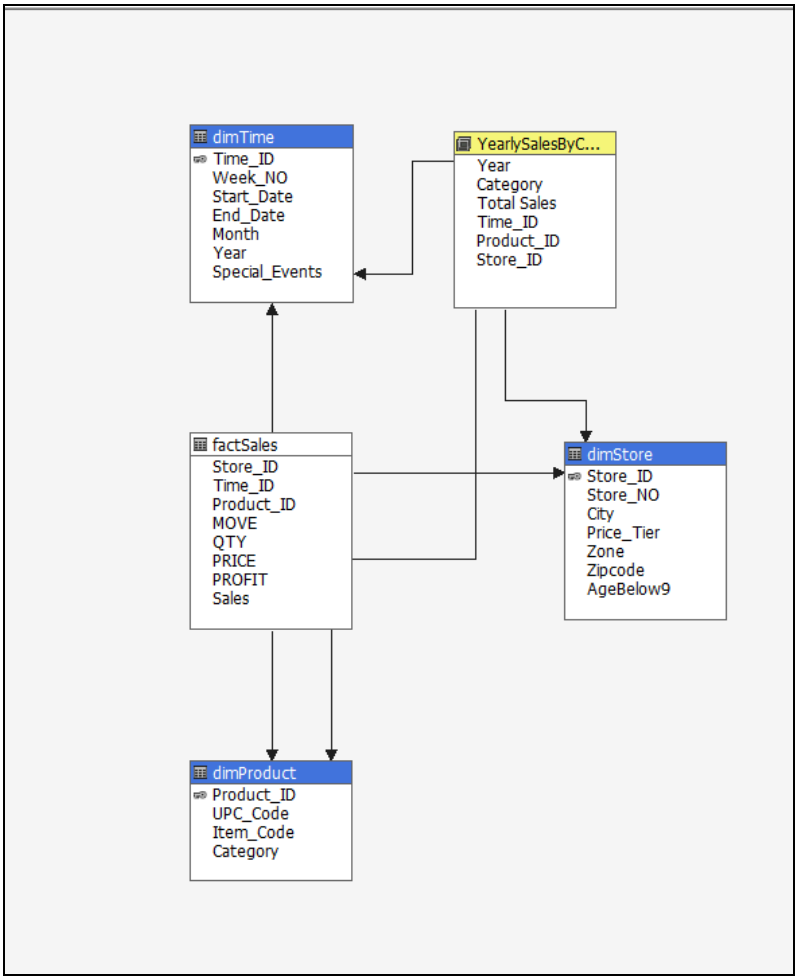


Figure 5.2.3.7 Cube structure - Sales Data Mart

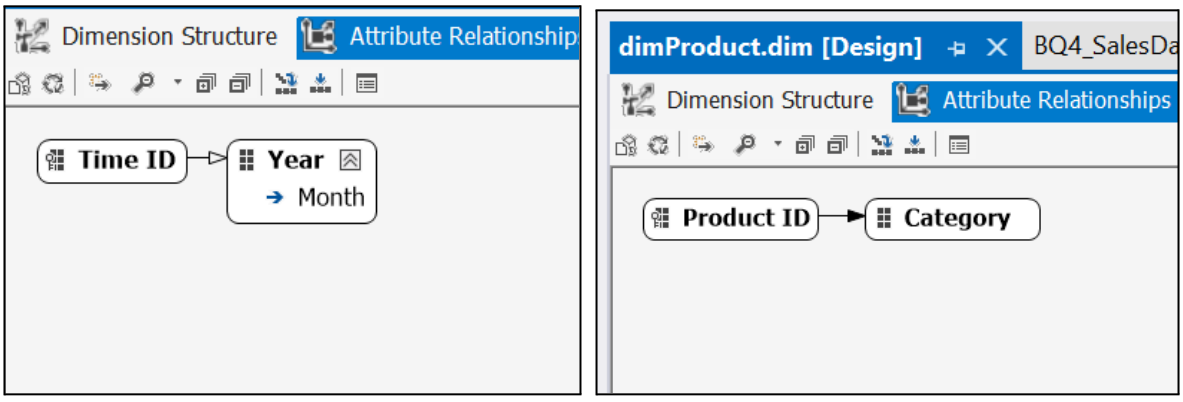


Figure 5.2.3.8 Creating hierarchies for dimTime and dimProduct

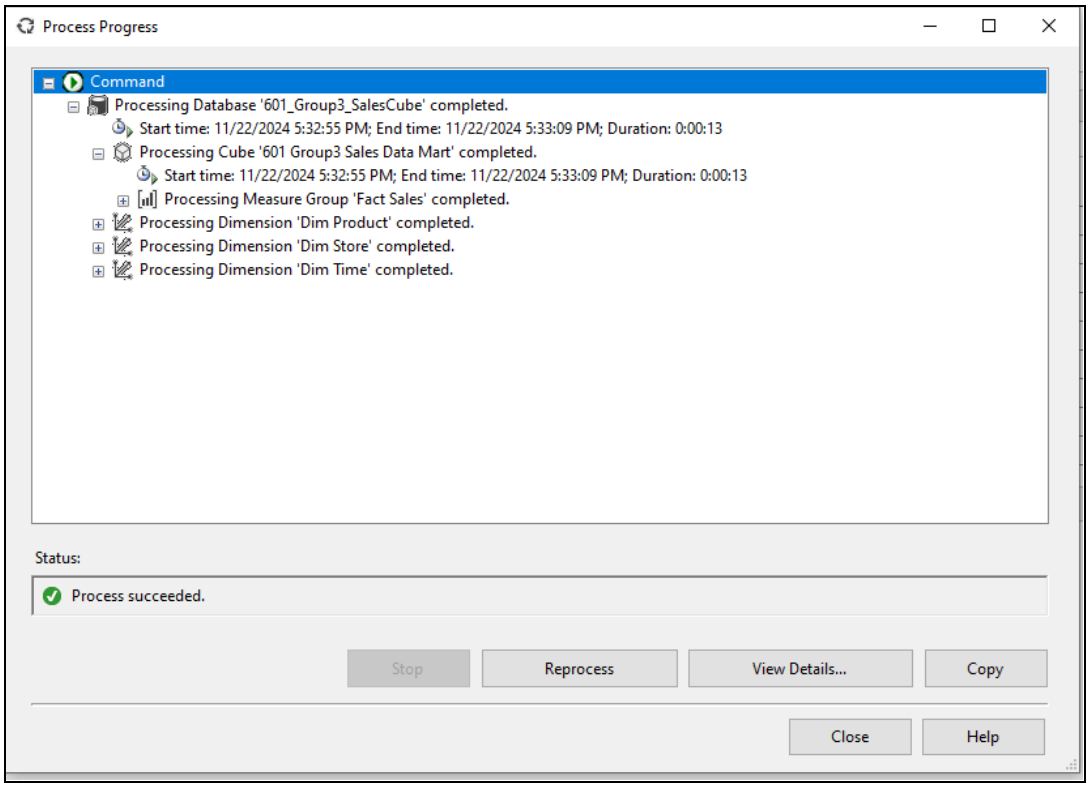


Figure 5.2.3.9 Processing the cube

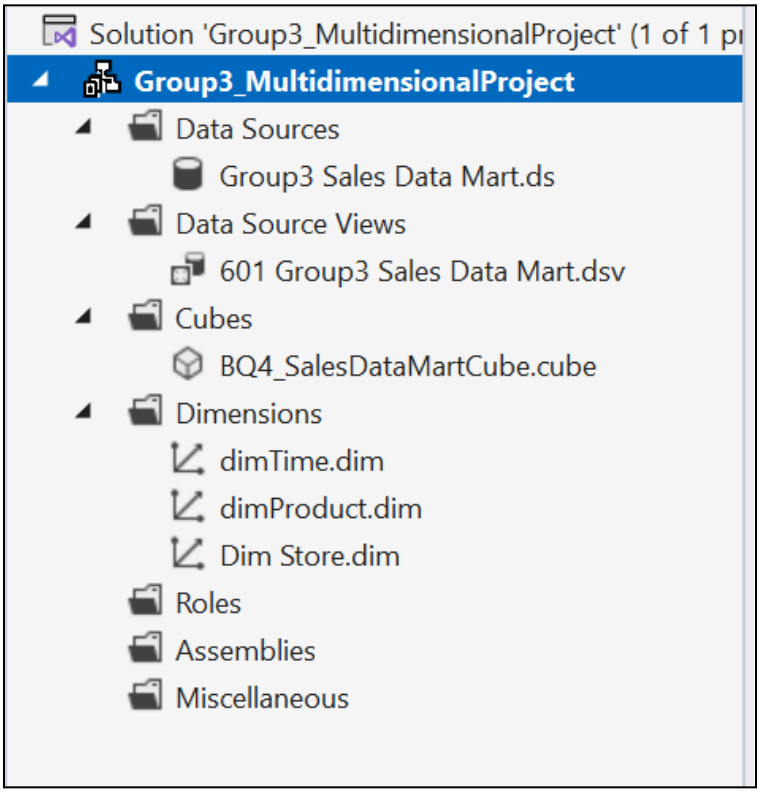


Figure 5.2.3.10 Various elements created in SSAS for answering the business question

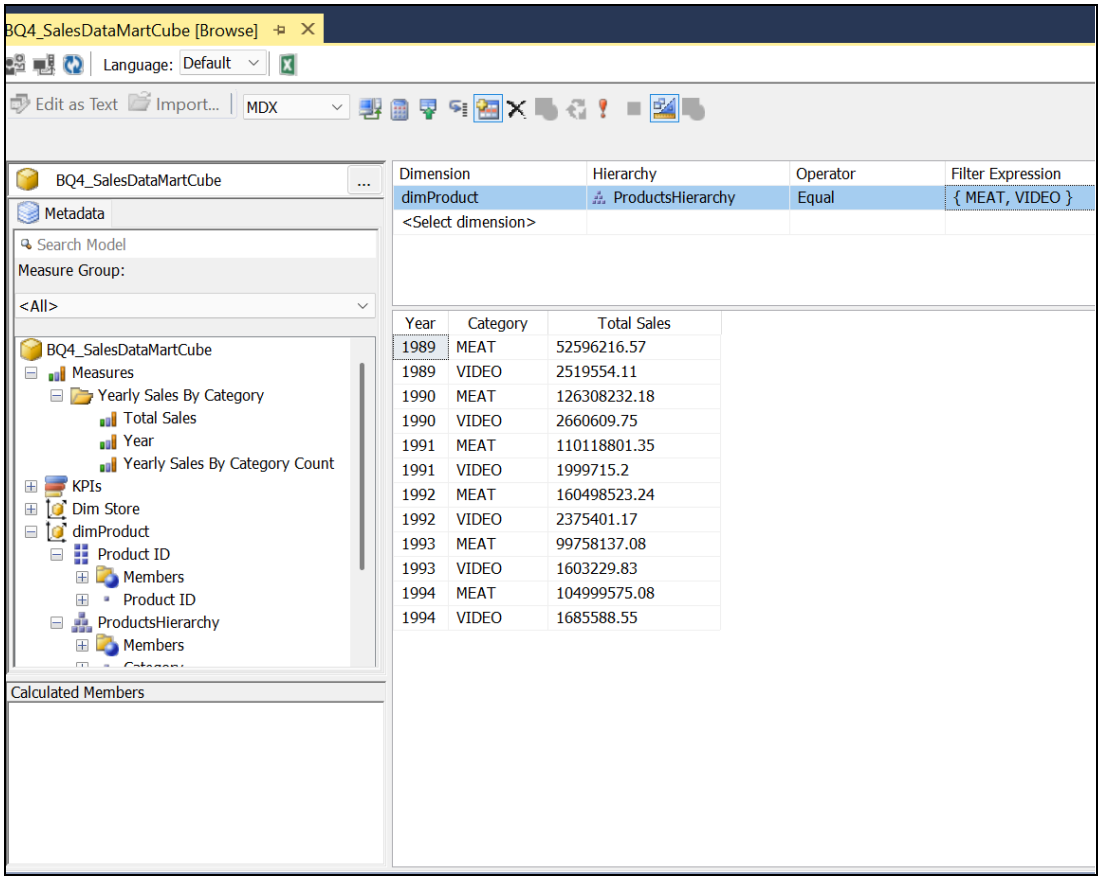


Figure 5.2.3.11 Final report using only SSAS

Conclusion

This BQ is an easy to compare BQ since only 2 products and 6 years are in question. Hence the SSAS reports are sufficient for us to check how the total sales of products like Meat and Video are trending between the years 1989-1994. If a more visual output is required by the management, a SSRS or Tableau chart can be created over this cube. From the output of the SSAS report, we can see that Meat has had an increasing trend over the years. The management can use this information to anticipate similar or slightly more stock for the upcoming years. This will allow the management to be aware of the annual movement of different products in the DFF stores.

5.2.4 Tableau Report for BQ5

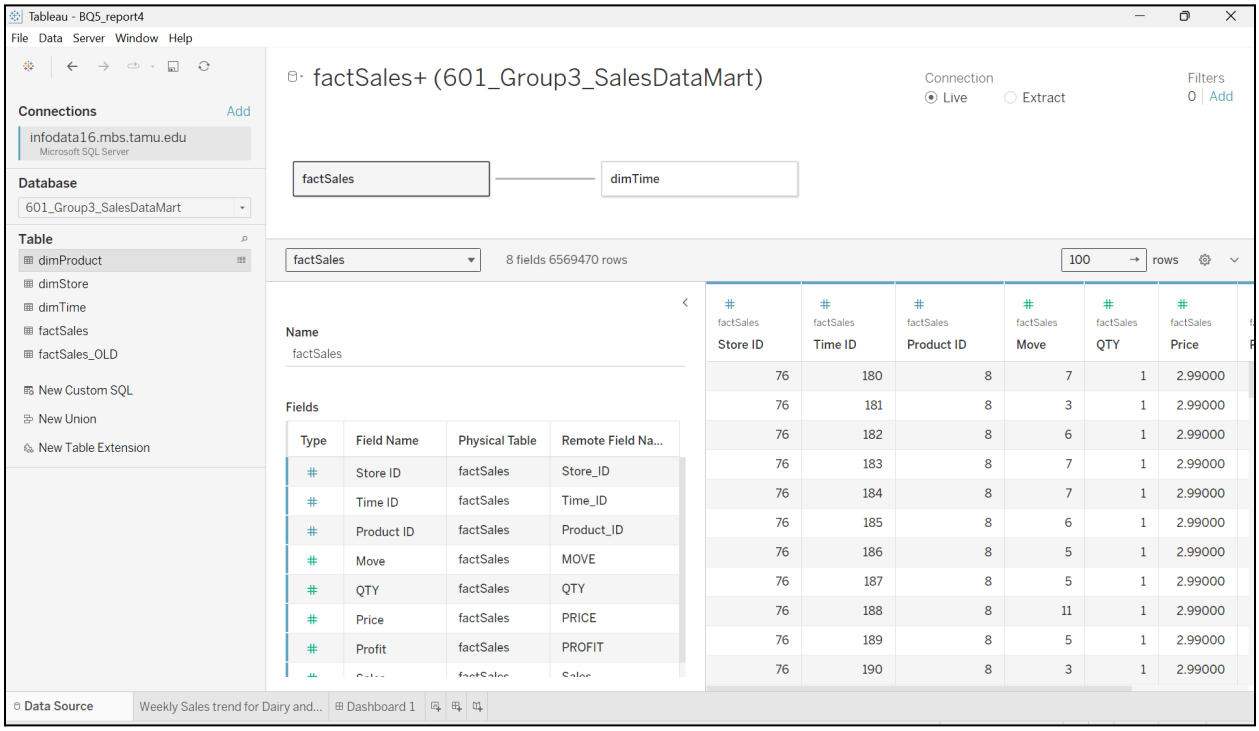


Figure 5.2.4.1 Connected Tableau to Sales Data Mart present on the server

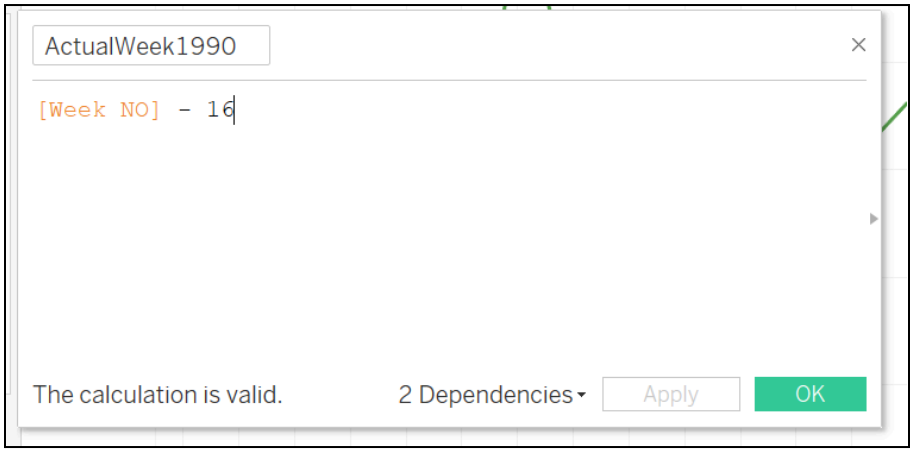


Figure 5.2.4.2 Creating a calculated field for better understanding

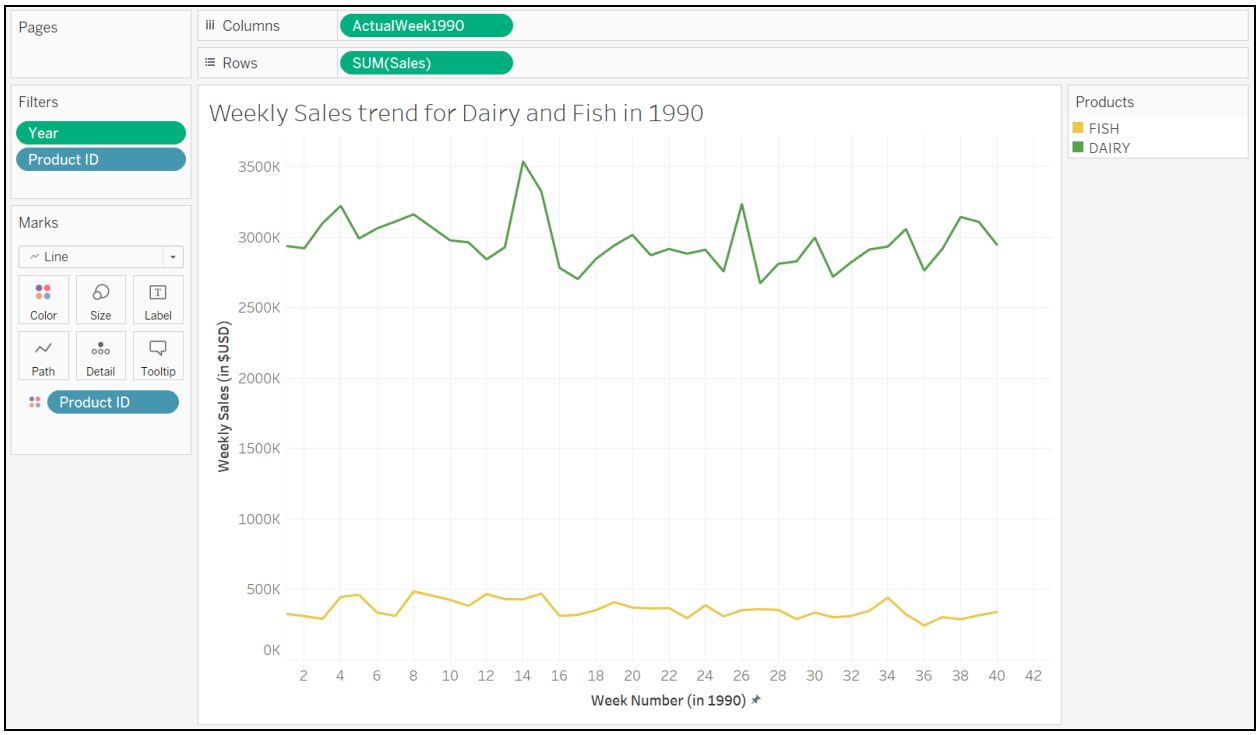


Figure 5.2.4.3 Building the visualization in Tableau

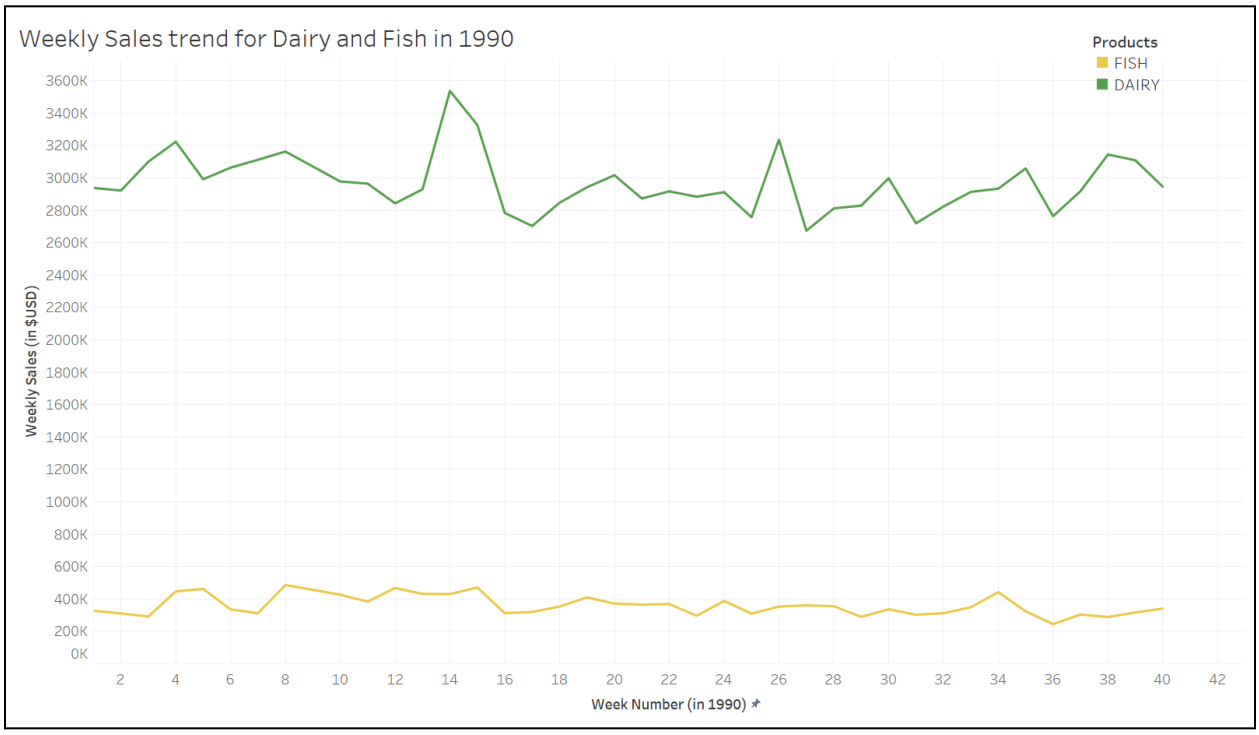


Figure 5.2.4.4 Final Dashboard Report on Tableau

Conclusion

From the graphs obtained via tableau, we can see patterns in the sales of perishable items like dairy and fish. This gives the management an idea as to which weeks of the year are these items more popular. From these charts,we also see that the peaks and descents of both these products are on similar weeks. This is new information that the management can use to analyze further and increase the over sales by increasing stocks during these weeks of upcoming years. Bundling offers can also be done for improving the sales.

5.2.5 Tableau on top of SSAS Report for BQ6

Creating Cube on SSAS

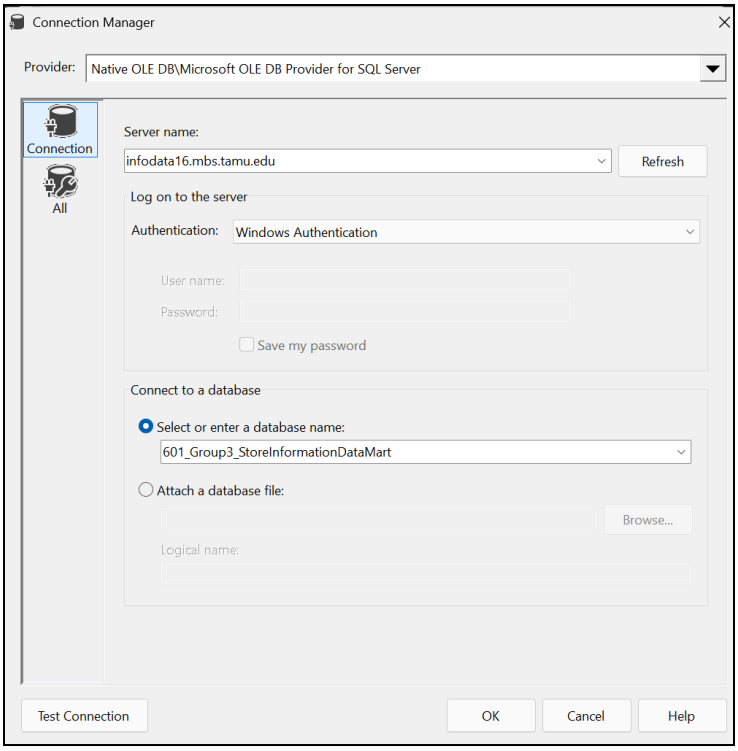


Figure 5.2.5.1 Creating a Data Source for Store-Information Data Mart

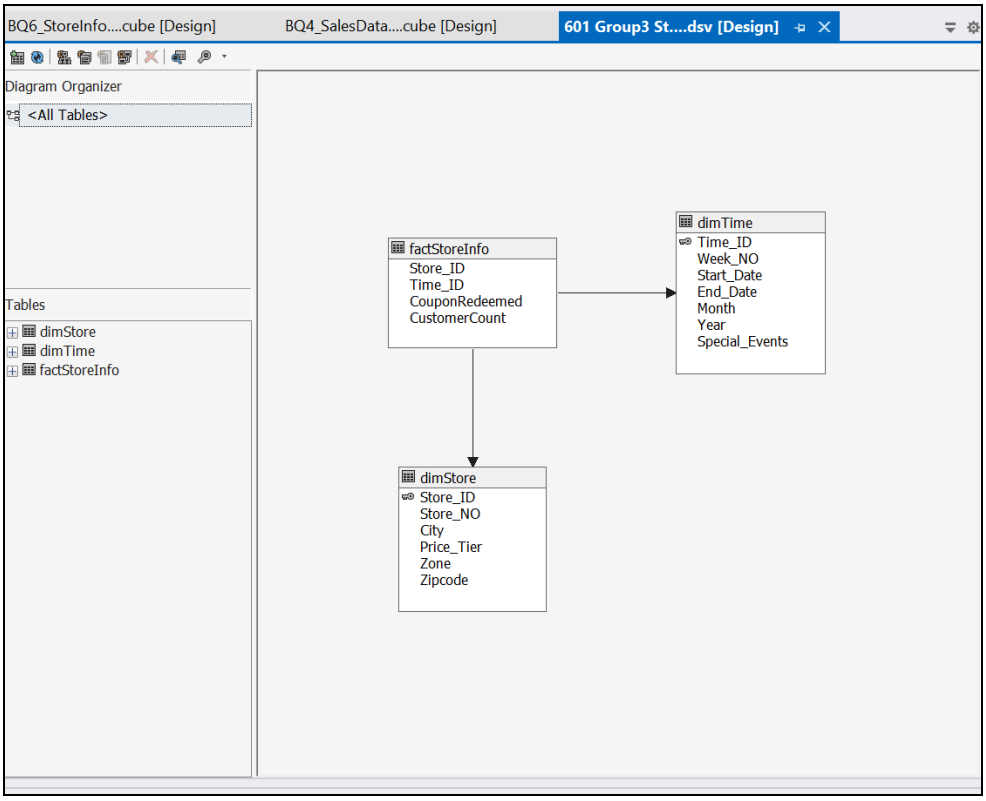


Figure 5.2.5.2 Creating a Data Source View for Store-Information Data Mart

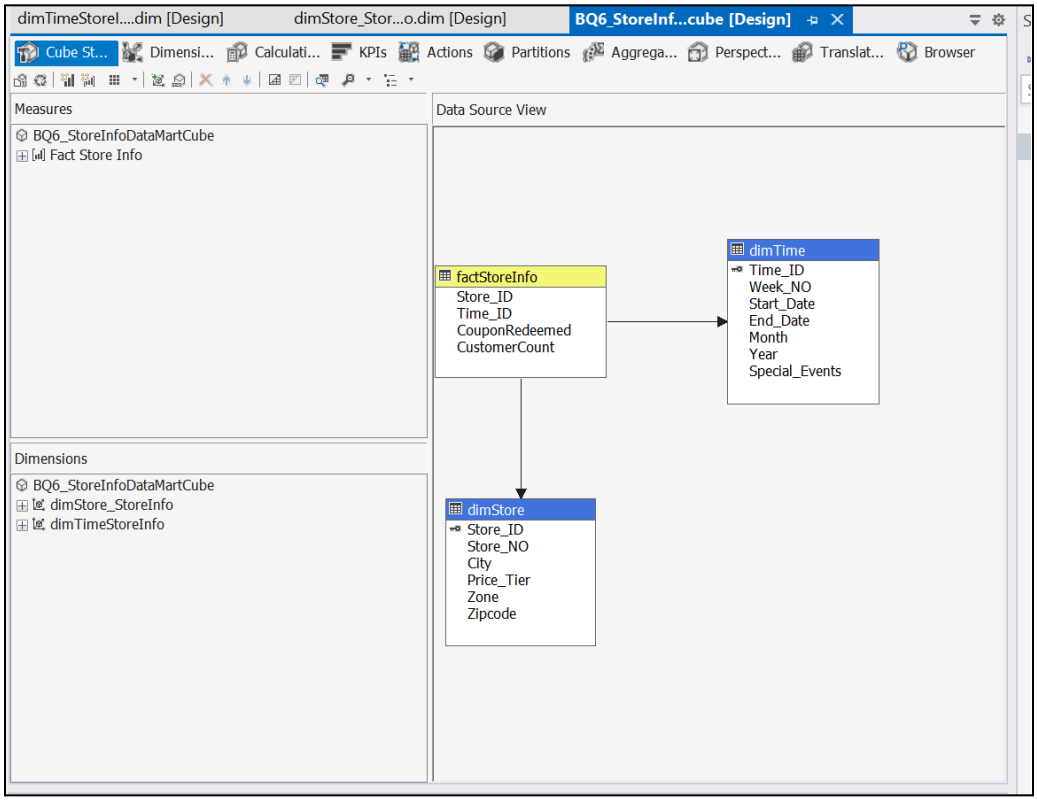


Figure 5.2.5.3 Cube Structure for Store-Information Data Mart

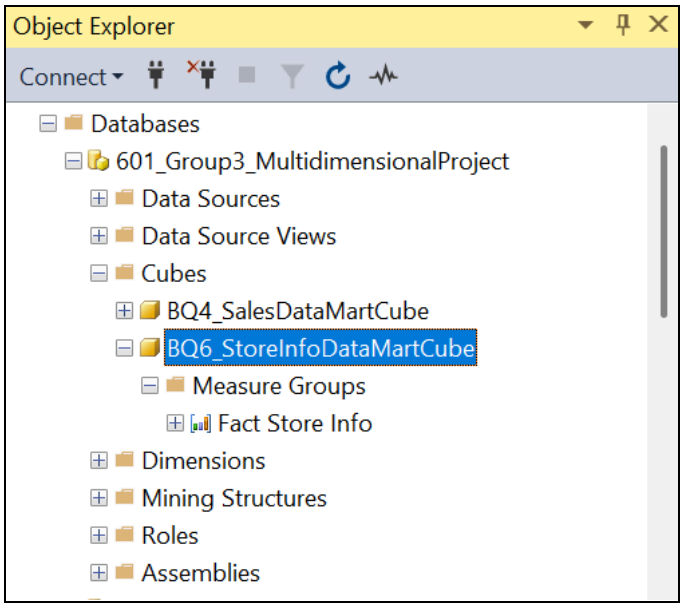


Figure 5.2.5.4 Deployed cube in the server

Creating Report on Tableau

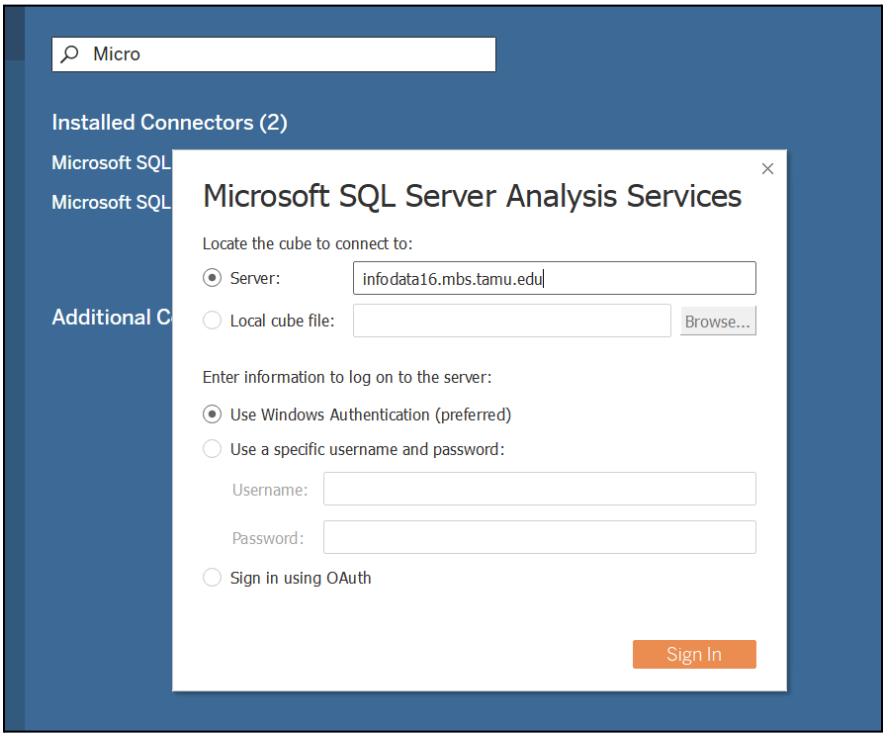


Figure 5.2.5.5 Connecting to the server on Tableau

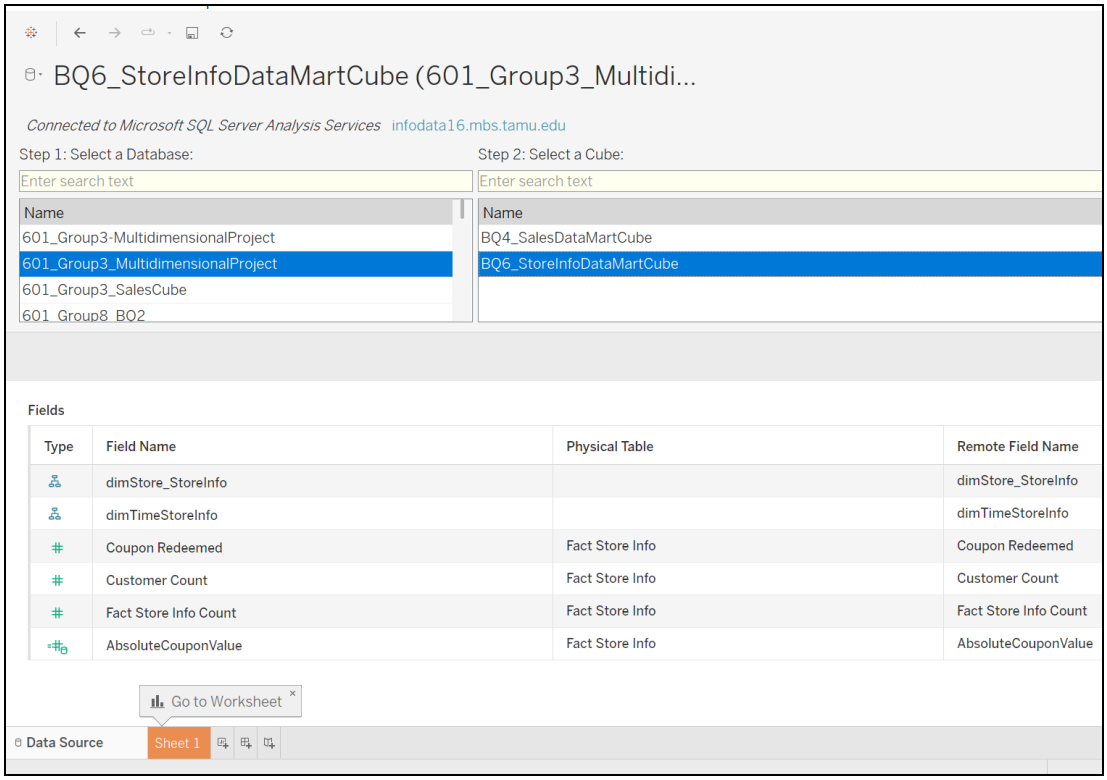


Figure 5.2.5.6 Choosing the cube from the server on Tableau

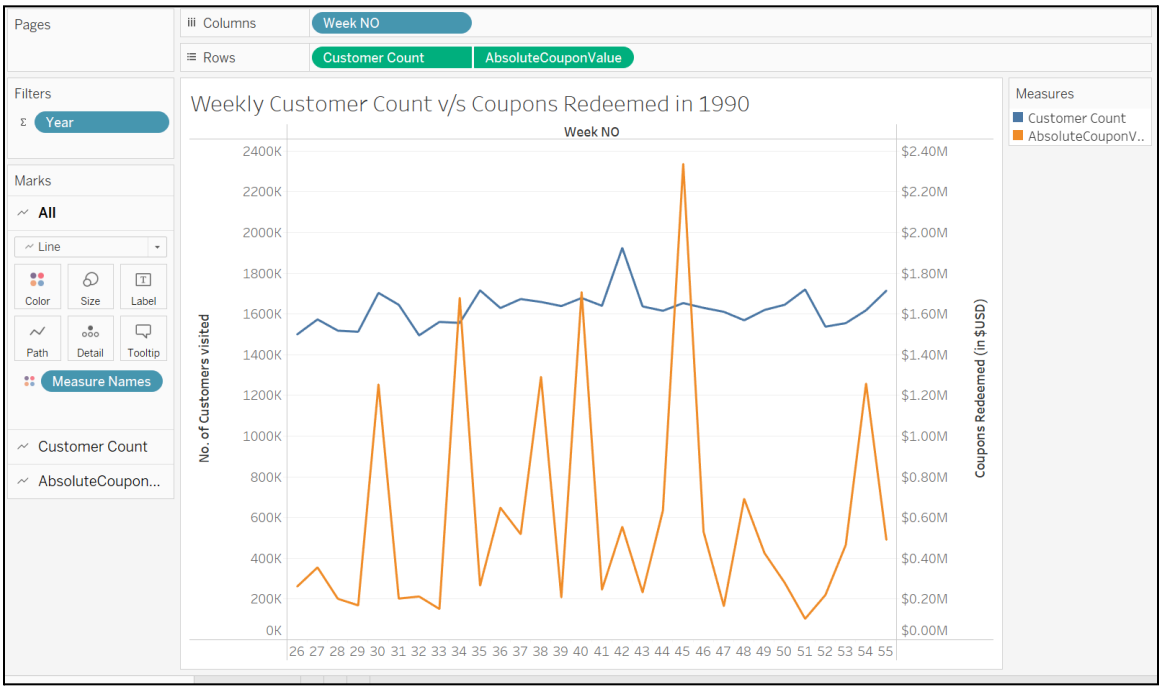


Figure 5.2.5.7 Building the visualization on Tableau

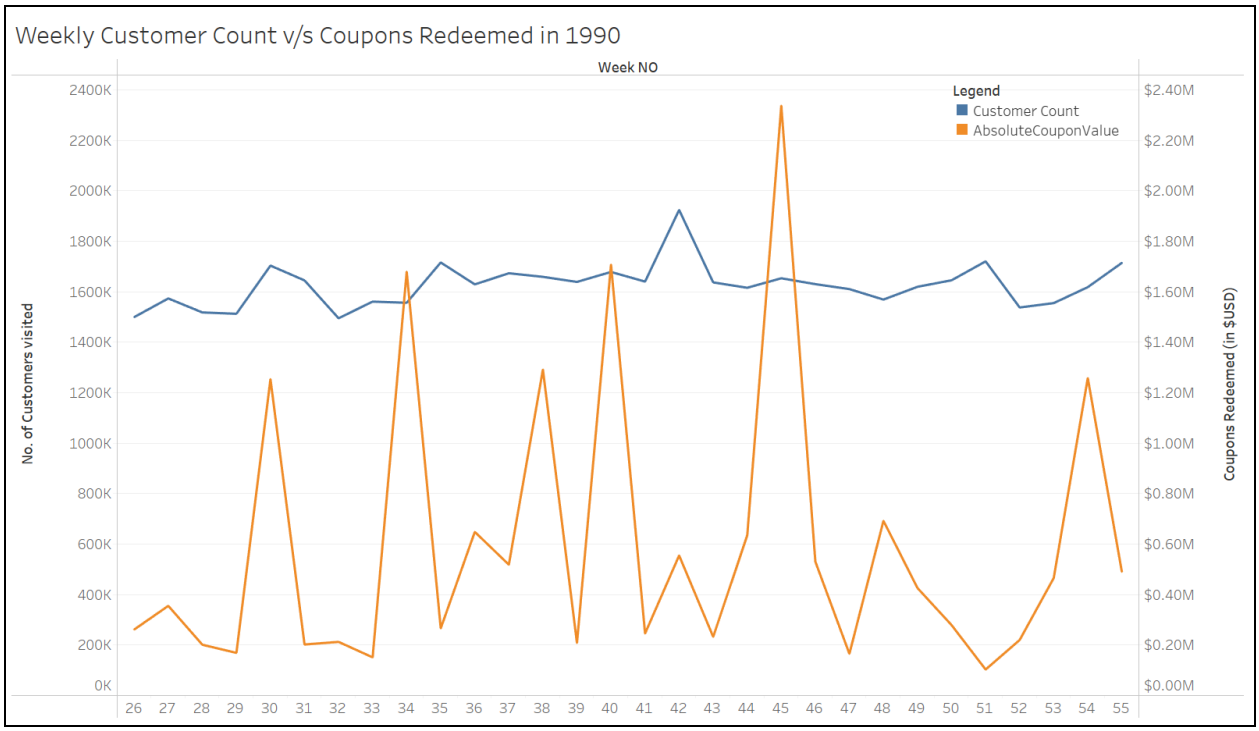


Figure 5.2.5.8 Final Dashboard Report on Tableau

Conclusion

We expect that coupons increase the store traffic. However, to ensure that the introduction of the coupons are bringing in the expected foot value, we need to rely on the graphs like above. From these graphs we can see that the peaks of both the lines are not aligning. This means that either the coupons offered during those periods are not lucrative or coupons in general do not have a correlation with the store's foot value. Analysts can further drill down and analyse this which will help the management understand how to invest in coupons in the future.

References

1. Strauss, 1998, "A Marketing Professor's Shopping List" (https://magazine.wharton.upenn.edu/issues/spring-1998/a-marketing-professors-shopping-list/)
2. Carpenter, Moore, 2006, "Consumer demographics, store attributes, and retail format choice in the US grocery market" (https://www.emerald.com/insight/content/doi/10.1108/09590550610667038/full/html)
3. Jaravaza, Chitando, 2013, The Role of Store Location in Influencing Customers' Store Choice

(https://www.scholarlinkinstitute.org/jetems/articles/The%20Role%20of%20Store%20Location.pdf)
4. Tan, 2024, “Data Analytics: Challenges in Retail Basic Data Analytics” (https://off-grid.sg/data-analytics-challenges-in-retail-basic-data-analytics/)