MGS657 – OLAP and Data-warehousing
Project Report

Submitted by,

Apoorva John - 50204543 - ajohn4@buffalo.edu

Neeraja Narayan - 50208881 – neerajan@buffalo.edu

Preethy Thomas - 50204728 - preethyt@buffalo.edu

Table of Contents

1. INTRODUCTION

1.1 Document Purpose

The purpose of this document is to arrive at key KPI's involved in the functioning of the Sales domain of a large-scale retail chain store, to establish an OLAP system and a data warehouse that fetches its data from an OLTP system with the help of freely available ETL tools in the market, to do an analysis of the data in the OLAP data warehouse with the help of freely available visualization tools in the market and to arrive at conclusions that could help achieve the KPI's identified.

1.2 Intended Audience

The main intended audience of this report include the key stakeholders of a similar retail chain business like CEO's, VP's of Sales domain etc. who could make use of the technologies mentioned and conceptualize the ideas proposed at a larger scale to arrive at fruitful conclusions regarding their business. Other stakeholders also include data architects, data analysts, technical architects, business analysts etc., who could make use of the information available in this report to model potential BI solutions for retail chain companies.

1.3 Definition of Success

Success in this scenario is defined as providing end to end analysis of KPI's to come up with suggestions and form conclusions from the data set utilized, which in this example could be to forecast future sales for a month, to identify the top n customers by the quantities they order, sales they made and number of visits they made to store etc. that could help in arriving at fruitful conclusions for the growth and potential expansion of the store across multiple regions.

1.4 Definition of Failure

Failure in this scenario is defined as being unable to come up with an accurate prediction about the forecast of sales, accurate prediction about top n customers, top n products etc. This might occur due to the following reasons:

a. Incomplete and incorrect data
b. ETL issues
c. Incorrect method used for analysis yielding incorrect analysis.

1.5 Goals

The following are the goals related to the sales domain of the company for the upcoming year:

a. To understand the stores with good sales and to promote the sales of other stores using similar marketing strategy.
b. To predict the future sales of the company for the next month which can be then used to arrive at possible conclusions on procuring and management of stock.
c. To understand the status of the company over years based on their sales to see if there has been an improvement or not.
d. To understand the top customers of the company that could help in marketing the products accordingly to these customers.
e. To understand the top products of the company that could help in stocking more of the them, deciding marketing strategies for them, trying to improve their sales, understand why a particular product is being sold more and use a similar marketing tactic on the products that are being sold less, thus helping in improving the sales of all products in a holistic manner.
f. To understand if the company is losing out on their top customers and products by doing a year over year analysis.

1.6 Current Key Business Issues:

The following are the current key business issues faced by the company:

a. The presence of an immense amount of data related to sales, stored on OLTP systems, preventing the analysis while concurrent inserts and updates are happening due to which data is not being put to use.
b. Lack of a clear understanding about the stores or regions at which the company has low sales over years.
c. Lack of a clear understanding of the top customers who visit the company and purchase from the company.
d. Lack of a clear understanding of the top products that are sold by the company and the products that generate maximum sales revenue.
e. Lack of a clear understanding about the growth of sales over years.
f. Lack of a clear prediction of future sales.

1.7 Problem Definition:

Assess the sales data over years to come up with a prediction of sales for the coming months, and also to analyze the data to understand the top n customers and top n products for the company so that the company can decide its future marketing strategies. Assuming each lost top customer as an opportunity, the company can analyze its potential opportunities. Understanding the trend of sales across its different store sizes, the company can come up with custom marketing strategies for its top customers, top products etc.

1.8 Assumptions:

Data utilized is across 3 years viz., 2006, 2007 and 2008 and is not real data but mimics the real data.

2. PROJECT PLANNER

Task Name	Start	End	Duration (days)
Searching for Dataset	10/21/2016	10/25/2016	4
Defining the KPI's	10/25/2016	10/26/2016	1
Defining the Queries	10/25/2016	10/26/2016	1
Analysis of raw data	10/27/2016	11/2/2016	6
Creation of master data	11/23/2016	11/25/2016	2
Creation of OLTP Schema	11/25/2016	11/27/2016	2
Data load to OLTP schema	11/27/2016	11/28/2016	1
Creation of OLAP schema	11/29/2016	11/29/2016	0
ETL to load data to OLAP schema	11/30/2016	12/3/2016	3
Reporting and analysis in Tableau	12/3/2016	12/7/2016	4
Final report creation	12/15/2016	12/16/2016	1

3. BUS MATRIX

Dimension / Business Process	Product	Store	Customer	Date
Sales	X	X	X	X
Product Stratification	X			X
Customer Stratification			X	X
Sales Forecast		X		X

4. KPIS

The following are the proposed KPI's that could help improve the sales of the retail store chain.

a. Sales maximization
b. Strengthening customer base
c. Product catalog design

5. RAW DATA

5.1 Transaction Data

Raw Data Set on retail store sales was taken from the website of Dunnhumby. The link is as follows: https://www.dunnhumby.com/sourcefiles. The following are the details of the data from the dataset.

Column name	Description	Type	Sample values
shop_week	Identifies the week of the basket	Char	Format is YYYYWW where the first 4 characters identify the fiscal year and the other two characters identify the specific week within the year (e.g. 200735). Being the fiscal year, the first week doesn't start in January. (See time.csv file for start/end dates of each week)
shop_date	Date when shopping has been made. Date is specified in the yyyymmdd format	Char	20060413, 20060 412
shop_weekday	Identifies the day of the week	Num	1=Sunday, 2=Monday, …, 7=Saturday
shop_hour	Hour slot of the shopping	Num	0=00:00-00:59, 1=01:00-01:59, …23=23:00-23:59
Quantity	Number of items of the same product bought in this basket	Num	Integer number
spend	Spend associated to the items bought	Num	Number with two decimal digits
prod_code	Product Code	Char	PRD0900001, PRD0900003
prod_code_10	Product Hierarchy Level 10 Code	Char	CL00072, CL00144
prod_code_20	Product Hierarchy Level 20 Code	Char	DEP00021, DEP00051
prod_code_30	Product Hierarchy Level 30 Code	Char	G00007, G00015
prod_code_40	Product Hierarchy Level 40 Code	Char	D00002, D00003
cust_code	Customer Code	Char	CUST0000001624, CUST0000001912
cust_price_sensitivity	Customer's Price Sensitivity	Char	LA=Less Affluent, MM=Mid Market, UM=Up Market, XX=unclassified

Column name	Description	Type	Sample values
cust_lifestage	Customer's Lifestage	Char	YA=Young Adults, OA=Older Adults, YF=Young Families, OF=Older Families, PE=Pensioners, OT=Other, XX=unclassified
basket_id	Basket ID. All items in a basket share the same basket_id value.	Num	994100100000020, 994100100000344
basket_size	Basket size	Char	L=Large, M=Medium, S=Small
basket_price_sensitivity	Basket price sensitivity	Char	LA=Less Affluent, MM=Mid Market, UM=Up Market, XX=unclassified
basket_type	Basket type	Char	Small Shop, Top Up, Full Shop, XX
basket_dominant_mission	Shopping dominant mission	Char	Fresh, Grocery, Mixed, Non Food, XX
store_code	Store Code	Char	STORE00001, STORE00002
store_format	Format of the Store	Char	LS, MS, SS, XLS
store_region	Region the store belongs to	Char	E02, W01, E01, N03

The transaction data obtained from the website was in the form of .csv files with 117 files corresponding to 117 weeks from 200607 to 200806 having around 2 million records.

5.2 Master Data

After further analysis of the raw dataset, we could understand that there were 4000+ unique product codes and customer codes in the dataset.

Master data like name, etc. was made up for both the customers and products and maintained in separately created .csv files.

5.2.1 Customer Master Data

The following are the details about the metadata of customer master data that was randomly populated in excel to sync with the existing transaction data:

Column name	Description	Type	Sample values
First Name	First Name of the customer	Char	Gale, Henry etc.
Last Name	Last Name of the customer	Char	Krom, Hennis etc.
Sex	Gender of the customer	Char	f or m
Phone	Primary phone number of the customer	Char	740-436-3571, 937-478-1091 etc.
Birthday	Birthday of the customer	Date	2/4/1993, 1/26/1950 etc.
E-mail address	Primary e-mail address of the customer	Char	Gale.M.Krom@pookmail.com, etc
Address	Primary address of the customer	Char	1000 CALLE H etc
State Code	State code of the primary address	Char	PR, NJ etc
City	City of the customers primary address	Char	CAGUAS, TINTON FALLS etc.
Postal Code	Postal code of the customers primary address	Char	725, 7727 etc.
Region	Store Region at which the customer's primary address belongs to.	Char	E, N, S, W
Price Sensitivity	Reactivity of customer to slight fluctuations in price of products.	Char	HIGH, MEDIUM, LOW
Customer Code	The unique identifier number given to the customers.	Char	CUST0000977889, CUST0000346601 etc.

5.2.2 Item Master Data

The following are the details about the metadata of item master data that was randomly populated in excel to sync with the existing transaction data:

Column name	Description	Type	Sample values
ITM_KEY	Unique Key of the Item	Number	477947, 667886 etc.
ITM_NBR	Unique number of item	Number	5201603, 5927298 etc.

CATGY_ID	Category ID of the item.	Number	1, 2 etc
MAJ_ID	Category Major ID of the item.	Number	21, 22 etc.
MNR_ID	Category Minor ID of the item.	Number	32, 75 etc.
BRND_CD	Brand code of the item.	Char	SANJMAR, USSURG etc.
ITM_STS_CD	Status code of the item indicating if item is active or inactive.	Char	A or I
ITM_DESC	Description of Item.	Char	GLOVE CUT RESISTANT, ENDO MESH 3X5 etc
ITM_EXT_DESC	Extended Description of Item	Char	GASTROSTOMY TUBE etc.
PCK_DESC	Pack description of item	Number	1,3, 300 etc
SZ_DESC	Size description of item.	Char	MEDIUM, EA, etc
HAZ_IND	Indicator which states if Item is hazardous or not.	Char	Y or N
CATCH_WGT_IND	Catch weight indicator of item if item is sold in pounds or cases. If catch weight indicator is 'Y', item is sold in pounds. If its 'N' its sold in cases.	Char	Y or N
CHLD_NUTR_ITM_IND	It states if Item helps in child nutrition.	Char	Y or N
UNIT_PR_CASE_QTY	Explains how many units of item are there in one case.	Char	1, 3, 300 etc
ITM_STS_EFF_DT	Explains when the item status was turned to active or inactive.	Date	4/13/2004, etc.
ITM_NET_WGT_VAL	Net weight value of item.	Number	0.14, 27.94 etc.
ITM_GRS_WGT_VAL	Gross weight value of item.	Number	0.14, 27 etc.
CUBE_SZ_VAL	Cube size value of item.	Number	0.2, 0.06 etc
ITM_STRG_CD	Storage Code of the item which indicates if item has to be stored in dry or frozen or cold condition.	Char	D, F, C etc.
UNIT_PRICE	Unit price of the item per case.	Number	15, 9, 8, 3 etc.

5.2.3 Item CMIM Data:

The following are the details about the metadata of Item CMIM viz., Item Category Major, Intermediate and Minor that was randomly populated in excel to sync with the created Item and existing transaction data:

Column name	Description	Type	Sample values
CATGY_ID	Category ID of the item.	Number	1, 2 etc
MAJ_ID	Category Major ID of the item.	Number	21, 22 etc.
MNR_ID	Category Minor ID of the item.	Number	32, 75 etc.
INTRM_ID	Category Intermediate ID of the item.	Number	78, 99 etc.
CATGY_DESC	Category Description of the item	Char	PRODUCE, SUPP EQUIPMENT etc.
MAJ_DESC	Major category Description of the Item	Char	VEGETABLE FRESH, CONTAINER PANS etc.
MNR_DESC	Minor category Description of the Item	Char	USA OLIVE OIL, HOT DRINK etc.
INTRM_DESC	Intermediate category description of the item	Char	MUSHROOMS, SNACK/CEREAL BAR etc

5.2.4 Item Brand Data:

The following are the details about the metadata of Item Brand that was randomly populated in excel to sync with the created Item and existing transaction data:

Column name	Description	Type	Sample values
BRND_ID	Brand ID of the Item	Char	PAINFRN, etc.
BRND_DESC	Brand Description of the Item	Char	PAINFRANCE (PASTRY) etc.

5.2.5 Store Data:

Store data was derived from the existing transaction data by pulling out just the store code, store format and store region fields.

Column name	Description	Type	Sample values
store_code	Store Code	Char	STORE00001, STORE00002
store_format	Format of the Store	Char	LS, MS, SS, XLS
store_region	Region the store belongs to	Char	E02, W01, E01, N03

6. OLTP SYSTEM:

An OLTP database was created in Oracle EE 12.1.0.2 database. The database was hosted in AWS and the end point and connection details of the database is as follows:

End point: olapproject.cfagn0ajamif.us-east-1.rds.amazonaws.com

User ID: Admin

Password: pa55w0rd

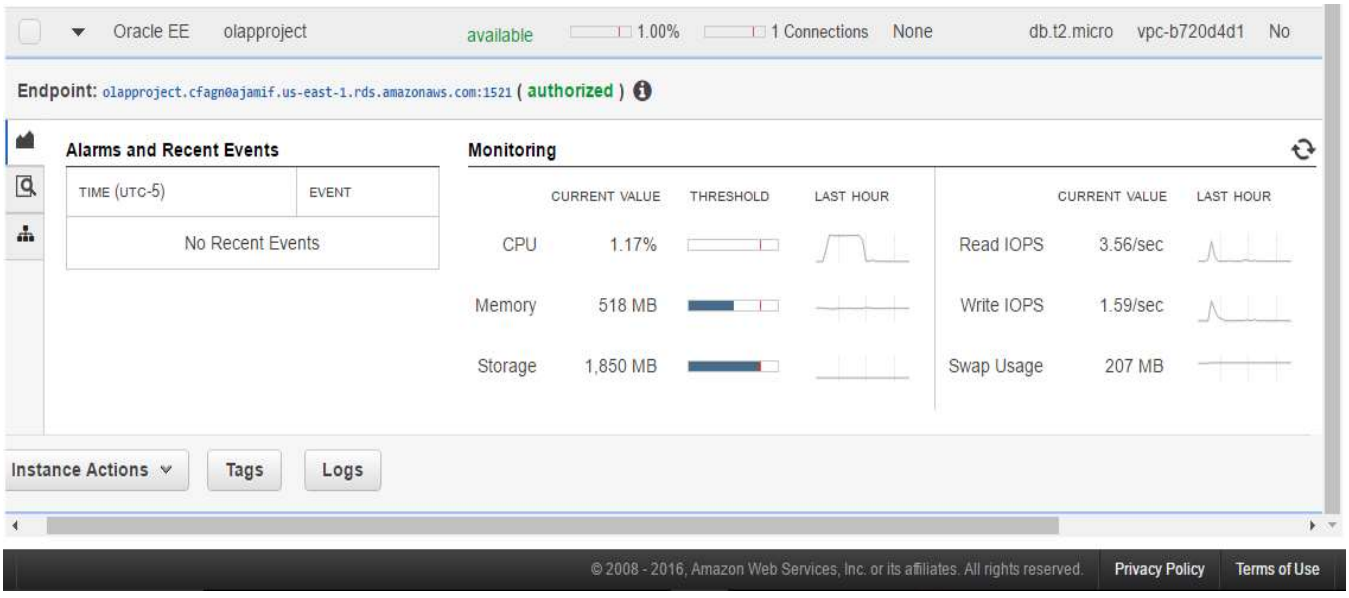
DB Name: ORCL

Database was not configured to have high availability and automated snapshots.

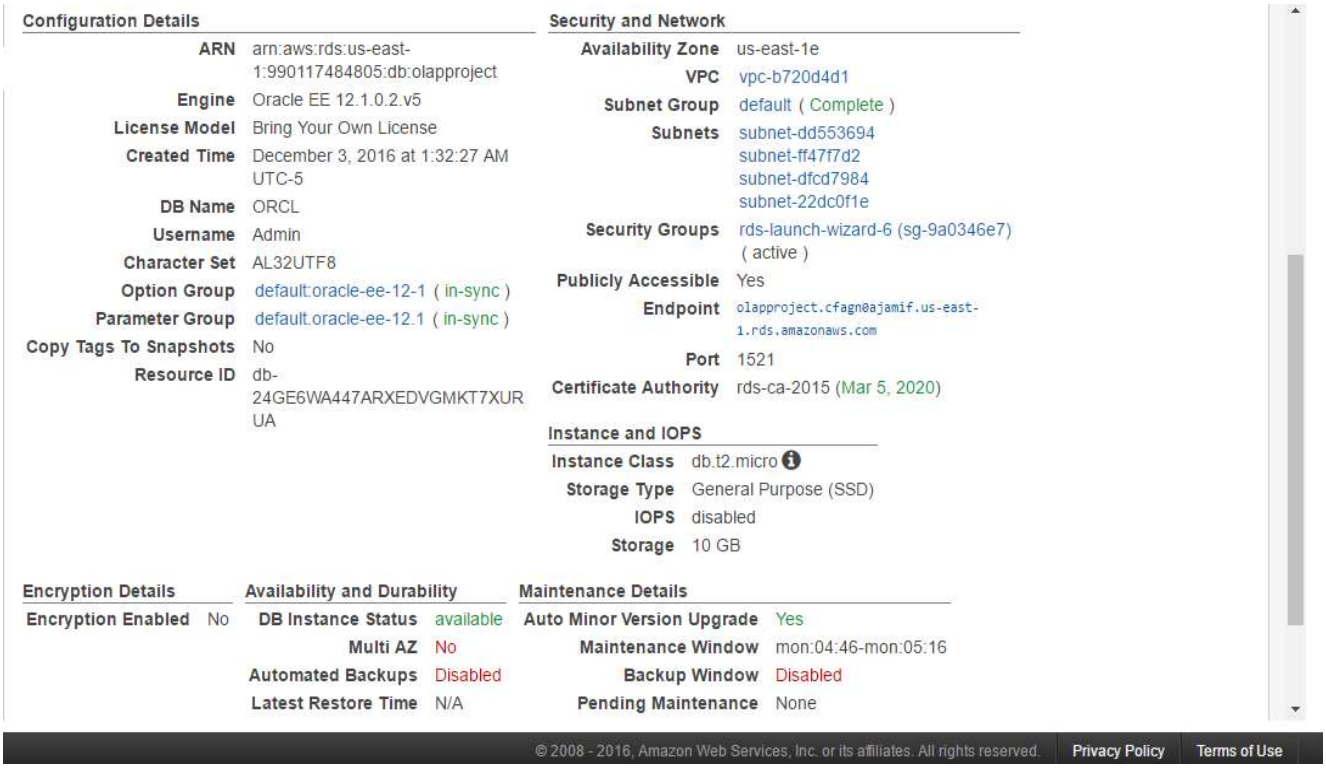
A storage of 1850 MB was provisioned and 518 MB was utilized by the data.

Database was also configured to be publicly accessible.

The following is the screenshot showing the instance storage available and monitoring from the amazon web services web page:



The following is the screenshot of the instance properties from AWS RDS:



Building of OLTP DB was taken into consideration due to the following two points:

a. Difficulty in manipulation of huge chunks of data in excel.
b. Portray an end-to-end flow in any OLAP system starting from an OLTP DB ending with an OLAP DW involving an ETL tool to transform and load the data.

The data in OLTP schema was stored in the following tables:

6.1 Customer:

The metadata of customer table is as follows:

COLUMN_NAME	DATA_TYPE	NULLABLE
FIRST_NAME	VARCHAR2(20 BYTE)	Yes
LAST_NAME	VARCHAR2(20 BYTE)	Yes
SEX	VARCHAR2(2 BYTE)	Yes
PHONE	VARCHAR2(20 BYTE)	Yes
BIRTHDAY	DATE	Yes

EMAIL_ADDRESS	VARCHAR2(70 BYTE)	Yes
ADDRESS	VARCHAR2(70 BYTE)	Yes
STATE_CD	VARCHAR2(2 BYTE)	Yes
CITY	VARCHAR2(60 BYTE)	Yes
POSTAL_CD	NUMBER(38,0)	Yes
REGION	VARCHAR2(1 BYTE)	Yes
CUST_CODE	VARCHAR2(15 BYTE)	No
PRICE_SENSITIVITY	VARCHAR2(20 BYTE)	Yes

6.2 Item

The metadata of item table is as follows:

COLUMN_NAME	DATA_TYPE	NULLABLE
ITM_KEY	NUMBER(38,0)	No
ITM_REC_EFF_DT	DATE	Yes
ITM_REC_TRM_DT	DATE	Yes
ITM_NBR	NUMBER	Yes
CATGY_ID	NUMBER	Yes
MAJ_ID	NUMBER	Yes
INTRM_ID	NUMBER	Yes
MNR_ID	NUMBER	Yes
BRND_CD	VARCHAR2(25 BYTE)	Yes
ITM_STS_CD	VARCHAR2(3 BYTE)	Yes
ITM_DESC	VARCHAR2(70 BYTE)	Yes
ITM_EXT_DESC	VARCHAR2(70 BYTE)	Yes
PCK_DESC	VARCHAR2(10 BYTE)	Yes
SZ_DESC	VARCHAR2(10 BYTE)	Yes
HAZ_IND	VARCHAR2(3 BYTE)	Yes
CATCH_WGT_IND	VARCHAR2(3 BYTE)	Yes
CHLD_NUTR_ITM_IND	VARCHAR2(3 BYTE)	Yes
UNIT_PR_CASE_QTY	NUMBER	Yes
ITM_STS_EFF_DT	DATE	Yes

ITM_NET_WGT_VAL	NUMBER	Yes
ITM_GRS_WGT_VAL	NUMBER	Yes
CUBE_SZ_VAL	NUMBER	Yes
ITM_STRG_CD	VARCHAR2(3 BYTE)	Yes
CURR_REC_IND	VARCHAR2(3 BYTE)	Yes
UNIT_PRICE	NUMBER	Yes

6.3 Brand

The metadata of brand table is as follows:

COLUMN_NAME	DATA_TYPE	NULLABLE
BRND_CD	VARCHAR2(20 BYTE)	No
BRND_DESC	VARCHAR2(70 BYTE)	Yes

6.4 Item CMIM

The metadata of CMIM table is as follows:

COLUMN_NAME	DATA_TYPE	NULLABLE
CATGY_ID	NUMBER	No
ITM_CATGY_DESC	VARCHAR2(70 BYTE)	Yes
MAJ_ID	NUMBER	No
ITM_MAJ_DESC	VARCHAR2(70 BYTE)	Yes
INTRM_ID	NUMBER	No
ITM_INTRM_DESC	VARCHAR2(70 BYTE)	Yes
MNR_ID	NUMBER	No
ITM_MNR_DESC	VARCHAR2(70 BYTE)	Yes

6.5 Sales Order:

The metadata of Sales Order table are as follows:

COLUMN_NAME	DATA_TYPE	NULLABLE
SHOP_DATE	NUMBER	Yes
SHOP_HOUR	NUMBER	Yes
CUST_CODE	VARCHAR2(15 BYTE)	Yes
STORE_CODE	VARCHAR2(11 BYTE)	Yes
ORDER_ID	VARCHAR2(70 BYTE)	No

6.6 Sales Order Line

The metadata of Sales Order Line table is as follows:

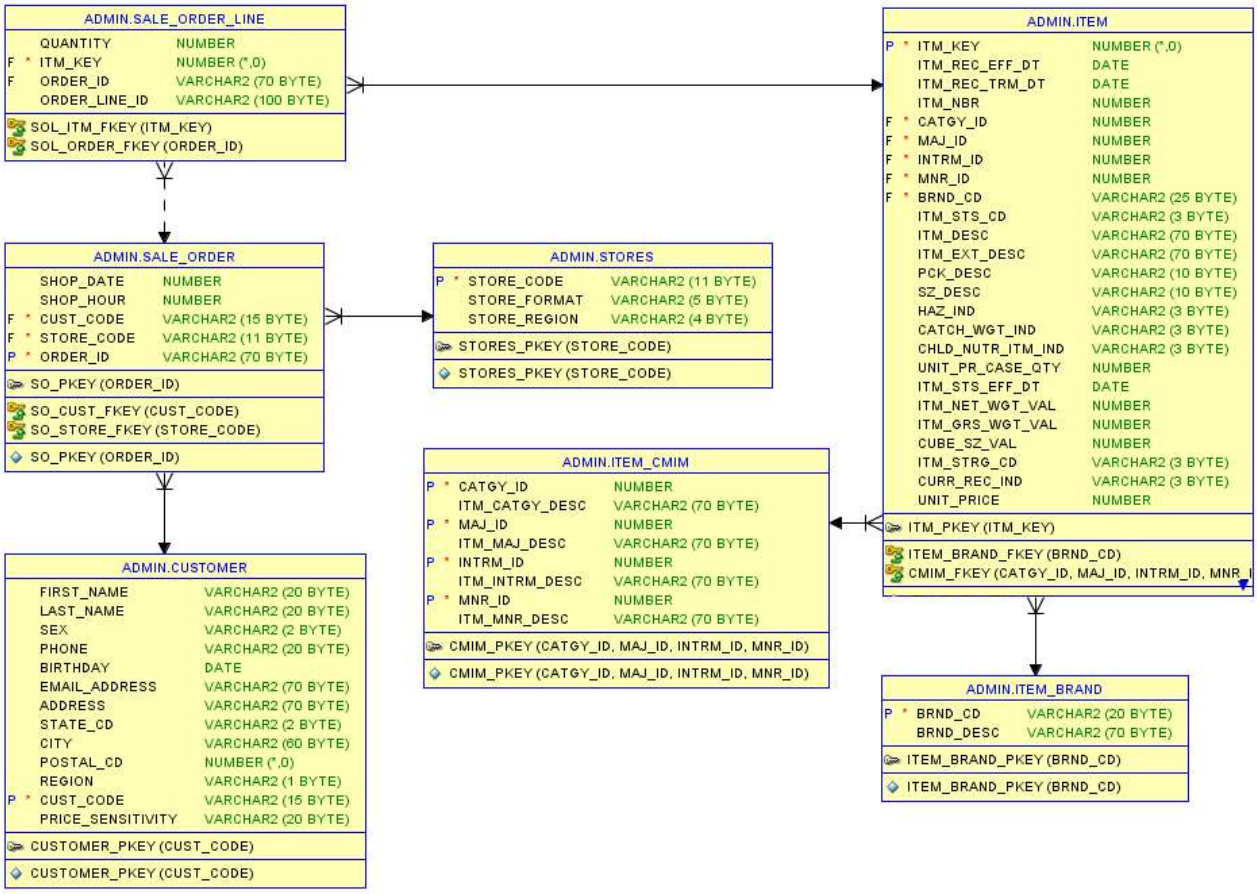
COLUMN_NAME	DATA_TYPE	NULLABLE
QUANTITY	NUMBER	Yes
ITM_KEY	NUMBER	Yes
ORDER_ID	VARCHAR2(70 BYTE)	Yes
ORDER_LINE_ID	VARCHAR2(100 BYTE)	Yes

6.7 Store Data

The metadata of Store data is as follows:

COLUMN_NAME	DATA_TYPE	NULLABLE
STORE_CODE	VARCHAR2(11 BYTE)	No
STORE_FORMAT	VARCHAR2(5 BYTE)	Yes
STORE_REGION	VARCHAR2(4 BYTE)	Yes

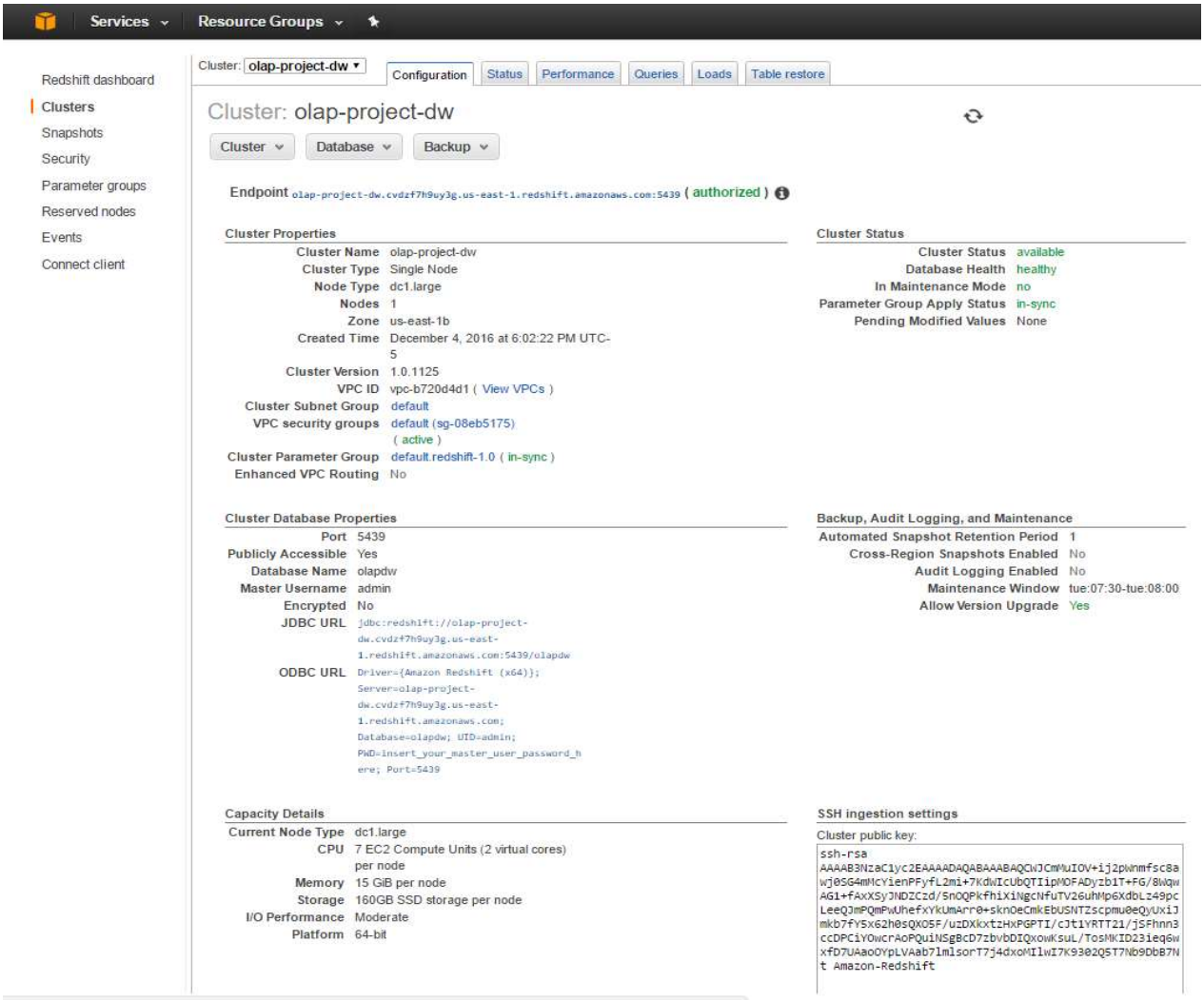
The E-R diagram of the OLTP database is as follows:


ADMIN.SALE_ORDER_LINE
QUANTITY NUMBER
F * ITM_KEY NUMBER (*,0)
F ORDER_ID VARCHAR2 (70 BYTE)
ORDER_LINE_ID VARCHAR2 (100 BYTE)
SOL_ITM_FKEY (ITM_KEY)
SOL_ORDER_FKEY (ORDER_ID)
ADMIN.ITEM
P * ITM_KEY NUMBER (*,0)
ITM_REC_EFF_DT DATE
ITM_REC_TRM_DT DATE
ITM_NBR NUMBER
F * CATGY_ID NUMBER
F * MAJ_ID NUMBER
F * INTRM_ID NUMBER
F * MNR_ID NUMBER
F * BRND_CD VARCHAR2 (25 BYTE)
ITM_STS_CD VARCHAR2 (3 BYTE)
ITM_DESC VARCHAR2 (70 BYTE)
ITM_EXT_DESC VARCHAR2 (70 BYTE)
PCK_DESC VARCHAR2 (10 BYTE)
SZ_DESC VARCHAR2 (10 BYTE)
HAZ_IND VARCHAR2 (3 BYTE)
CATCH_WGT_IND VARCHAR2 (3 BYTE)
CHLD_NUTR_ITM_IND VARCHAR2 (3 BYTE)
UNIT_PR_CASE_QTY NUMBER
ITM_STS_EFF_DT DATE
ITM_NET_WGT_VAL NUMBER
ITM_GRS_WGT_VAL NUMBER
CUBE_SZ_VAL NUMBER
ITM_STRG_CD VARCHAR2 (3 BYTE)
CURR_REC_IND VARCHAR2 (3 BYTE)
UNIT_PRICE NUMBER
ITM_PKEY (ITM_KEY)
ITEM_BRAND_FKEY (BRND_CD)
CMIM_FKEY (CATGY_ID, MAJ_ID, INTRM_ID, MNR_I
ADMIN.SALE_ORDER
SHOP_DATE NUMBER
SHOP_HOUR NUMBER
F * CUST_CODE VARCHAR2 (15 BYTE)
F * STORE_CODE VARCHAR2 (11 BYTE)
P * ORDER_ID VARCHAR2 (70 BYTE)
SO_PKEY (ORDER_ID)
SO_CUST_FKEY (CUST_CODE)
SO_STORE_FKEY (STORE_CODE)
SO_PKEY (ORDER_ID)
ADMIN.STORES
P * STORE_CODE VARCHAR2 (11 BYTE)
STORE_FORMAT VARCHAR2 (5 BYTE)
STORE_REGION VARCHAR2 (4 BYTE)
STORES_PKEY (STORE_CODE)
STORES_PKEY (STORE_CODE)
ADMIN.ITEM_CMIM
P * CATGY_ID NUMBER
ITM_CATGY_DESC VARCHAR2 (70 BYTE)
P * MAJ_ID NUMBER
ITM_MAJ_DESC VARCHAR2 (70 BYTE)
P * INTRM_ID NUMBER
ITM_INTRM_DESC VARCHAR2 (70 BYTE)
P * MNR_ID NUMBER
ITM_MNR_DESC VARCHAR2 (70 BYTE)
CMIM_PKEY (CATGY_ID, MAJ_ID, INTRM_ID, MNR_ID)
CMIM_PKEY (CATGY_ID, MAJ_ID, INTRM_ID, MNR_ID)
ADMIN.CUSTOMER
FIRST_NAME VARCHAR2 (20 BYTE)
LAST_NAME VARCHAR2 (20 BYTE)
SEX VARCHAR2 (2 BYTE)
PHONE VARCHAR2 (20 BYTE)
BIRTHDAY DATE
EMAIL_ADDRESS VARCHAR2 (70 BYTE)
ADDRESS VARCHAR2 (70 BYTE)
STATE_CD VARCHAR2 (2 BYTE)
CITY VARCHAR2 (60 BYTE)
POSTAL_CD NUMBER (*,0)
REGION VARCHAR2 (1 BYTE)
P * CUST_CODE VARCHAR2 (15 BYTE)
PRICE_SENSITIVITY VARCHAR2 (20 BYTE)
CUSTOMER_PKEY (CUST_CODE)
CUSTOMER_PKEY (CUST_CODE)
ADMIN.ITEM_BRAND
P * BRND_CD VARCHAR2 (20 BYTE)
BRND_DESC VARCHAR2 (70 BYTE)
ITEM_BRAND_PKEY (BRND_CD)
ITEM_BRAND_PKEY (BRND_CD)

7 OLAP Datawarehouse

For the OLAP DW we decided to implement it on Amazon Redshift. It is designed with the ability to handle analytics workloads on large scale datasets. It is based on PostgreSQL 8.0.2 and is able to handle connections to a variety of BI and data integration tools.

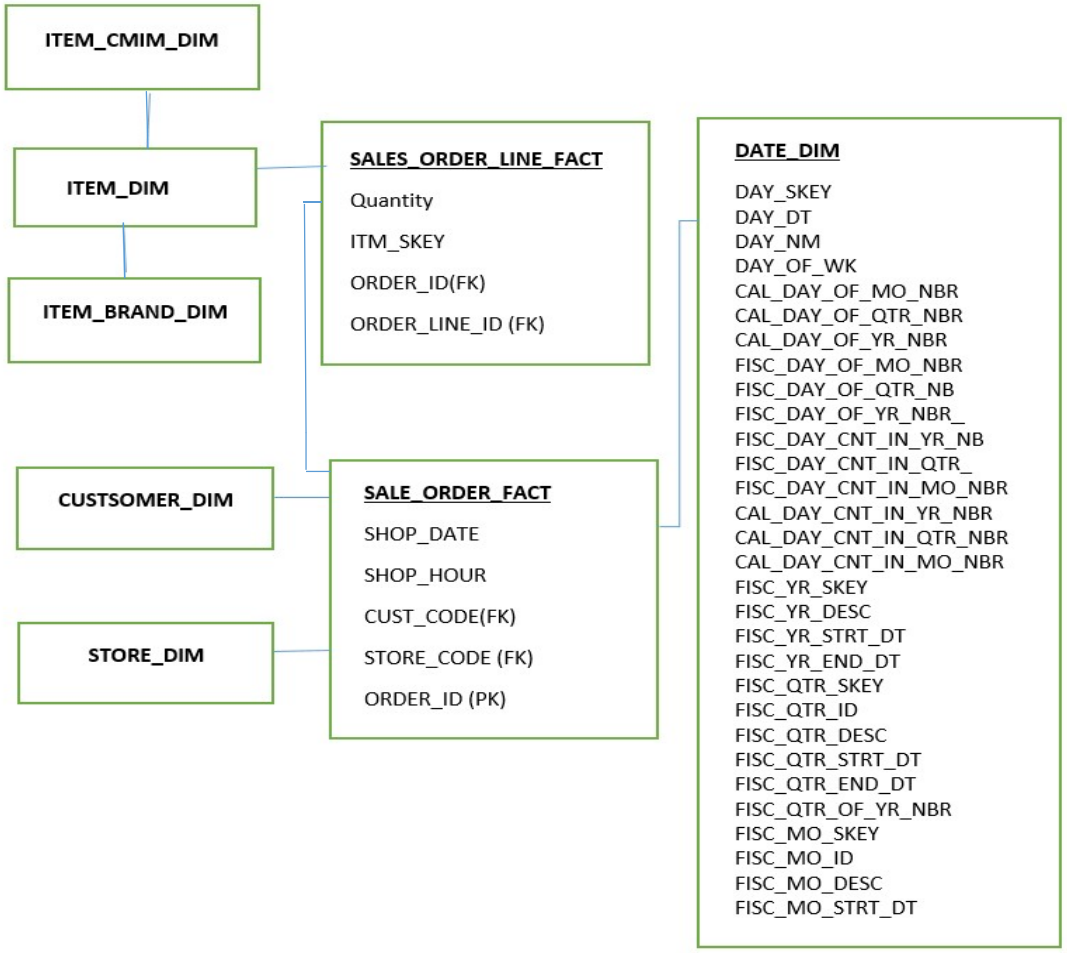
The details of the datawarehouse created are as below.
End point: olap-project-dw.cvdzf7h9uy3g.us-east-1.redshift.amazonaws.com:5439
User ID: admin
Password: Pa55w0rd
DB Name: olapdw
Database was configured on a single node cluster type and was automated to have snapshots with a retention period of 1 day.



After setting this up, we created the OLAP fact and dimension tables using Aginity Workbench for Redshift as the SQL database development tool.

8 OLAP Schema

The multidimensional OLAP database is optimized for snowflake schema. The tables are connected as below.



The two fact tables Salesorder_Fact and Sales_Order_Line_Fact are linked based one the order_ID. The shop_date of the sales_order_line_fact table is linked with the day_dt of the Date dimension whereas the itm_skey is mapped to the item_skey in the item_dim. The item_dim is further linked to item_brand_dim based on the key brand code which is brnd_cd. It is also linked to item_cmim_dim based on the key category ID i.e. catgy_id. Similarly, the other fact table sale_order_fact is mapped to the customer dimension and store dimension based on cust_code and store_code respectively.

9 ETL method

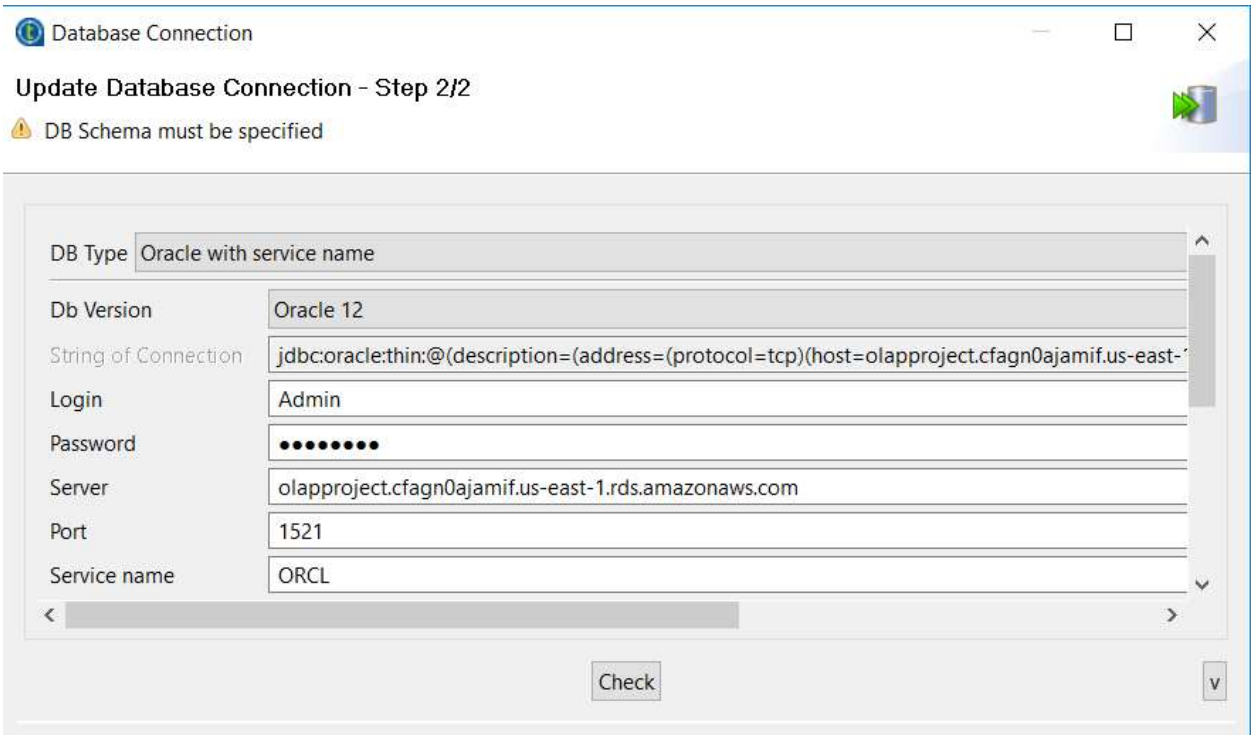
Extraction, Transformation and Loading that is ETL processes are one of the most critical components for feeding a data warehouse. While mostly invisible to users of a business intelligence platform, an ETL process retrieves data from operational systems and pre-processes it for further analysis by reporting and analytics tools.

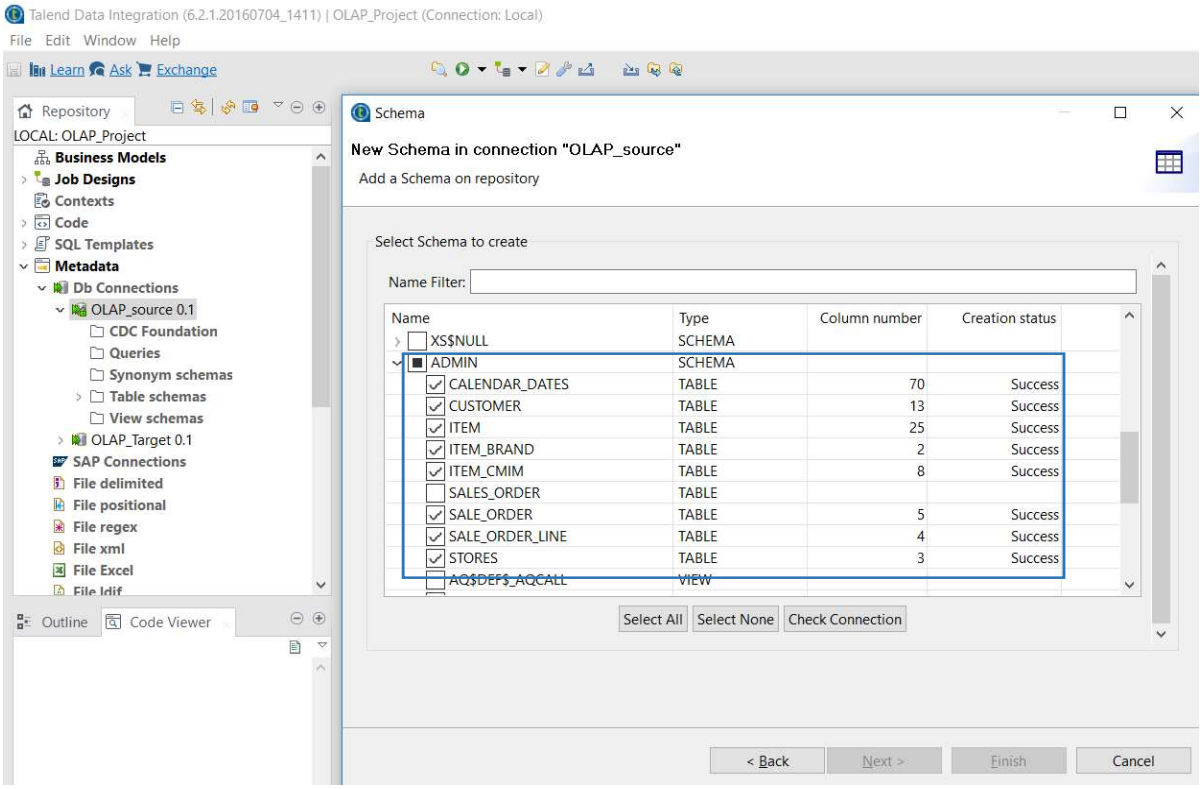
ETL can be used to acquire a temporary subset of data for reports or other purposes, or a more permanent data set may be acquired for other purposes such as, the population of a data mart or data warehouse; conversion from one database type to another; and the migration of data from one database or platform to another. The accuracy and timeliness of the entire business intelligence platform relies on the ETL processes.

For our project implementation, we adopted the ETL tool Talend Open Studio for Data Integration. The tool has an easy and simple database connectivity with the OLTP database we opted for i.e. Oracle and also the OLAP database which was Amazon Redshift.

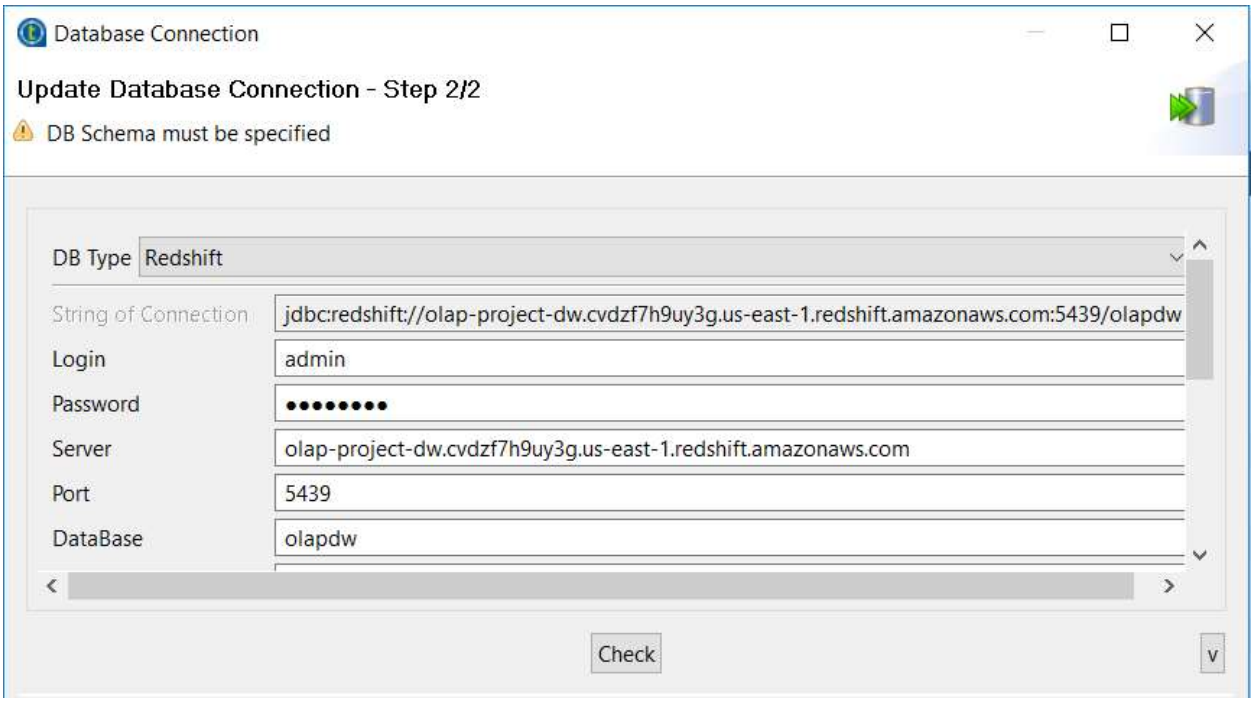
9.1 Extract

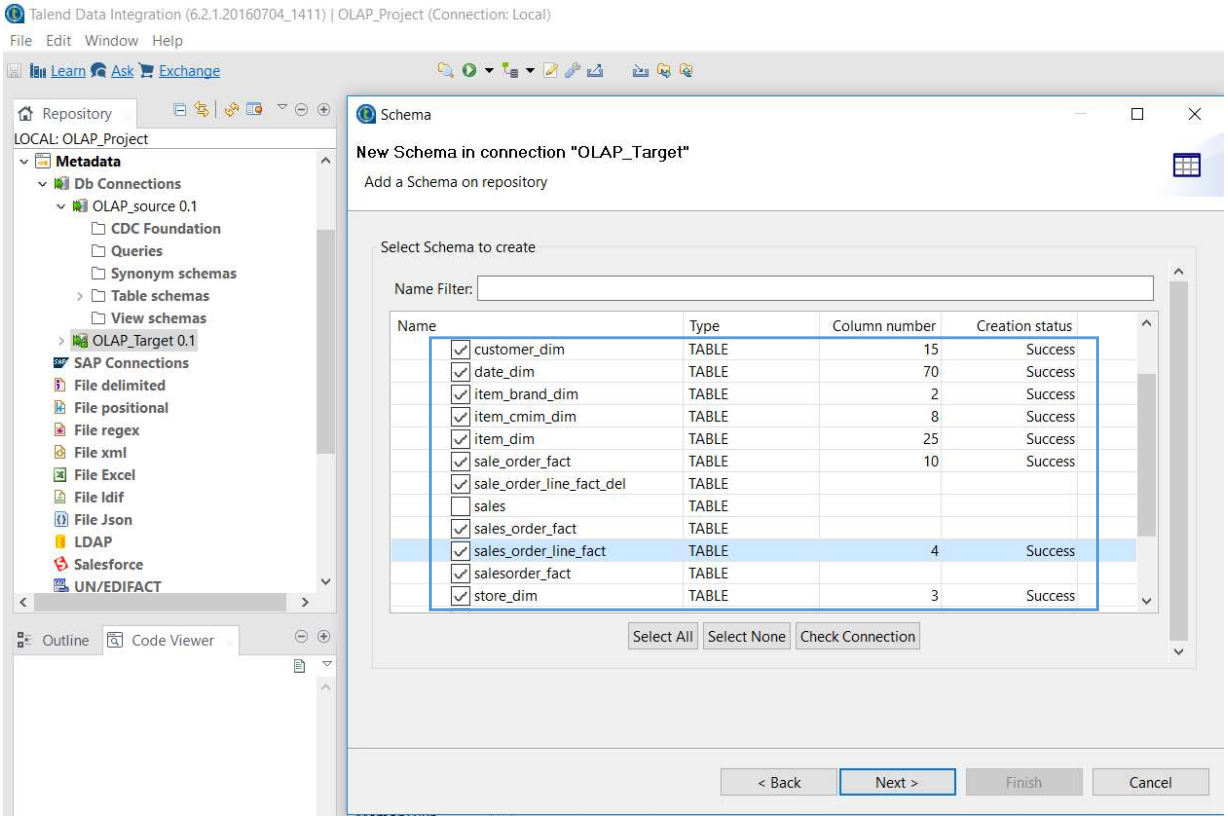
The first part of an ETL process involves extraction of the data from production applications and databases (ERP, CRM, RDBMS, files, etc.). Here, we had to extract data from our OLTP database, for which first the database connections are set up. Following this the schema is retrieved, only the tables necessary for further processing is selected.





The same needs to be done for the target database to enable the load process. Here we establish connection with the AWS Redshift and retrieve the schema.



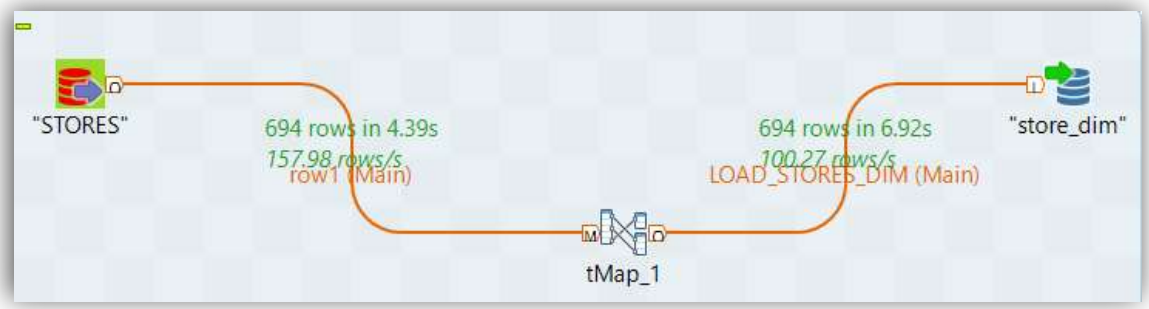


The data had to be now cleaned and converted into the necessary format before loading it into Redshift.

9.2 Transform

The transform function works with the acquired data using rules or lookup tables, or creating combinations with other data to convert it to the desired state. We had to majorly manipulate the data to make sure the data types were similar to the target database's tables.
For each transformation in Talend we had to create a Standard Job Design. The following transformations were made for the six dimension and the two fact:

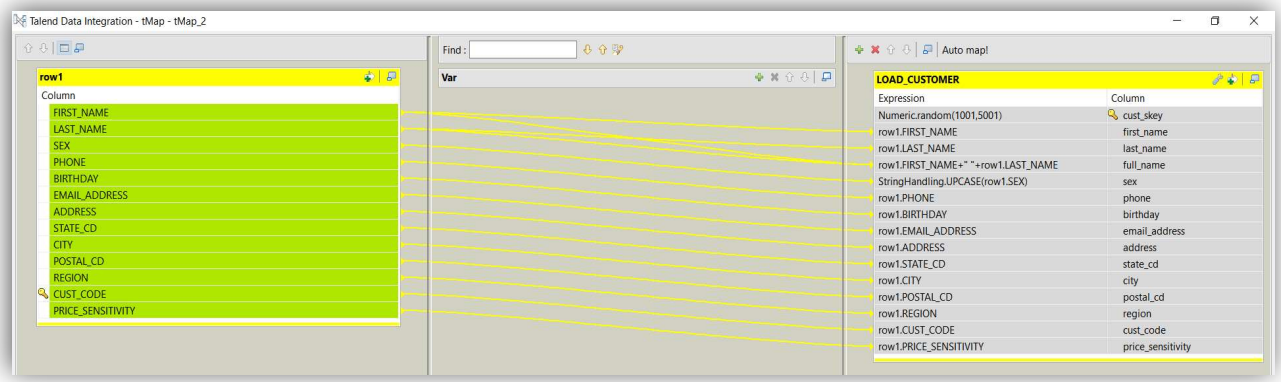
9.2.1 Store Dimension



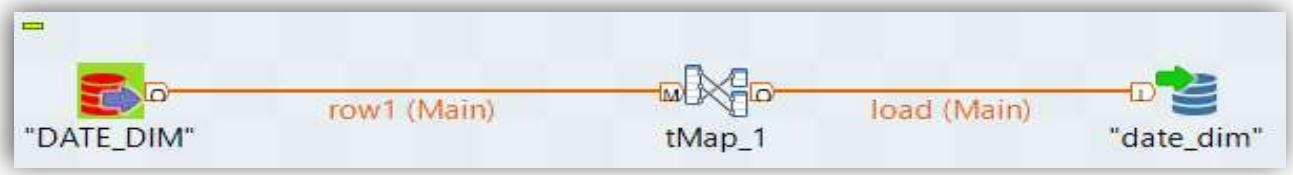
The transformation involved a simple one to one mapping to the three attributes store code, store format and store region.

9.2.2 Customer Dimension

The transformation mapping here involved conversion to uppercase of the values of the attribute sex and populating the new field Full_Name by concatenating the first and second names. We also added a customer key here by randomly generating numbers for each row.

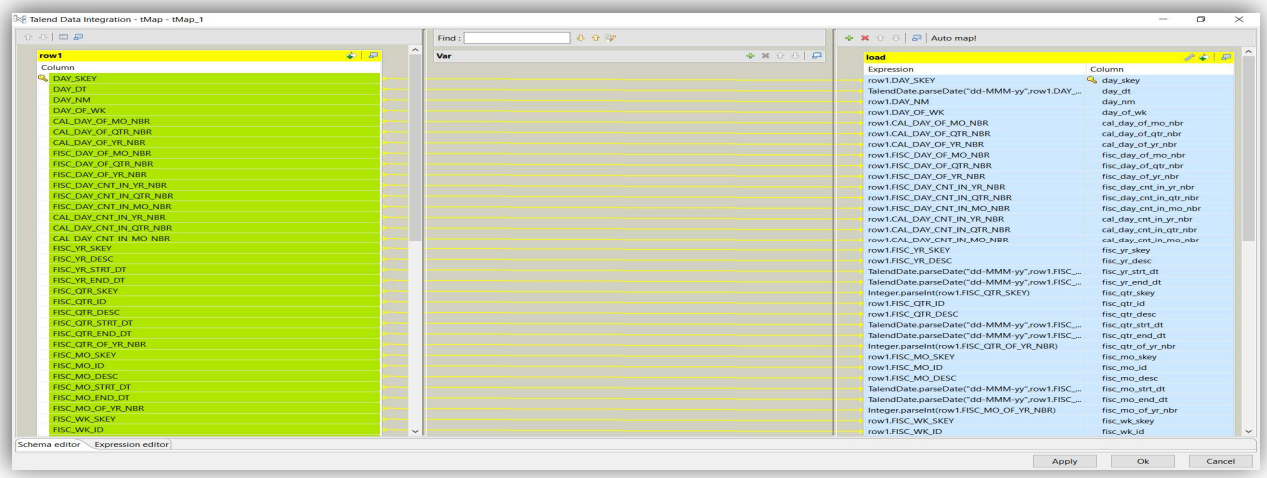


9.2.3 Date Dimension

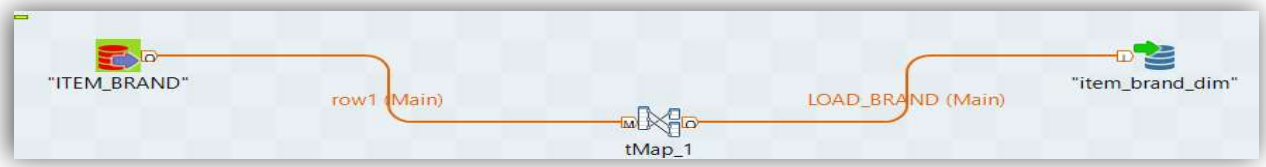


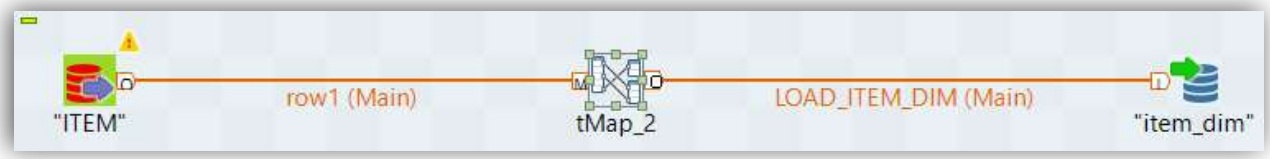
The transformation mapping here involved data type conversions, so that all the dates are in the DD – MM – YY date format. We also performed some conversions from string data type to integer using the talend conversion commands as below:

```
TalendDate.parseDate("dd-MMM-yy",row1.CAL_QTR_STRT_DT)
TalendDate.parseDate("dd-MMM-yy",row1.CAL_QTR_END_DT)
Integer.parseInt(row1.CAL_QTR_OF_YR_NBR)
```



9.2.4 Item Dimension/ Item Brand Dimension/ Item CMIM Dimension



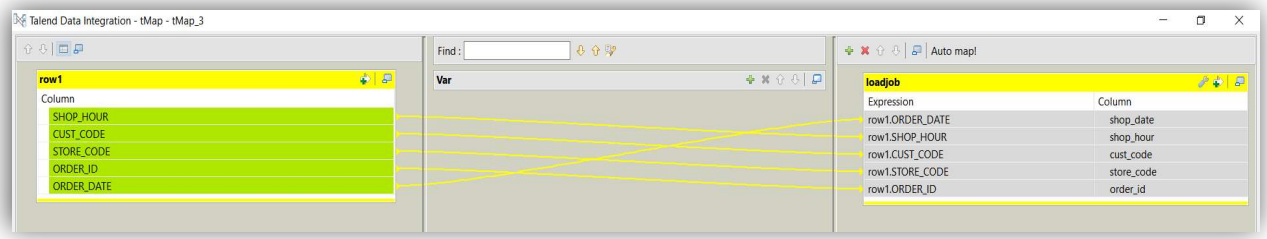


The attributes of the Item and Item Brand dimension had direct one to one mapping here with the target attributes. And in the mapping for Item CMIM we have conversion of data types to integers.

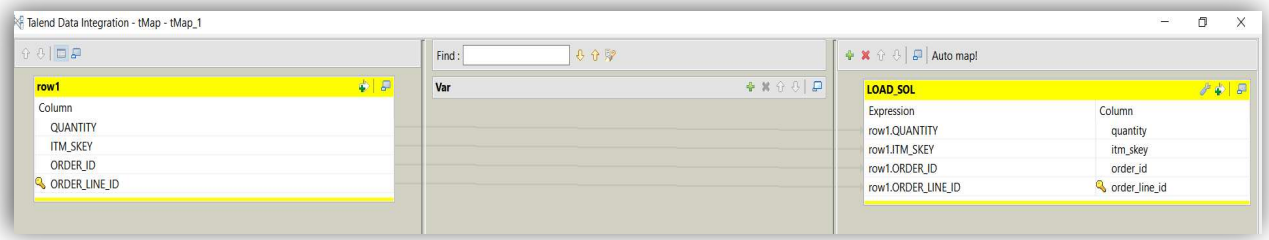
9.2.5 Sales Order Fact / Sales Order Line Fact

The transformation for the fact table had no particular conversions as the datatypes were all in sync. Hence, we performed a direct mapping as shown below.

SalesOrder Fact



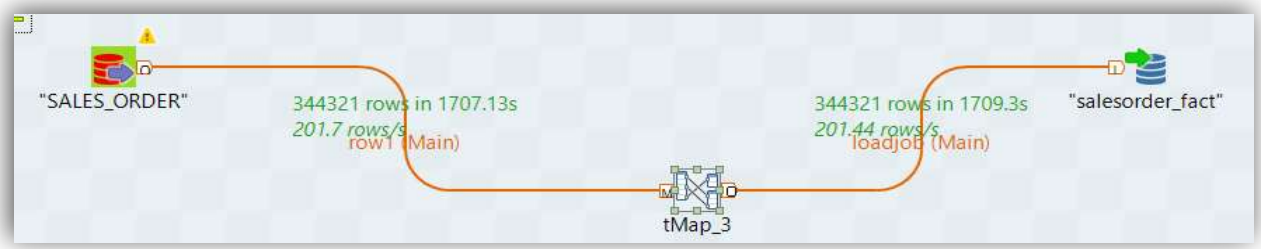
Sales Order Line Fact



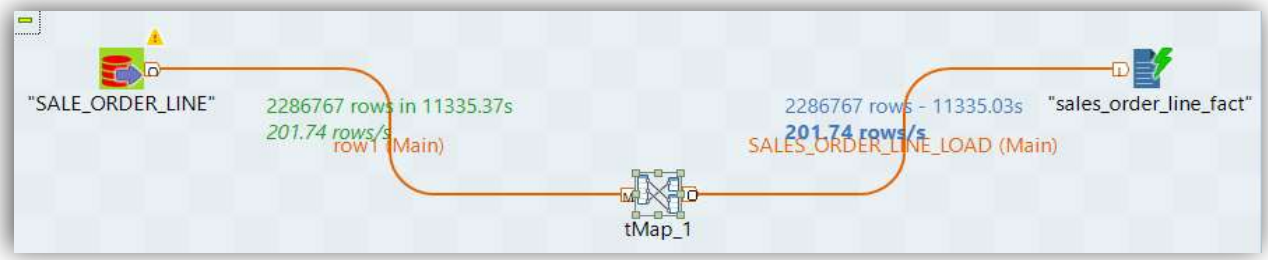
9.3 Load

Finally, the load function is used to write the resulting data (either all of the subset or just the changes) to a target database, which may or may not previously exist. We made use of all full loads to populate the tables already created in Redshift. The jobs created in the transform stage for each table is then executed to complete the load of data.The two major loads we had performed were for the two fact tables, the Salesorder_Fact and Sales_Order_Line_Fact.

The load to the Salesorder_fact table was for over 300 thousand records and took around half hour for the load process. The details are as shown below:

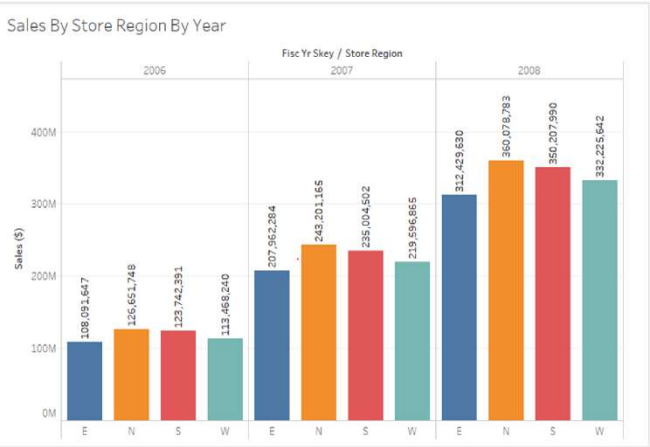


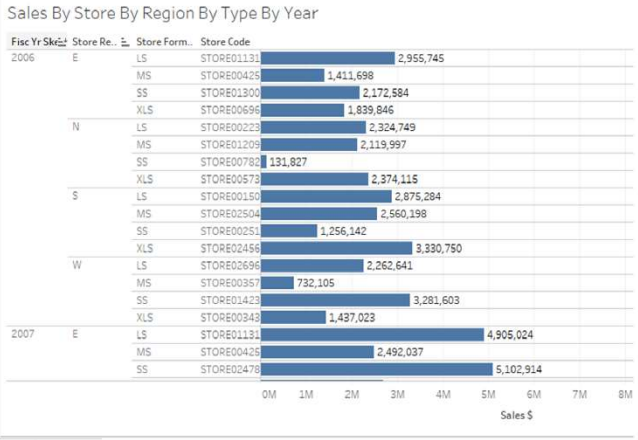
The second load to the Sales_Order_Line_Fact table involved more than 2 million records and this load took approximately 4 hours to complete.



10 Reporting and Analysis:

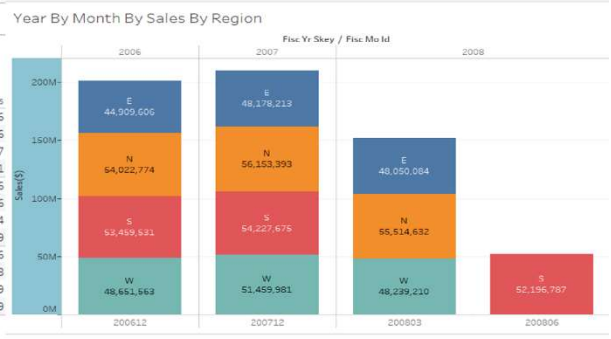
10.1 Dashboard – Sales Revenue

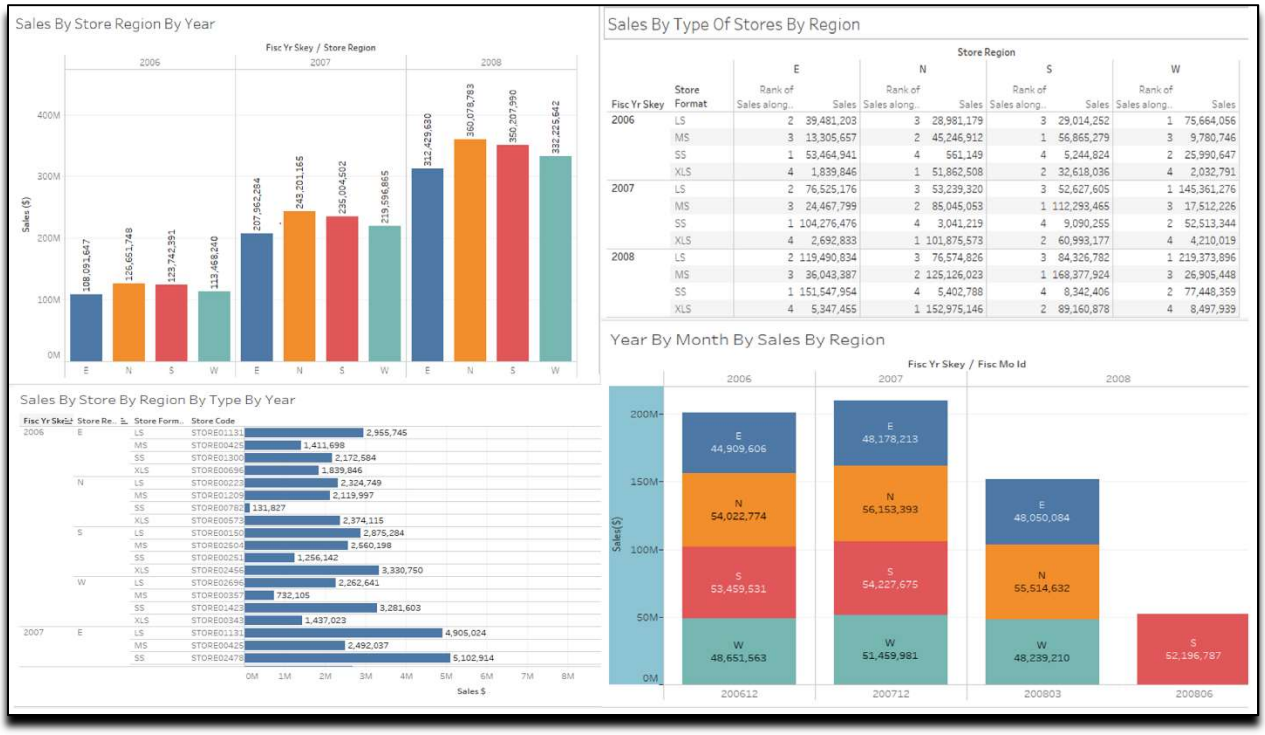




Sales By Type Of Stores By Region

		Store Region							
		E		N		S		W	
Fisc Yr Skey	Store Format	Rank of Sales along..	Sales	Rank of Sales along..	Sales	Rank of Sales along..	Sales	Rank of Sales along..	Sales
2006	LS	2	39,481,203	3	28,981,179	3	29,014,252	1	75,664,056
	MS	3	13,305,657	2	45,246,912	1	56,865,279	3	9,780,746
	SS	1	53,464,941	4	561,149	4	5,244,824	2	25,990,647
	XLS	4	1,839,846	1	51,862,508	2	32,618,036	4	2,032,791
2007	LS	2	76,525,176	3	53,239,320	3	52,627,605	1	145,361,276
	MS	3	24,467,799	2	85,045,053	1	112,293,465	3	17,512,226
	SS	1	104,276,476	4	3,041,219	4	9,090,255	2	52,513,344
	XLS	4	2,692,833	1	101,875,573	2	60,993,177	4	4,210,019
2008	LS	2	119,490,834	3	76,574,826	3	84,326,782	1	219,373,896
	MS	3	36,043,387	2	125,126,023	1	168,377,924	3	26,905,448
	SS	1	151,547,954	4	5,402,788	4	8,342,406	2	77,448,359
	XLS	4	5,347,455	1	152,975,146	2	89,160,878	4	8,497,939

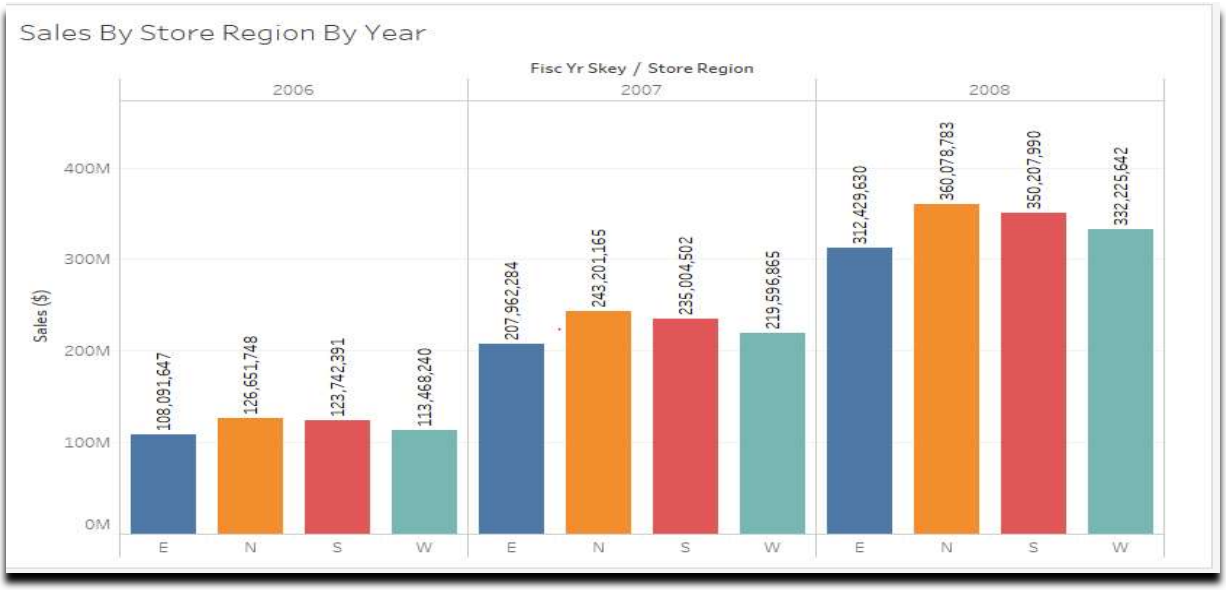




Overview of this dashboard:

The business owners will get an overview of the revenue being generated by the retail chain. The sales figures can be viewed with respect to various parameters, viz., store region, store type, store code and month. We have used a drill down approach to measure the revenue generated from a larger perspective, i.e., from sales region wise to arriving at a granular level detail of individual stores in the regions making the highest profit.

Below are the details pertaining to each chart in the above dashboard:



a) Inference from the above chart:

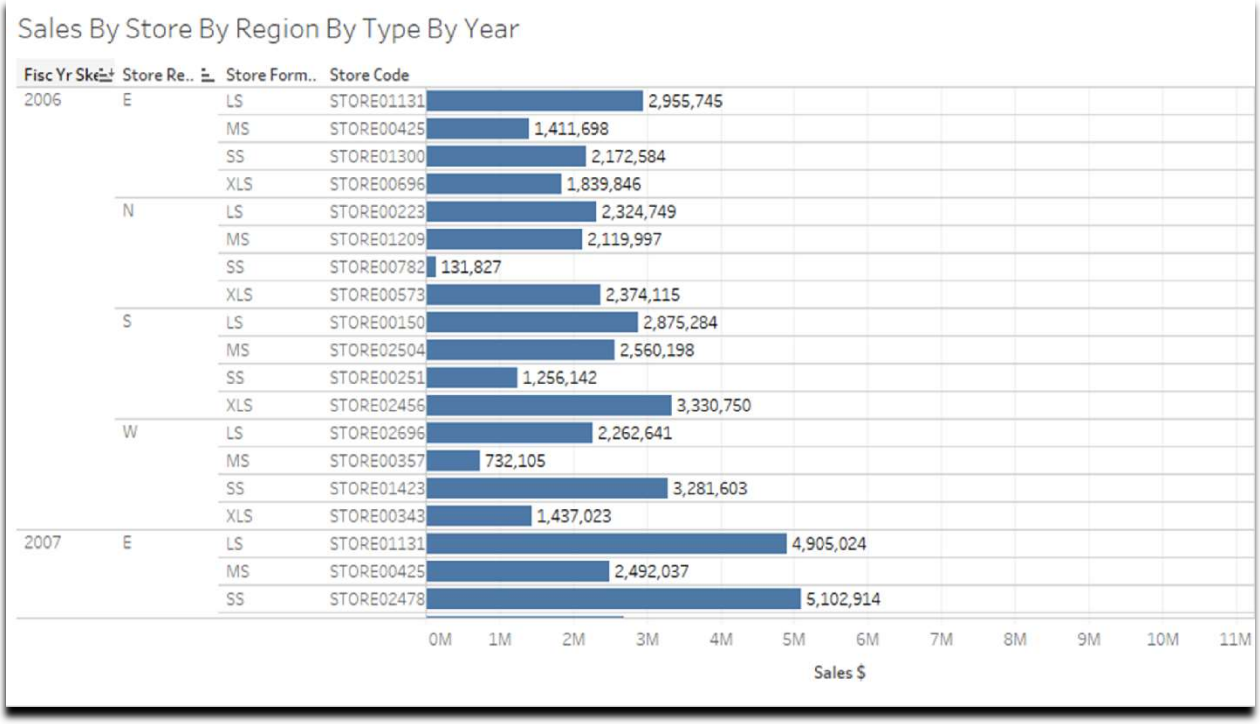
In the above chart, sales generated from each region across the span of three years has been displayed. As we can see the stores in the northern region have been making the highest profit all the three years. We also observe that the sales of this retail chain has been growing steadily each year. This indicates that the retail chain has been able to gather considerable customer attention and the marketing and promotions strategy in place are proving its merit.

Sales By Type Of Stores By Region

		Store Region							
		E		N		S		W	
Fisc Yr Skey	Store Format	Rank of Sales along..	Sales	Rank of Sales along..	Sales	Rank of Sales along..	Sales	Rank of Sales along..	Sales
2006	LS	2	39,481,203	3	28,981,179	3	29,014,252	1	75,664,056
	MS	3	13,305,657	2	45,246,912	1	56,865,279	3	9,780,746
	SS	1	53,464,941	4	561,149	4	5,244,824	2	25,990,647
	XLS	4	1,839,846	1	51,862,508	2	32,618,036	4	2,032,791
2007	LS	2	76,525,176	3	53,239,320	3	52,627,605	1	145,361,276
	MS	3	24,467,799	2	85,045,053	1	112,293,465	3	17,512,226
	SS	1	104,276,476	4	3,041,219	4	9,090,255	2	52,513,344
	XLS	4	2,692,833	1	101,875,573	2	60,993,177	4	4,210,019
2008	LS	2	119,490,834	3	76,574,826	3	84,326,782	1	219,373,896
	MS	3	36,043,387	2	125,126,023	1	168,377,924	3	26,905,448
	SS	1	151,547,954	4	5,402,788	4	8,342,406	2	77,448,359
	XLS	4	5,347,455	1	152,975,146	2	89,160,878	4	8,497,939

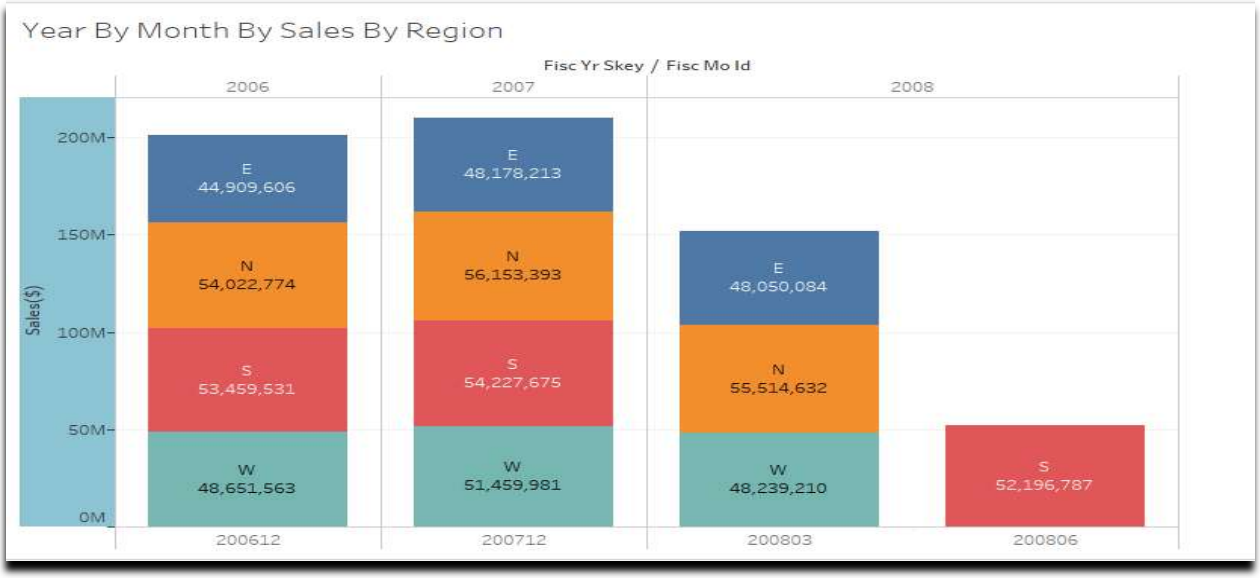
b) Inference from the above chart:

In the above chart, we have drilled down to the next level of finding the type of store (LS, MS, SS, XLS) in all the four regions (North, South, East, West) across all the three years (2006, 2007, 2008) that makes the highest profit. The store types have been ranked to find out the highest and lowest measure of sales with rank 1 awarded to the store type that generates the maximum sales and rank 4 to the store type that generated the minimum sales. As seen from the chart, store type SS in the Eastern region, XLS in the northern region, MS in the southern region and LS in the western region has been making the highest profit and this has been consistent all the three years.



c) Inference from the above chart:

In the above chart, we have drilled down to the next level of finding the individual stores in each store type (LS, MS, SS, XLS) in all the four regions (North, South, East, West) across all the three years (2006, 2007, 2008) that makes the highest profit. As we can see, the store code of stores that make the highest profit has been displayed.



d) Inference from the above chart:

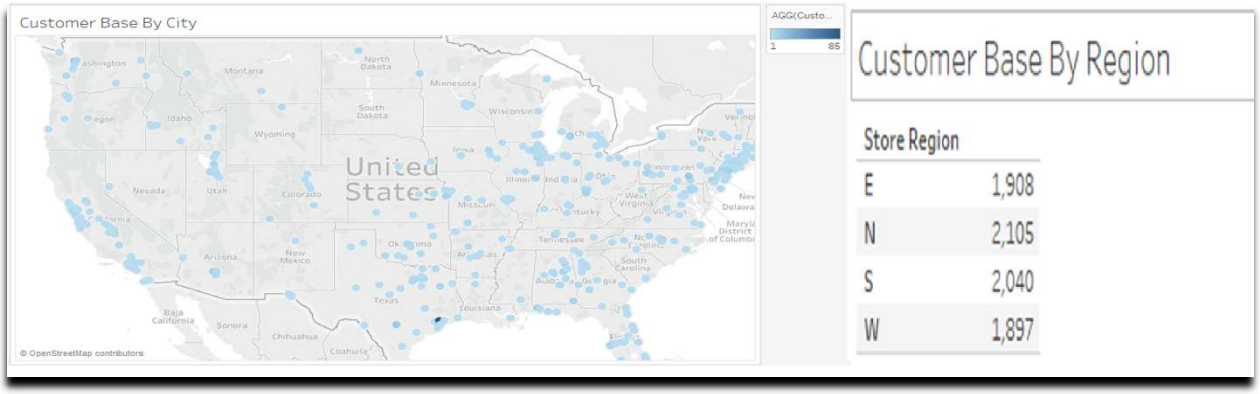
In the above chart, monthly sales analysis of all the four regions (East, North, South, West) has been done for all the three years. As seen for the year 2006 and 2007 all the four regions generate the maximum revenue in the month of December. There is a variation to this in the year 2008 where we see that the

Eastern, Northern and Western regions generate the maximum revenue in the month of March and in the southern region we observe the highest sales figure in the month of June.

Key decision making points:

- The management should focus on developing and increasing the number of stores of type SS in the eastern region, store type XLS in the northern region, MS in the southern region and LS in the western region as this seems to find traction with the customer base of the respective regions.
- The management can award incentives to the individual stores of the different formats making the highest profit in all the 4 regions. This would foster a competitive drive among all the stores to maximize profits.
- The maximum sales in the month of December in the first two years, i.e., 2006and 2007 can be attributed to the fact that December is a holiday and festive month and customers frequently visit the store during this month. The management should focus on developing robust marketing and promotions strategy for this particular month.
- The monthly analysis of store sales in chart d exhibits a diversion from the sales trend of 2006 and 2007 in the year 2008 where we see the maximum sales being generated in the months of March and June. The exact reason for this change should be looked into which could possibly be due to a strong competitor emerging in the market in the year 2008 and diminishing the market share of the retail chain. Enhancing the sales and marketing strategy to best suit the changing market dynamics should be worked upon.

10.2 Dashboard – Customer Analysis



a) Inference from the above charts:

From the above chart, we observe that this retail chain has the maximum number of customers in the northern region. The customer base is very strong in the northern and southern regions of the United States of American when compared to the eastern and western regions. Houston in the state of Texas holds the maximum number of customers as observed from the dark blue spot in the map.

Top 10 Customers By Visting Frequency For Each Fiscal Year

		Fisc Yr Skey					
		Rank_Visting_Frequency alo..			Visit_Frequency		
Full Name	Cust Code	2006	2007	2008	2006	2007	2008
Abdul Moreno	CUST0000473868			8			129
Antonio Sifford	CUST0000648788			7			129
Brianna Savage	CUST0000693928	6	5	3	48	96	136
Carl Jones	CUST0000746359		9			81	
Darlene Battle	CUST0000558339	2	7		59	88	
David Madsen	CUST0000595675	10	4		45	97	
Hector Dominguez	CUST0000751821			2			141
Louis Chancy	CUST0000004239	9			45		
Mary Keely	CUST0000487479	7	3	10	46	107	126
Rachel Trueblood	CUST0000732836	8	1	1	45	117	143
Rodney Brown	CUST0000305191			4			135
Rodney Parrish	CUST0000115970		8			84	
Ronald Clark	CUST0000837146			9			128
Terry Tonkin	CUST0000975940			6			132
Vincent Dube	CUST0000975272	3	6		56	93	
Warren Carter	CUST0000685886		10			81	
William Mack	CUST0000609232	5			49		
William Pinion	CUST0000088415	1	2	5	65	114	133
William Ware	CUST0000858503	4			55		

b) Inference from the above chart:

From the above chart we can observe that customer analysis has been done to arrive at the top 10 customers by visiting frequency for all the 3 years (2006, 2007, 2008). Customers have been ranked with rank 1 given to the customer with the maximum number of visits and rank 10 to the customer with the least number of visits. Examples of significant observations that hold value is –

- Customer William Pinion whose visits has significantly lowered in comparison to other customers (Dip in the customer rank from 1 in the year 2006, rank 2 in the year 2007 and rank 2008 in the year 2008)
- Customer Rachel Trueblood whose visits has significantly improved in comparison to other customers (Rise in the customer rank from 8 in the year 2006 to rank 1 in the year 2007 and 2008)

Top 10 Customers By Sales

		Fisc Yr Skey					
		Sales			Unique Sales Rank along Table (Down)		
Full Name	Cust Code	2006	2007	2008	2006	2007	2008
Barbara Acosta	CUST0000365429	984,986	2,047,235	3,668,543	5	2	2
Calvin Pike	CUST0000999385	836,705			10		
Clarence Causey	CUST0000530119	1,334,920			2		
Danielle Jackson	CUST0000016087	1,061,731			3		
Eric Barney	CUST0000763030		1,976,320	3,197,761		4	4
Helen Banker	CUST0000440403	847,986	1,626,199		9	10	
Helen Stewart	CUST0000088575		2,038,701			3	
John Miller	CUST0000895836		1,755,518			7	
Keith Watson	CUST0000497628		1,772,666	2,863,049		6	5
Kelly Avendano	CUST0000041792			2,801,119			7
Lakia Turner	CUST0000411707	887,496		2,711,775	7		8
Mary Shannon	CUST0000768393			3,476,151			3
Melissa Trussell	CUST0000708813	868,852			8		
Randall Lundgren	CUST0000483138			2,668,759			9
Renee Cotton	CUST0000186979	1,967,984	3,115,825	4,258,767	1	1	1
Sadie Buchanan	CUST0000530406	990,791	1,655,207		4	9	
Thomas Martinez	CUST0000386495		1,746,997	2,826,637		8	6
Timothy Sloan	CUST0000376140			2,506,159			10
William Mack	CUST0000609232	944,054	1,885,359		6	5	

c) Inference from the above chart:

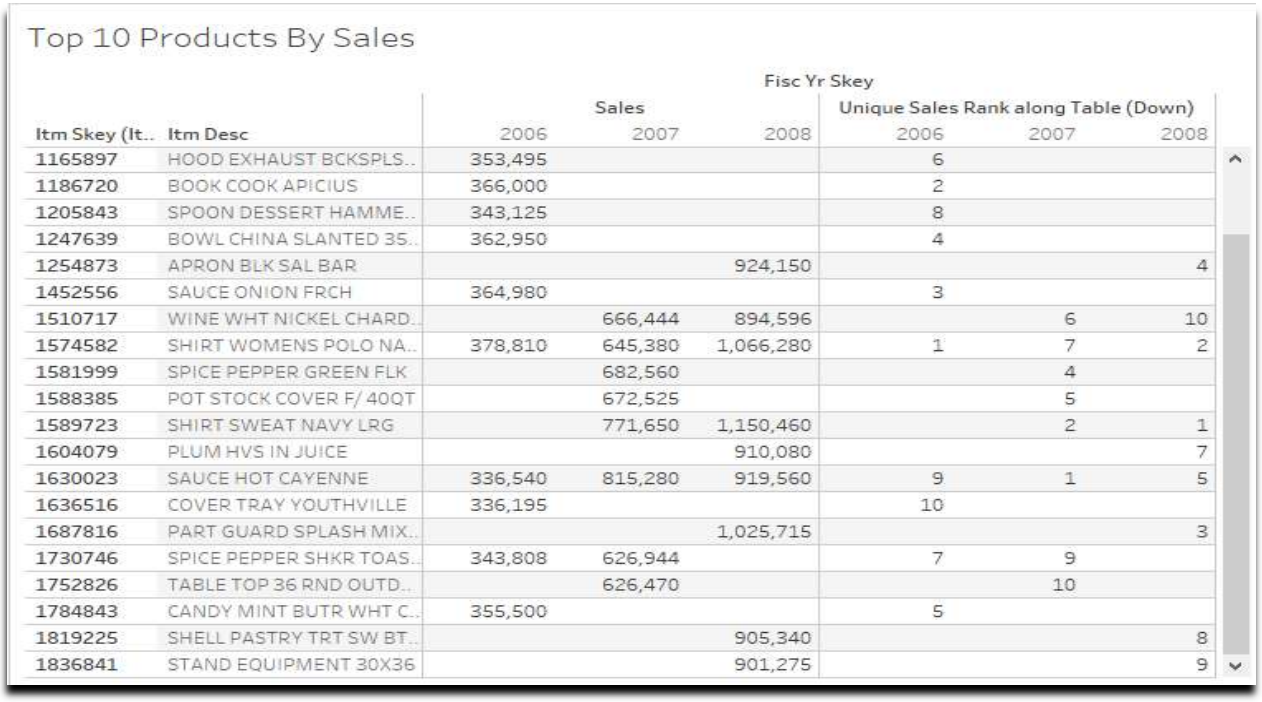
From the above chart we can observe that customer analysis has been done to arrive at the top 10 customers by sales for all the 3 years (2006, 2007, 2008). Customers have been ranked with rank 1 given to the customer with the highest purchase value and rank 10 to the customer with the lowest purchase value. Example of significant observation that holds value is –

- Customer Renee Cotton has been ranked 1 consistently which indicates that this particular customer has generated the highest purchase amount through all the years.

Key Decision Making Points:

- Management should focus on customers who are showing a downward trend in terms of frequency of visits by targeting those particular customers through exclusive offers (like loyalty cards, vouchers, special discounts, etc.) tailored to suit their needs.
- Management could request for further analysis to check the customers purchase pattern to figure out if the customer is spending larger amounts on low priced products or on high priced products. Once this is done, a strategy could be built around keeping a good number of varieties of those products in terms of brands and models and giving promotional offers on them to keep the top ranked customers spending unwavering.

10.3 Dashboard – Product Analysis



Top 10 Products By Sales

		Fisc Yr Skey					
		Sales			Unique Sales Rank along Table (Down)		
Itm Skey (It..	Itm Desc	2006	2007	2008	2006	2007	2008
1165897	HOOD EXHAUST BCKSPLS..	353,495			6		
1186720	BOOK COOK APICIUS	366,000			2		
1205843	SPOON DESSERT HAMME..	343,125			8		
1247639	BOWL CHINA SLANTED 35..	362,950			4		
1254873	APRON BLK SAL BAR			924,150			4
1452556	SAUCE ONION FRCH	364,980			3		
1510717	WINE WHT NICKEL CHARD..		666,444	894,596		6	10
1574582	SHIRT WOMENS POLO NA..	378,810	645,380	1,066,280	1	7	2
1581999	SPICE PEPPER GREEN FLK		682,560			4	
1588385	POT STOCK COVER F/ 40QT		672,525			5	
1589723	SHIRT SWEAT NAVY LRG		771,650	1,150,460		2	1
1604079	PLUM HVS IN JUICE			910,080			7
1630023	SAUCE HOT CAYENNE	336,540	815,280	919,560	9	1	5
1636516	COVER TRAY YOUTHVILLE	336,195			10		
1687816	PART GUARD SPLASH MIX..			1,025,715			3
1730746	SPICE PEPPER SHKR TOAS..	343,808	626,944		7	9	
1752826	TABLE TOP 36 RND OUTD..		626,470			10	
1784843	CANDY MINT BUTR WHT C..	355,500			5		
1819225	SHELL PASTRY TRT SW BT..			905,340			8
1836841	STAND EQUIPMENT 30X36			901,275			9

a) Inference from the above chart:

From the above chart we can observe that product analysis has been done to arrive at the top 10 products by sales for all the 3 years (2006, 2007, 2008). Products have been ranked with rank 1 given to the product that has generated the maximum sales and rank 10 to the customer that has generated the lowest sales figure. We see that the list of products that make to the top 10 differs each year and only 2 products, i.e., Product ID – 1574582 and 1630023, hold a place in the top 10 list across the three years.

Top 10 Product By Quantity Ordered

		Fisc Yr Skey					
		Quantity			Rank_Quantity_Ordered alo..		
Itm Skey (It..	Itm Desc	2006	2007	2008	2006	2007	2008
445611	ANCHOVY FILET EASY OP..			75,524			5
455613	COVER TABLE TULIP MAR..			72,590			9
555006	SHEAR KITCHEN POULTRY..			71,675			10
559096	CONNECTOR GAS QUICK D..	29,280			6		
987551	OIL EMULSIFIED GLYERINE	29,072			7		
1085745	GARLIC CHOPPED IN OIL P..		52,456			7	
1149337	NUT PISTACHIO GREEN M..		51,192			8	
1204824	HANDLE MOP FIBRGLS SP..		50,630			9	
1225248	PLATE DINNER SATURN 1..			75,030			6
1258851	CANDLE LAMP LACE DESI..			72,895			8
1359421	SINK STNLS MODULE		52,460			6	
1360587	SIGN CUSTOM INSERT F/ ..		53,070			5	
1364554	FLOUR PEA CHICK		53,088			4	
1389296	SALT SEA FINE PORTUGU..		64,464			1	
1397580	CEREAL GRANOLA GOJI A..	27,176			9		
1439390	PLATE PLAS CLR 7.5 ETCH..			91,821			2
1446198	FLOUR MIX T55 VIRON PL..		61,620			2	
1484563	SAUCE CREAM PARMESAN	31,916			4		
1512655	BOWL MOLDED SQR BAM..			109,190			1
1526071	GLASS FLUTE SENSATION		50,630			10	

b) Inference from the above chart:

From the above chart we can observe that product analysis has been done to arrive at the top 10 products by quantity purchased for all the 3 years (2006, 2007, 2008). Products have been ranked with rank 1 given to the product whose units purchased is the maximum and rank 10 to the product whose units purchased is the minimum. We see that the list of products that make to the top 10 differs each year.

Top Brand By Quantity Ordered

		Fisc Yr Skey					
		Quantity			Rank_Quantity_Ordered along Table (Down)		
Brnd Cd	Brnd Desc	2006	2007	2008	2006	2007	2008
BHB/NPM	BUCKHD BF/NEWPRT PRD..	373,041	732,786	1,106,107	9	9	9
BRGKING	BURGER KING		716,423	1,093,963		10	10
MARKO	MID WEST MARKO	627,080	1,277,950	1,804,075	3	3	3
OHIO F	OHIO FARMS	510,330	1,007,112	1,511,307	5	5	5
PACKER	PACKER	5,469,553	10,609,584	15,705,970	1	1	1
STEELIT	STEELITE INTL USA	505,385	956,785	1,411,235	6	6	6
SYS CLS	SYSCO CLASSIC	428,003	815,474	1,221,831	8	8	8
T HORTN	TIM HORTON'S RESTAUR..	1,021,235	1,994,606	3,012,723	2	2	2
ULT SRC	THE ULTIMATE SOURCE	368,135			10		
VOLLRTH	VOLLRATH	452,315	855,220	1,317,600	7	7	7
WENDYS	WENDY'S HAMBURGERS ..	592,131	1,128,897	1,641,332	4	4	4

c) Inference from the above chart:

From the above chart we can observe that product analysis has been done to arrive at the top 10 brands by quantity purchased for all the 3 years (2006, 2007, 2008). Brands have been ranked with rank 1 given to the brand whose product units purchased is the maximum and rank 10 to the brand whose product units purchased is the minimum. We see that the list of brands that make to the top 10 remains consistent over the span of the 3 years with the same set of brands reaching the top 10 list.

Key Decision Making Points:

- From chart (a) the management can get the list of their best selling products and capture additional sales by promoting the popularity of these products through attractive online ads, flyer distribution in stores, adopting innovative product placement strategy plans in the store, special discounted rates.
- From chart (b) the management can get the list of products that are ordered the most in terms of quantity. Demand forecast should be done for these products and robust inventory management should be in place to maintain stock.
- From chart (c) the management can focus on the brand promotion of these specific brands and establish strong ties with the wholesalers that sell these brands.