**Design and Implementation of a Data Warehouse for a Retail Store with Store-level Data**

**Consulting Report-2**

*Logical Design the Data Warehouse Schema*

**Group 11**

George, Alpha

Rafi, Parvez

Shetty, Sharvil

Table of Contents

[1. Business Questions 2](#_Toc478470936)

[2. Proposed schema 2](#_Toc478470937)

[3. Kimball’s Matrix for Data Marts 6](#_Toc478470938)

[4. Star Schema representation 6](#_Toc478470939)

[5. Mapping tables 8](#_Toc478470940)

[6. Business question justification corresponding to proposed schema 12](#_Toc478470941)

[7. Conclusion 15](#_Toc478470942)

### Business Questions

1. What is the trend of wine sales during Christmas holiday season?
2. Which stores have more popularity among kids and elderly groups?
3. Which stores attract people who earn below poverty line?
4. What is the effect of introducing coupons on total number of customer visits?
5. What is the trend of a product demand in different price-tiers?

### Proposed schema

The schema being proposed to answer the above business questions is a combination of the five dimensions viz. Store, Product, Time, Demographic, Coupon and two fact tables- SALES\_FACT which will hold the data from the MOVEMENT source file and STORE\_VISITS\_FACT which will hold the data from the CCOUNT source file.

The dimension STORE\_DIM dimension contains details about all the DFF stores. This dimension is populated with data from the Dominic’s research project manual, which contains the details about each store under the DFF retail chain. It has an inherent hierarchy in Zone, City and Store number.



*Fig 1: Store dimension table*

The PRODUCT\_DIM dimension contains the data related to all the products sold at DFF. The UPC files for each product act as the source for this data. It is a collection of data from each of the categories of products. The product category column is populated from the name of the csv file from which the data is being loaded.



*Fig 2: Product dimension table*

The TIME\_DIM dimension contains the data related to the occurrence of an event. The attribute TIME\_ID is an auto incremented surrogate key, acting as the primary key for the table. As stated in the manual, the data contains observations from week 0 to 400, based on this fact, the field WEEK is auto populated. The WEEK acts as the base to calculate MONTH and YEAR as described in the transformation for the time dimension.



*Fig 3: Time dimension table*

The COUPON\_DIM dimension stores all the types of coupons there are available at the stores. This value is derived from the header column of the CCOUNT.csv file. It is a manual task to separate out the names of the fields, which represent the coupon usage in the dataset.



*Fig 4: Coupon dimension table*

The DEMOGRAPHIC\_DIM dimension is derived from the DEMO.csv table. It stores all the demographics related to each store. It contains clearly defined statistics for various demographic factors, which are directly mapped from the source table.



*Fig 5: Demographic dimension table*

The SALES\_FACT table stores the values derived from the movement datasets. Product, Store and Time are linked to this fact table. This fact table is used to address business questions 1 and 5.



*Fig 6: Sales fact table*

The STORE\_VISITS\_FACT is designed to address business question 2, 3 and 4. This fact table enables us to answer the queries related to introduction of coupons and customer visits. The CUSTOMER\_COUNT stores the number of customers who visited the store.



*Fig 7: Store visits fact table*

### Kimball’s Matrix for Data Marts

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data mart** |  | **Dimension** |  |  |  |
|  | STORE\_DIM | TIME\_DIM | PRODUCT\_DIM | DEMOGRAPHIC\_DIM | COUPON\_DIM |
| Sales | x | x | x |  |  |
| Store visits | x | x |  | x | x |

### Star Schema representation



*Fig 8: Star schema for Sales data mart*



*Fig 9: Star schema for Store visits data mart*

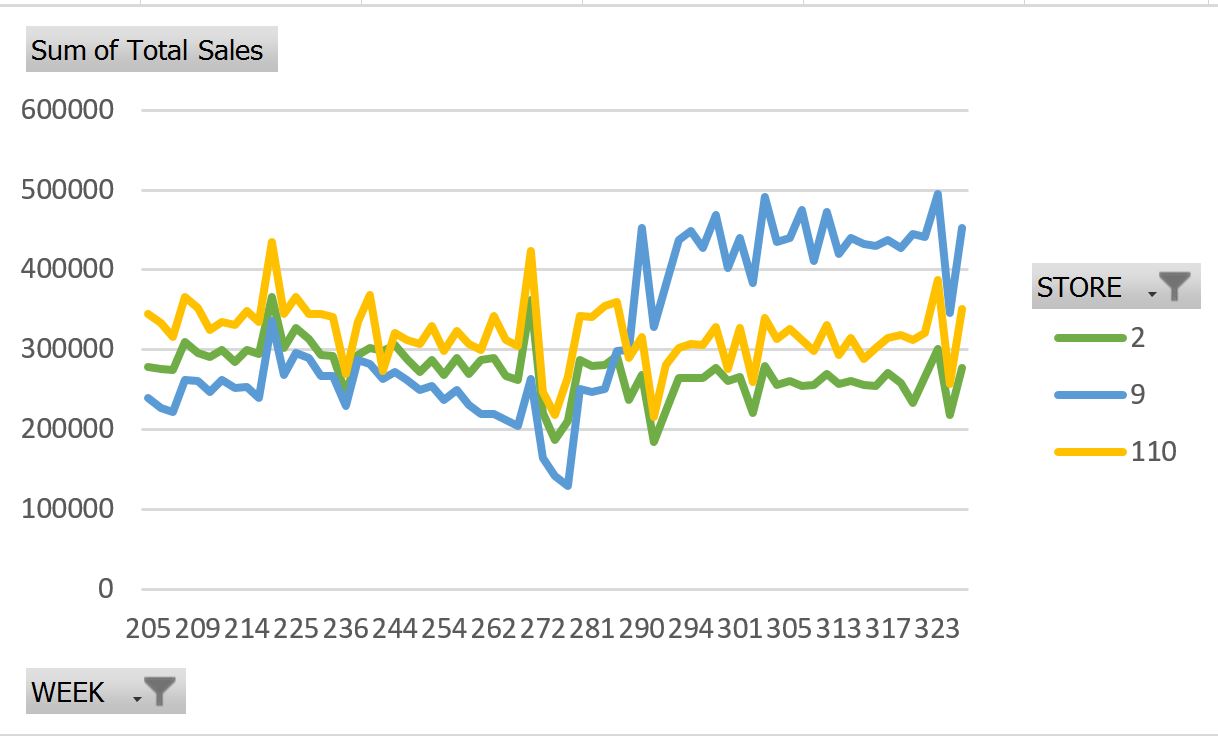
### Mapping tables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dimension: Product** | | | | | | |
| **DW Target Table** | **DW Target Column** | **Target Datatype** | **Source System/Table** | **Source Column** | **Transformation/Business Rule** | **Error Handling Rules** |
| PRODUCT\_DIM | PRODUCT\_ID | int |  | surrogate key |  |  |
| UPC\_CODE | int | UPCXXX.csv | upc |  |  |
| PRODUCT\_CATEGORY | varchar | UPCXXX.csv |  | XXX in source file name corresponds to the category code. e.g. If source csv file is named 'upcana', PRODUCT\_CATEGORY will be 'ana' |  |
| ITEM\_CODE | varchar | UPCXXX.csv | nitem |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **Dimension: Demographic** | | | | | | |
| **DW Target Table** | **DW Target Column** | **Target Datatype** | **Source System/Table** | **Source Column** | **Transformation/Business Rule** | **Error Handling Rules** |
| DEMOGRAPHIC\_DIM | DEMO\_ID | int |  | surrogate key |  |  |
| INCOME\_LEVEL | varchar | DEMO.csv | income | Income in source column is log of median income (M.I.), which ranges from 9.87 to 11.24. e9.87=$19,341 and e11.24 = $76,114. We define 3 levels for INCOME\_LEVEL:  1. *Low*: <10.3 (M.I. less than $30,000 ) 2. *Medium*: 10.3 to 11.0(M.I. between $60,000) 3. *High*: >11.0 (M.I. greater than $60,000) | If income field is null in source table, update INCOME\_LEVEL with NA |
| POOR\_% | float | DEMO.csv | poverty |  |  |
| BELOW\_9\_% | float | DEMO.csv | age9 |  |  |
| ABOVE\_60\_% | float | DEMO.csv | age60 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **Dimension: Coupon** | | | | | | |
| **DW Target Table** | **DW Target Column** | **Target Datatype** | **Source System/Table** | **Source Column** | **Transformation/Business Rule** | **Error Handling Rules** |
| COUPON\_DIM | COUPON\_ID | int |  | surrogate key |  |  |
| COUPON\_REDEEMED | float | ccount |  | Sum up amount of coupons redeemed across all product categories. i.e. COUPON\_REDEEMED = BAKCOUP+BULKCOUP+COSMCOUP+FISHCOUP+… | for any missing coupons redeemed value across a product category, assume the value to be 0 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **Dimension: Store** | | | | | | |
| **DW Target Table** | **DW Target Column** | **Target Datatype** | **Source System/Table** | **Source Column** | **Transformation/Business Rule** | **Error Handling Rules** |
| STORE\_DIM | STORE\_ID | int |  | Surrogate key |  |  |
| STORE\_NUMBER | int | Dominick’s Stores | Store |  |  |
| PRICE\_TIER | varchar | Dominick’s Stores | Price Tier |  | If a row doesn’t have a value for the price tier, set the target as ‘NA’ |
| ZONE | int | Dominick’s Stores | Zone |  | If a row doesn’t have a value for the price tier, set the target as ‘NA’ |
| CITY | varchar | Dominick’s Stores | City |  |  |
| ZIPCODE | int | Dominick’s Storess | Zip Code |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **Dimension: Time** | | | | | | |
| **DW Target Table** | **DW Target Column** | **Target Datatype** | **Source System/Table** | **Source Column** | **Transformation/Business Rule** | **Error Handling Rules** |
| TIME\_DIM | TIME\_ID | int | Surrogate key |  |  |  |
| WEEK | int | Week’s Decode Table | Week # | Populated from 1 to 400 |  |
| MONTH | int | Week’s Decode Table | Start | The start column is of format MM/DD/YY. Split it to get MM |  |
| YEAR | int | Week’s Decode Table | Start | The start column is of format MM/DD/YY. Split it to get YY |  |
|  | EVENT | float | Week’s Decode Table | Special Events |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **Fact: Store Visits** | | | | | | |
| **DW Target Table** | **DW Target Column** | **Target Datatype** | **Source System/Table** | **Source Column** | **Transformation/Business Rule** | **Error Handling Rules** |
| STORE\_VISITS\_FACT | DEMO\_ID | int | DEMOGRAPHIC\_DIM | DEMO\_ID | Foreign key corresponding to primary key DEMO\_ID of DEMOGRAPHIC\_DIM dimension |  |
| COUPON\_ID | int | COUPON\_DIM | DEMO\_ID | Foreign key corresponding to primary key COUPON\_ID of COUPON\_DIM dimension |  |
| STORE\_ID | int | STORE\_DIM | STORE\_ID | Foreign key corresponding to primary key STORE\_ID of STORE\_DIM dimension |  |
| TIME\_ID | int | TIME\_DIM | TIME\_ID | Foreign key corresponding to primary key TIME\_ID of TIME\_DIM dimension |  |
| CUSTOMER\_COUNT | int | ccount | CUSTCOUN |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **Fact: Sales** | | | | | | |
| **DW Target Table** | **DW Target Column** | **Target Datatype** | **Source System/Table** | **Source Column** | **Transformation/Business Rule** | **Error Handling Rules** |
| SALES\_FACT | STORE\_ID | int | STORE\_DIM | STORE\_ID | Foreign key corresponding to primary key STORE\_ID of STORE\_DIM dimension |  |
| TIME\_ID | int | TIME\_DIM | TIME\_ID | Foreign key corresponding to primary key TIME\_ID of TIME\_DIM dimension |  |
| PRODUCT\_ID | int | PRODUCT\_DIM | PRODUCT\_ID | Foreign key corresponding to primary key PRODUCT\_ID of PRODUCT\_DIM dimension |  |
| MOVEMENT | int | movement files | move |  |  |
| QUANTITY | int | movement files | qty |  |  |
| SALES\_AMOUNT | float | movement files |  | SALES\_AMOUNT = Price \* Move / Qty for a given row |  |

### Business question justification corresponding to proposed schema

Please find the explanation of how we can answer all the five business question using the above two schemas below.

#### **BQ** **1: What is the trend of wine sales during Christmas holiday season?**



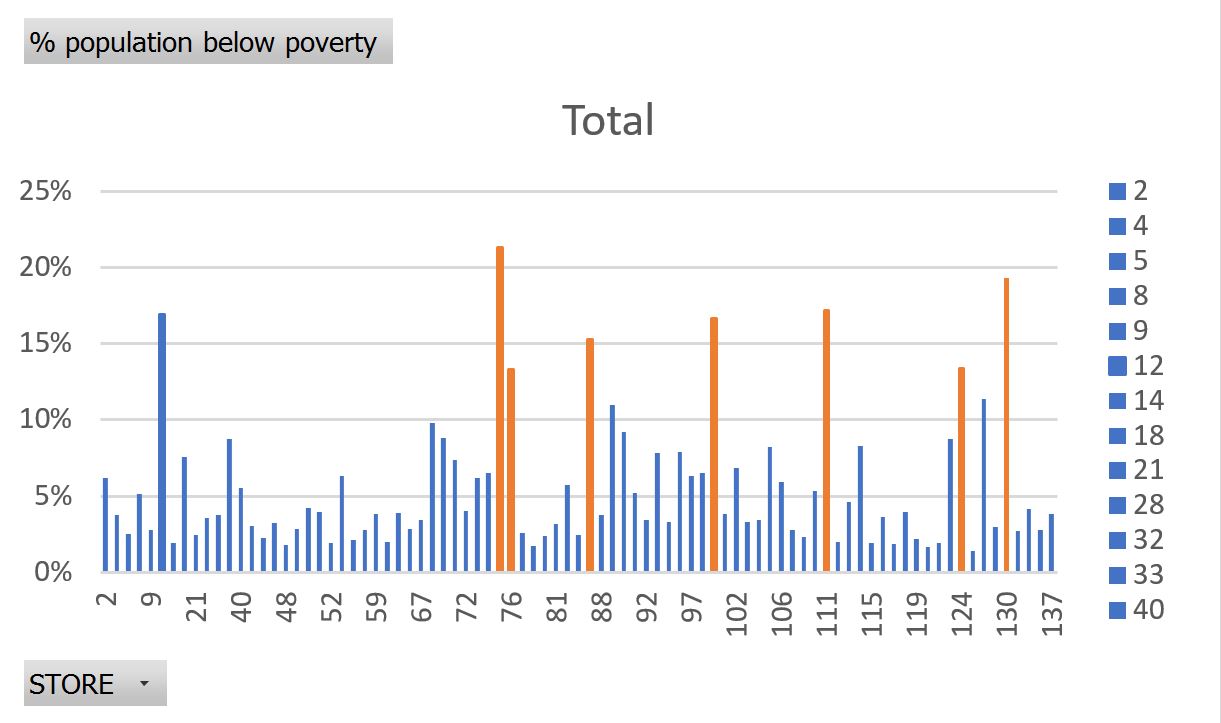
This business question is addressed by analyzing the Wine sales during the Christmas weeks as shown in the above graph. Same analysis can also be done with the Sales data mart shown above . Here the product under consideration is Wine, information related to that can be obtained from PRODUCT\_DIM dimension table and Week related information could be taken from WEEK attribute loaded in TIME\_DIM dimension table. Wine Sales related information corresponding to each week during Christmas week can be mapped from SALES\_FACT fact table in Sales data mart.

#### **BQ 2:** **Which stores have more popularity among kids and elderly groups?**



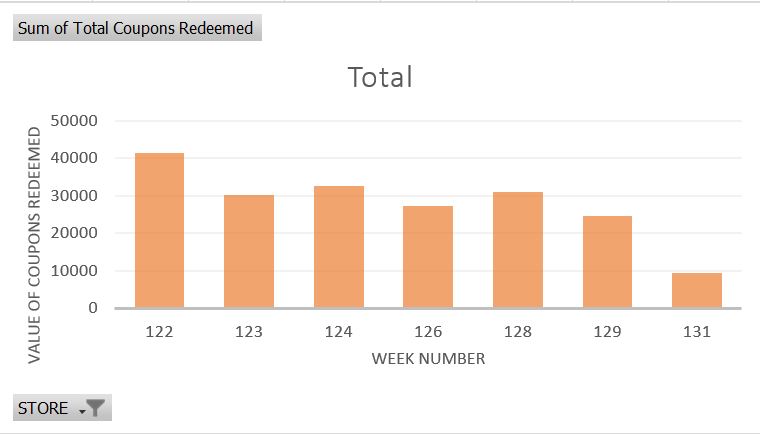
Store visits data mart can be used for analyzing this business question. Number of people below age 9 (BELOW\_9\_ %) and above 60(ABOVE\_60\_ %) are attributes in DEMOGRAPHIC\_DIM dimension table and Store Number plotted the above graph can be plotted from attribute STORE\_NUMBER from STORE\_DIM dimension table. Comparing the two values, the number of store visits, CUSTOMER\_COUNT is obtained from STORE\_VISITS\_FACT fact table. Above graph will help us analyze just the popularity among people above age 60. For answering this business question completely, popularity of stores among kids and elderly group will be analyzed separately in similar way.

#### **BQ 3:** **Which stores attract people who earn below poverty line?**



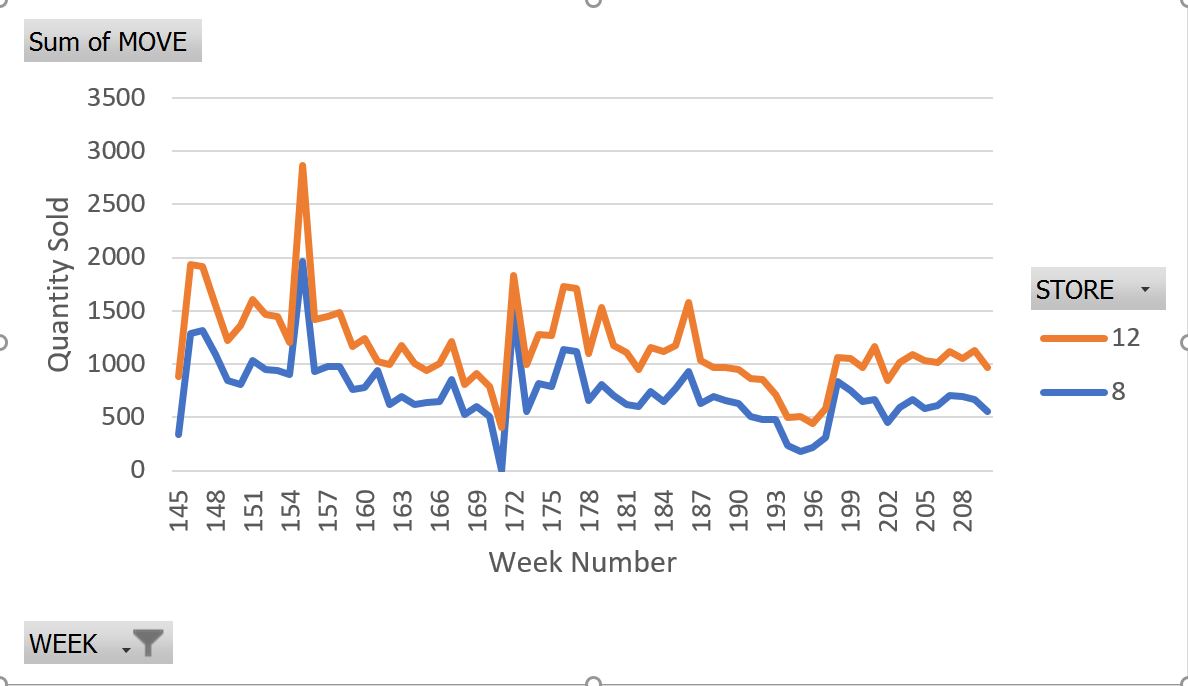
This business question can be analyzed similar to Business Question 2. Same Store visits data mart can be used for analysis. IS\_POOR attribute is available in DEMOGRAPHIC\_DIM table and store number STORE\_NUMBER is obtained from STORE\_DIM table. For each corresponding values of STORE\_NUM and IS\_POOR we can get the CUSTOMER\_COUNT from STORE\_VISITS\_FACT table given in the schema. This data can be used to analyze store attraction among people below poverty line.

#### **BQ 4:** **What is the effect of introducing coupons on total number of customer visits?**



This is comparatively complex business question to analyze even though it is easy to understand. For analyzing the question we need to plot and compare the Total number of customer visits and Total number of coupons redeemed on a weekly basis. Comparing the trend in both the graph’s we can reach at a conclusion of effect of coupons on customer visits. So here the Store visits data mart can be used again for analysis. CUSTOMER\_COUNT attribute from STORE\_VISITS\_FACT fact table can be plotted against the WEEK attribute loaded in TIME\_DIM dimension table form the schema. Similarly, for plotting the second graph we need calculate the value derived from the summation of coupon sales from different product category, which is given in CCount table. This summed up value is stored in COUPON\_DIM dimension table as COUPON\_REDEEMED attribute. COUPON\_REDEEMED attribute from COUPON\_DIM dimension table can be used for plotting the second graph. So now with the two graph’s plotted from the data loaded to dimension tables and fact tables we can easily find the effect of introducing coupons in total number of customer visits.

#### **BQ 5:** **What is the trend of a product demand in different price-tiers?**



This business question can be answered by plotting the weekly sales data of each Shop tiers (A,B.C) . For this Business question, we can use Sales data mart. In Sales data mart the PRICE\_TIER information is loaded in STORE\_DIM dimension table and WEEK related data is stored in TIME\_DIM dimension table. All the sales related information for all the shop price tiers can be collected from the SALES\_AMOUNT attribute in SALES\_FACT fact table. This information is plotted for each shop tiers and the distribution of sales in each shop tier can easily be analyzed from the plot.

### Conclusion

From the above explanation, it is clear that all selected business questions can be analyzed with 2 schema diagrams shown above. Both schemas include two fact tables - STORE\_VISITS\_FACT, SALES\_FACT and 5 dimension tables - STORE\_DIM, TIME\_DIM, PRODUCT\_DIM, DEMOGRAPHIC\_DIM, and COUPON\_DIM. Current fact tables, dimension tables and fact tables in each schema is designed to give faster performance. Still the performance can be increased further by using aggregate fact tables if needed. Need for the aggregate fact tables can be analyzed after properly loading the required data from all the 4 types of csv files to the fact and dimension tables designed in the current schemas and running queries on top of it. Need for the aggregation fact tables will be decided on the upcoming stages of the project.