December 8, 2014

IDS 521 – Advanced Database Management Systems
Acquire Valued Shoppers Challenge
Final Project Report
Prof. Ali Tafti

Mounica Sirineni: msirin2@uic.edu

Nikhila Valipe: nvalip2@uic.edu

Sai Lahari Jalaparthi: sjalap2@uic.edu

Contents

Overview.. 3
Description of the data 4
Detailed Enterprise Bus Matrix 6
Dimensional Models and Supporting Arguments 7
Sample Business Intelligence Reports 12
Technology and Implementation 13
Appendix 16
Bibliography 19

Overview

Business Process – Sales

Industry – E-commerce

Introduction

E-commerce, also known as Electronic Marketing, refers to buying and selling of goods and services over an electronic system, such as the internet. The history of E-commerce dates back into the age of 1970's. Today, this business has grown into a large chain. Therefore, business analysis for e-commerce is very important to maintain competitive advantage.

Objective

A well designed data warehouse stores historical data and turns it into information and knowledge. In the case of e-commerce, it helps in making right decisions by providing right information at the right time. Since technologies for e-commerce are developing rapidly, a data warehouse model will provide a significant contribution to its business intelligence.

Motivation

Our project deals with sales transactions of an e-commerce business. We have collected transactions for two years where consumers were not provided any incentives in first year and incentives were provided in second year. From this data, we try analyze customer purchase behavior. This analysis will help in attracting new customers and also retaining past customers. It also provides with the list of customers who did not use any incentives. We can also know percentage of sales for each product and take necessary actions on products which are being sold less.

Description of the Data

The data is about online sales transaction of e-commerce website. Our data source is Kaggle.com Website. The data is in the form of an excel sheet. The data comprises of various different types of fields and variables including both categorical and quantitative variables which are described below.

Column Name	Description
Customer Dimension	
Customer_id	The unique ID given to a customer
Customer_name	Name of the customer
Customer_address	Address of the customer
Customer_phonenum	Phone number of the customer
Date Dimension	
Date	The date of purchase
Date_time	The time of purchase
Day_of_week	The name of day of purchase
Day_Number_in_Month	The number of day of purchase in month
Day_Number_in_Year	The number of day of purchase in year
Holiday_Indicator	Indicates whether day of purchase is holiday or not
Weekday_Indicator	Indicates whether day of purchase is weekday or not
Calender_Month_Name	The name of month of day of purchase
Calender_Month_Number_in_year	The number of month of day of purchase in year
Quarter	The quarter number in the year
Year	Year of day of purchase
Offers Dimension	
Offer_id	An id representing a certain offer
Offer_code	Promotion code
Category	The product category
Quantity	The number of units one must purchase to get the discount
Company	An id of the company that sells the item
Offer value	The dollar value of the offer
Brand	An id of the brand to which the item belongs
Offer name	The name of the offer
Coupon type	The type of coupon offered
Offer_startdate	The start date a customer received the offer for product
Price_reduction_type	The type of price reduction implemented in that offer
Demographics Dimension	
Demographics_id	An ID representing demographics of a customer
Repeattrips	The number of times the customer made a repeat purchase
Repeater	A boolean, equal to repeattrips > 0
Age	Age of the customer

Income	Income of the customer
Store Dimension	
Store_id	An id representing a store
Chain_Number	An integer representing a store chain
Store_name	The name of the store
Store_address	The address of the store
Store_city	The city of the store
Store_state	The state of the store
Store_zipcode	The zipcode of the store
Store_Manager	The manager of the store
Store_phoneno	The phone number of the store
Product Dimension	
Product_id	An id representing the product
SKU_Number	The store keeping unit number of the product
Product_name	The name of the product
Brand	An id of the brand to which the item belongs
Department	An aggregate grouping of the Category
Category	The product category
Package type	The type of package of the product
Package size	The size of package of the product
Weight	The weight of the product
Storage type	The storage type of the product
Market Dimension	
Market_id	An id representing a geographical region
Market_name	The name of the market
Facts	
Transaction_ID	The ID representing the transaction
Company	An id of the company that sells the item which is degenerate dimension
Product size	The amount of the product purchase
Product measure	The units of the product purchase
Product quantity	The number of units purchased
Regular product amount per unit	The regular price of product per unit
Discount product amount per unit	The discount price of product per unit
Net product amount per unit	The net price of product per unit after making discount on regular price
Gross profit per unit	Profit per unit because of the discount offered on the product
Total regular product amount per unit	The regular price of product per unit multiplied by quantity of product purchased
Total discount product amount	The discount price of product per unit multiplied by the quantity of product purchased

Total net product amount	The net price of product per unit multiplied by the quantity of product purchased
Total gross profit	Profit per unit multiplied by the quantity of product purchased

The most important data fields are Product_id, Store_id, Market_id, Customer_id, Date, Offer_id, Demographic_id, Product quantity, Total regular product amount, Brand, Company #.

Detailed Enterprise bus matrix:

Common Dimensions

Business Processes	Date	Product	Market	Store	Customer	Offers	Demographics
Sales	X	X	X	X	X	X	
Offers tracking	X	x	X	X		X	
Frequent customer purchases	X		X		X	X	X
Returns	X	X		X	X		
Market Analysis	X	X	X		X	X	X
Profits	X	X	X	X		X	

Business Process:

Sales: In this process, we will find out how much sales each product has made, who has purchased the product, what offers are there on the product, in which market or store the sales are processed. Hence, we will use date, product, market, store, customer, offers dimensions.

Offers Tracking: This process gives information about the various offers available on the product in a particular store and market. Therefore, we use date, product, market, store and offers dimension.

Frequent customer purchases: If we want to know about the details of a customer and his purchase history we use this process. Date, market, customer, offers, demographics dimensions are used.

Returns: Using this process, we will find out the details of those products which customer has returned. So we use, date, product, store and customer dimensions.

Market Analysis: To predict the customer purchasing behavior, we do market analysis. Hence, date, product, market, customer, offers, demographic dimensions are involved.

Profits: If we want to know the profits on the sales, we use this business process. Date, product, store, market, offers dimensions are used.

Grain:

1. One row per transaction without offers in "Transactions" fact table.
2. One row per transaction with offers in "Transactions with offers" fact table.

Dimensional Models and Supporting Arguments

The key factor to the success of E-commerce is the analysis of Customer behavior which helps them to increase their sales. Another key step is sales of product. With the analysis of sales of product they can identify whether modifications are needed if the sales are very less for that. Now-a-days Marketing is more based on information than the sales power. In order to sustain in this competitive world, they need to acquire customers and retain them. So, we suggested a dimensional model which helps them to analyze the consumer behavior and generate reports based on the sales transactions.

Approach to the Dimensional Model

1. **Selection of Business Process**: With the available source data and keeping in mind the business objectives of the E-commerce, we selected the business process as the 'Online sales transactions'. This helps them to analyze consumer behavior and product sales in different situations.

2. **Declaration of Grain**: In our scenario we need to consider two cases.
    i. Customers were not offered incentives in the year 2012.
    ii. Customers are offered incentives in the year 2013 and compare these transactions with those in the previous year.

   Granularity level for the year 2012 is the sales transaction i.e.., one row per transaction without offers. As in the next case we need to have extra data to be captured, so we are defining the granularity level for that as the sales transaction with the offers captured in it i.e.., one row per transaction with offers.

   Even though there is a possibility of adding columns to the existing granularity in the year 2012, we don't want to disturb that data. If we add those columns we need to modify the data for that year by adding zero's to it. This disturbs our main motive of the analysis. So we suggested another granularity level for that which helps them with detailed drill down process than the previous.
3. **Identification of Dimensions:** With the declared level of granularity, we have identified the dimensions as below.

   Case 1 – Year 2012: Descriptive dimensions attributes are Date dimension to capture the date of sales transaction, Store dimension to identify from which store the transaction happened, Market dimension for geographical location identification, Customer dimension to know who purchased the product in that transaction and Product dimension

   Case 2 – Year 2013: Along with the above dimensions, we identified extra descriptive dimensions such as Offers dimension to capture incentive offered and Demographics mini dimension to Customer Dimension.
4. **Identification of Facts:** As the facts must be declared based on the granularity, we have identified two fact tables 'Transaction' fact table and 'Transaction with offers' fact table to maintain consistency.

   It is important for the company to know which customers did not avail the offers or which products are not at all sold when the incentives are offered. Therefore we suggest factless fact table which captures all the incentives offered to the customers on particular products.

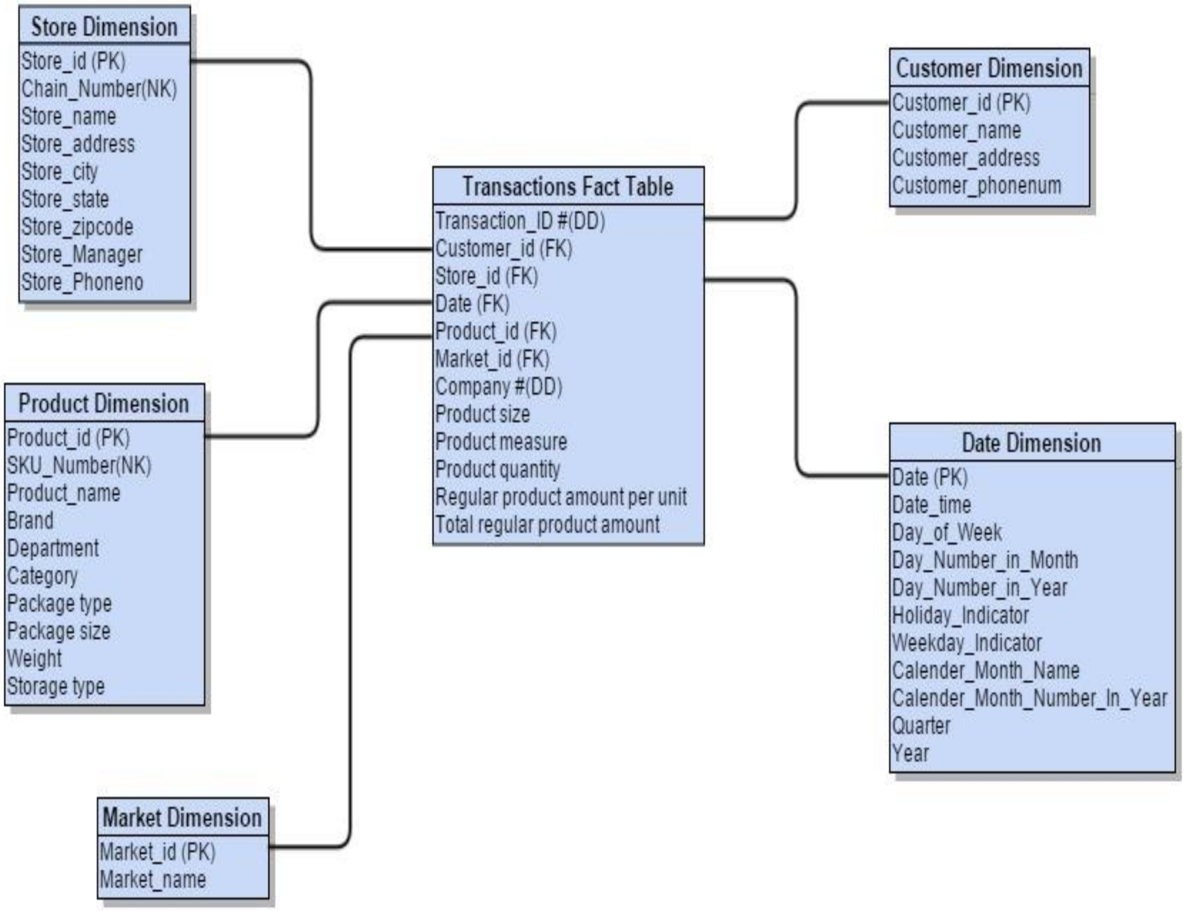


Fig: Dimensional model for year 2012

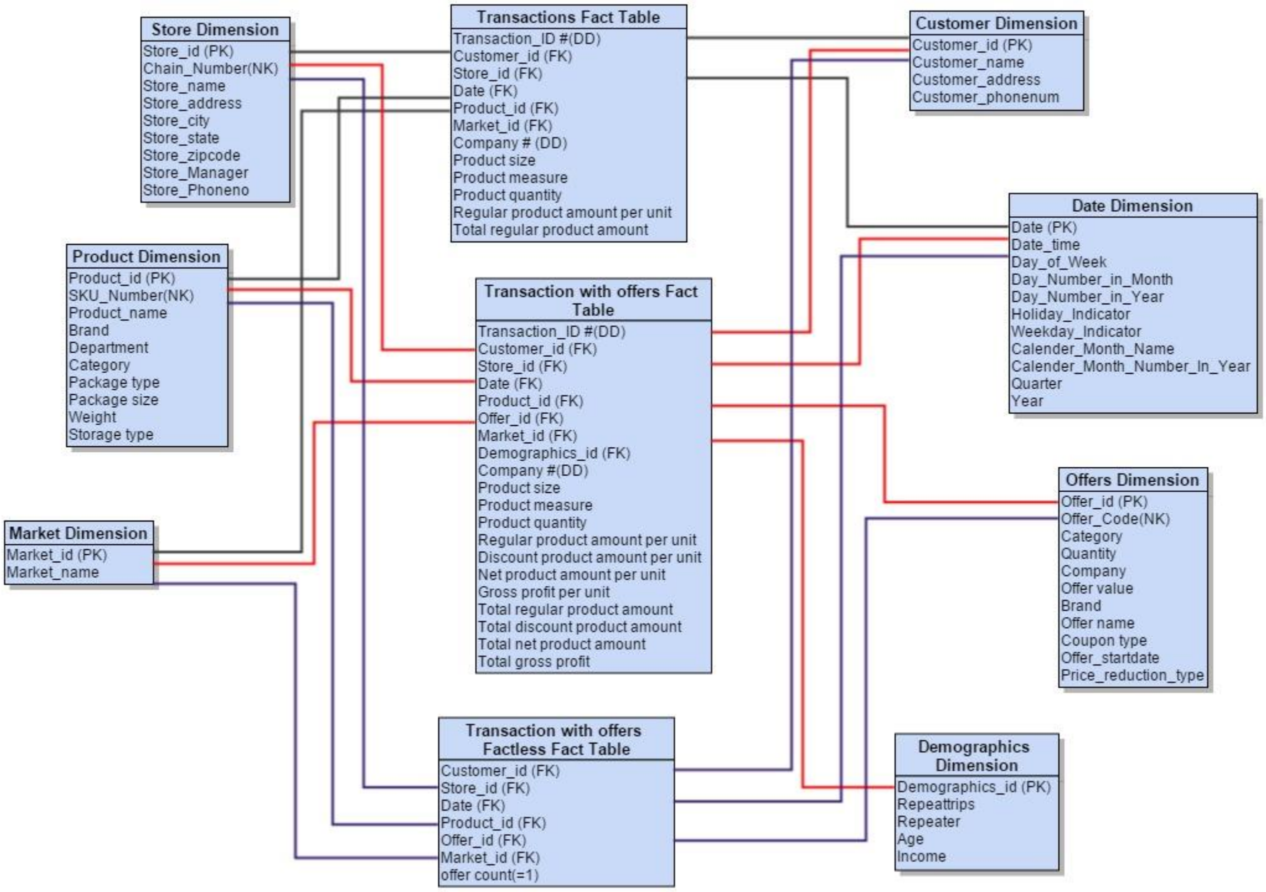


Fig: Dimensional model for year 2013 for transaction with offers given

Fact tables

The type of the snapshot we are using is the transaction snapshot which allows us to hold the data at detailed level.

- **Transaction fact table:**
  Granularity is one row per transaction. Transaction ID captured is considered as degenerate dimension. Product size, Product measure, Product quantity, Regular product amount per unit and Total regular product amount are also captured. We have foreign keys in the fact table for dimensions: Store, Product, Market, Customer and Date.
  Product size, Product measure, Regular product amount per unit are the non-additive facts. Product quantity and Total regular product amount are the additive facts.
- **Transaction with offers fact table:**
  Granularity is one row per transaction with offers. Transaction ID captured is considered as degenerate dimension. Extra facts that are mainly captured in this are Discount

product amount per unit, Net product amount per unit, Total discount product amount, Total net product amount. We have foreign keys in the fact table for dimensions: Store, Product, Market, Customer, Offers, Demographics and Date.
Gross profit per unit and Total gross profit are the derived facts.

Dimensions

Store Dimension: Store dimension is used for tracking from which store chain the product is sold in that transaction. Store_Id is the primary key and Chain number is the natural key. Chain number is considered as Natural key as it was generated by the operational source system.

Product Dimension: Product dimension represents details of the product. Product_id is the Primary key. The product is defined with Brand, Department and Category. This facilitates the business users through detailed drill down process at different brand, department and category levels. SKU Number is the natural key defined by the operation systems.

Market Dimension: Market dimension describes the geographical location from where the product was sold in that transaction. Market_id is the primary key for the dimension table.

Customer Dimension: Customer dimension is used for the purchase behavior analysis. Customer_id is the primary key. In each transaction in the fact table, customer ID is also captured. Using this, we can analyze which customer has frequent purchases. The slowly changing dimension that best fits for this dimension is Type 1 slowly changing dimension as any update to the customer database overwrites the existing.

Date Dimension: Date dimension has attributes which indicates whether the particular is Holiday, Week day. This helps the business users to analyze how the sales of the product on holidays or weekends.

Offers Dimension: Offers dimension describes the details regarding the offers given to the customers on the product. Offer_id is the primary key and Offer Code is the Natural Key. Offer code is the code generated by the operational source system.

Demographics Dimension: Demographics dimension is the mini dimension to the customer dimension. This is a Type 4 slowly changing dimension. We suggest this dimension to analyze the customer behavior at the income level.

Degenerate Dimensions

Transaction_ID is the grouping attribute used to group the products that are sold in a particular transaction, as it is quite common user might buy several products in a single transaction. There are no other attributes related to the Transaction_ID to keep it in a separate dimension table. Therefore it is considered as degenerate dimension in the fact table.

Company # is the degenerate dimension. This is considered as degenerate because there are no other attributes related to the company. Therefore it is mentioned in the fact table as degenerate dimension.

Transaction with offers factless Fact table

This factless fact table contains all the offers given to the customer for particular product. This helps the business user to report which offers are not availed by any customer. As this table doesn't contain any metrics we have included a dummy fact with Offer count as 1.

Other Design Decisions

For offers dimension, if there is any update to product or brand we cannot directly update it as it may lead to wrong situations. Therefore we suggest them to use Type 2 slowly changing dimension. For this type of dimension we need to have three required columns Effective date, Expiration date and Current row indicator. Among this we already have Start date which is effective date. They can add two columns to the dimension which doesn't affect the granularity.

For Product dimension, if there are any changes to department, brand or category then we suggest them to use Type 2 slowly changing dimension. For this dimension they doesn't have any of the three required columns. So they need to add three columns to this dimension.

Sample Business Intelligence Reports

The dimensional model developed can be used to answer to the following business questions that can provide useful insights to the sales of the company.

1. **What is the change in the behavior of the customer when the offers are available?**
   - We wanted to know the difference in customer purchasing power when the offers are provided to the products. We have taken sample 10 customers and analyzed their behavior.
   - From the reports generated, we concluded that customers purchased variety of products once the offers are available on the products. This shows that if discounts are applied on the products, customers are tend to get attracted and their purchasing power increases.
   - **Dimension table accessed**: Date, product, market, store, customer, offers, and demographics.
2. **What is the percentage difference in the sales of a brand when a discount was applied?**
   - For this question, we have taken 4 brands and analyzed their sales when offers are given on the product and when there were no offers.
   - From the reports, we can say that the sales of the 4 brands have increased dramatically when offers were promoted on the brands.
   - **Dimension tables accessed:** Date, product, market, store, offers

3. **What is change in the net sales year wise after the promotion was applied?**
    - We have analyzed the sales in the year 2012, where offers were not present and in the year 2013, where offers were present.
    - We can conclude that, 2012 has less sales compared to the year 2013, where promotions were offered to the customer.
    - **Dimension tables accessed:** Date, product, market, store, offers.
4. **How many customers didn't avail the offers?**
    - We will like to know those customers who didn't avail the offers and who didn't make any purchases. From this data, we can conclude whether if a promotion is applied on a product, there is any improvement in sales or not. We derived this table, using SAP Hana technology.
    - **Dimension tables accessed:** Date, product, market, store, offers, and customers.

Kindly refer to the sample reports in the appendix.

Other business questions:

- What is the purchase pattern for repeated users?
- What are the top 5 selling brands sales wise and revenue wise?
- What are the lowest 10% items sold in the current month?
- After applying promotions, which product sales have improved and which have deteriorated?
- What products have not been sold online from past one month?

Technology and Implementation

Tools used for Implementation

1. MySQL Workbench
2. Microsoft Excel Pivot table
3. SAP HANA Studio

MySQL Workbench:

Data collected for 2 years is in 'csv' format. In the data there were lot of transactions, unused information and it was difficult to transform to data into SAP HANA Studio as the size of data was 20GB. So we decided to clean the data and use MySQL Workbench for cleaning the data and use transactions for the months of March and April in 2012 and 2013 for analysis. After that we extracted the cleaned data in csv format.

Microsoft Excel Pivot table:

Microsoft excel pivot table is used for filtering the data and generation of reports. This layout has the following features

- Report filter: Report filter can be used when a filter is to be applied for an entire table i.e., to have only particular values chosen from the dropdown available
- Row Labels: Row labels can be used when the user need to filter one or more rows to be shown in the pivot table
- Colum labels: Column labels can be used when the user need to filter one or more columns to be shown in the pivot table
- Summation values: This can be used for summation of values based on the filter applied if it is a quantitative variable or it can be used to show the count for descriptive variables.

SAP HANA Studio:

SAP HANA has the modeler feature which can be used for slicing and dicing the data in the database tables and create views for the business scenarios. This helps the end users to generate reports and in the decision making process.

The modeling features enables the simulation of entities and the relationship between them. We have three views Attribute view, Analytic view and Calculation view. Attribute view is used for the dimensions and Analytic view is used for the fact tables. In the Analytic view we can join the dimensions with the fact tables. In the calculation view we can add attributes that need to be calculated from the attributes present in the table for example Net profit.

Attributes/fields in the data table can be represented in two ways - Attribute and Measure. The field is represented as ‘Attribute’ if they are the descriptive variables or if they are used for grouping purposes. The field is represented as ‘Measure’ if it is a quantifiable data.

After the validation of model, the end user can easily generate reports.

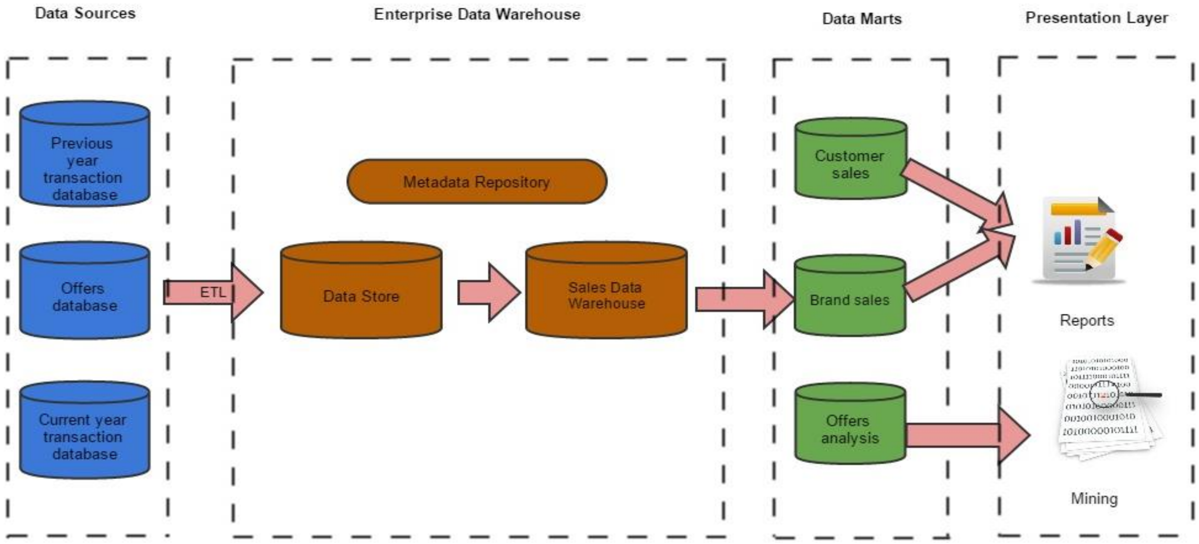
Advantages of using SAP HANA:

- SAP HANA has in-memory relational database management system.
- Loading the data is easy.
- Generation of end user reports/dashboards is easier
- Parallel processing
- Implementation and Integration of tables is easier.

Recommendation of tools:

For generation of reports they can use either SAP HANA Studio or Pivot table. SAP HANA Studio has more advantages than Pivot table. So, we recommend them to use SAP HANA Studio.

Architecture diagram



ETL Process:

Extract: Extract step involves the extraction of data and make it accessible for further processing. Source of data for our project is from kaggle.com (http://www.kaggle.com/c/acquire-valued-shoppers-challenge/data) in the .csv file format. Three files extracted are 'Offers', '2012 sales transaction' and '2013 sales transaction'. Once the data is extracted from the source we analyzed the data captured in the data set to know which business questions can be answered. Each row represents the sales transaction. In our data demographics related to customer is not present but we recommend this as add-on for building the data warehouse.

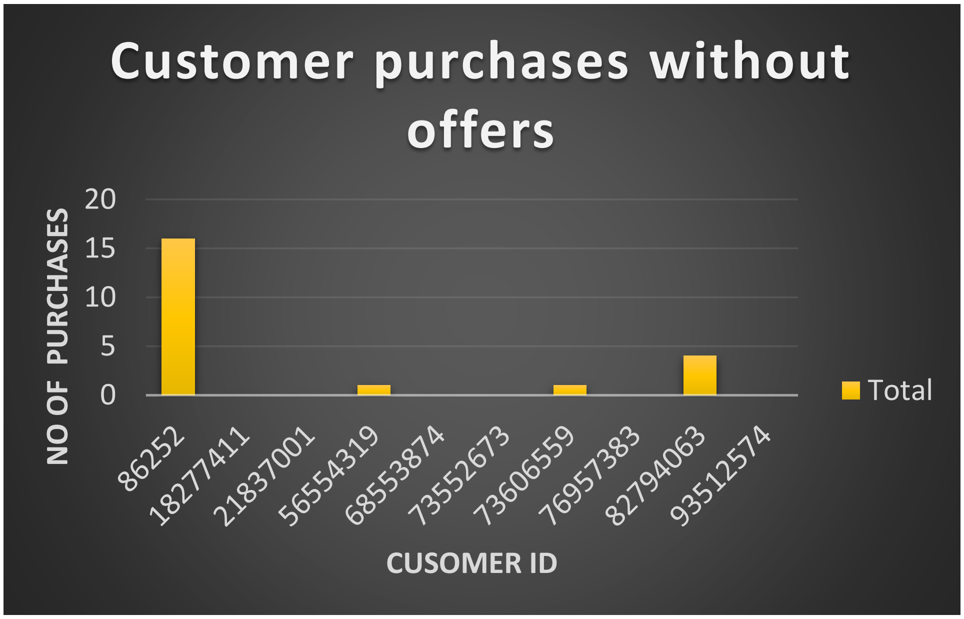
Transform: Transform step involves the process of conversion of extracted data from the previous form into the form that can be placed into another table for analysis. In our case, first we cleaned the given data using MySQL Workbench. Second, we grouped the tables 'Offers' and '2013 sales transactions' using the offer ID as in '2013 sales transactions' does not contain Product ID but the 'Offers' table has Product ID. This enabled us to see which product is related to that sales transaction. We added few columns which enables us to capture more information related to the business process. We converted the final tables into 'csv' format with semicolon delimiter which help us to write the data into target database.

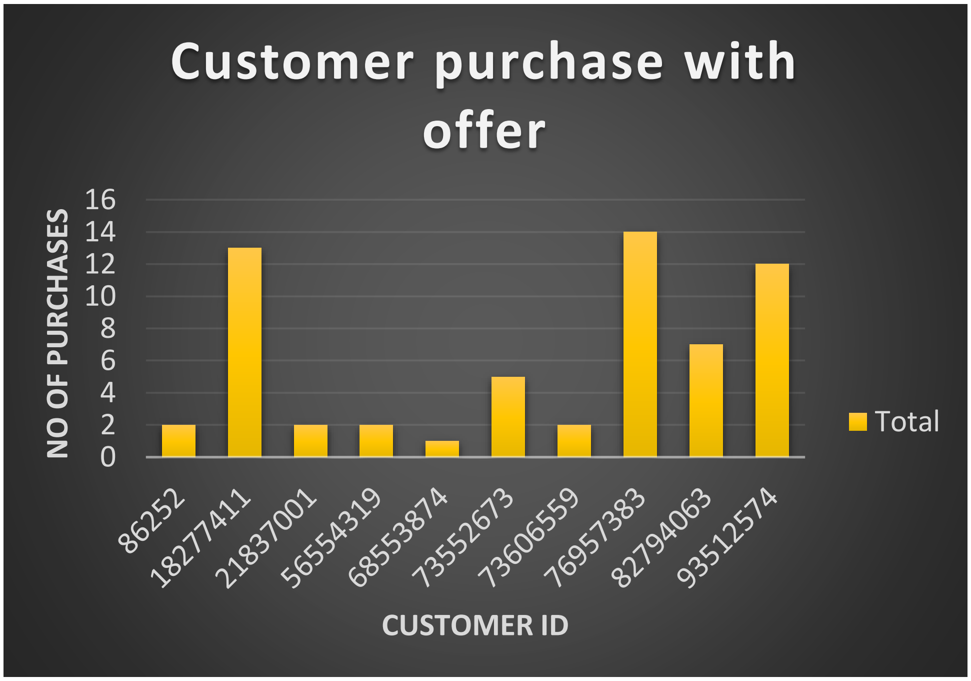
Load: Load step involves writing the data into target database. In our case, we have loaded the converted data into SAP HANA Studio. Reports were generated after creation of the dimensional model in the SAP HANA Studio.

Appendix:

Analysis and Reporting

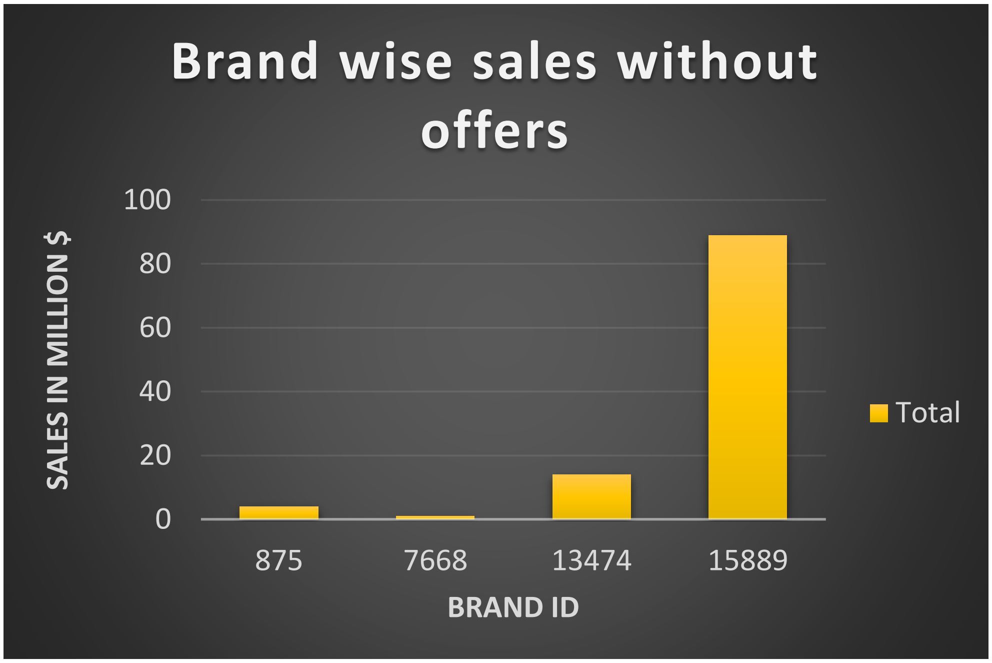
Sample BI Reports –

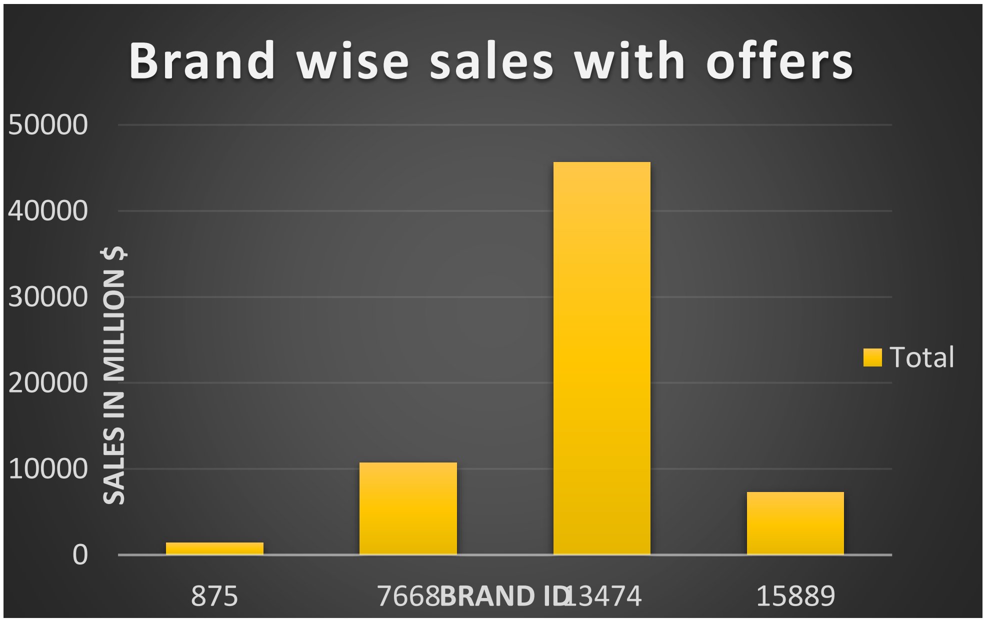




In the above graphs we have taken 10 customers and analyzed their purchasing behavior. Here, we see that, very few customer bought few products before promotion was applied. In the second graphs, we see that customers purchased more products when the promotion was applied.

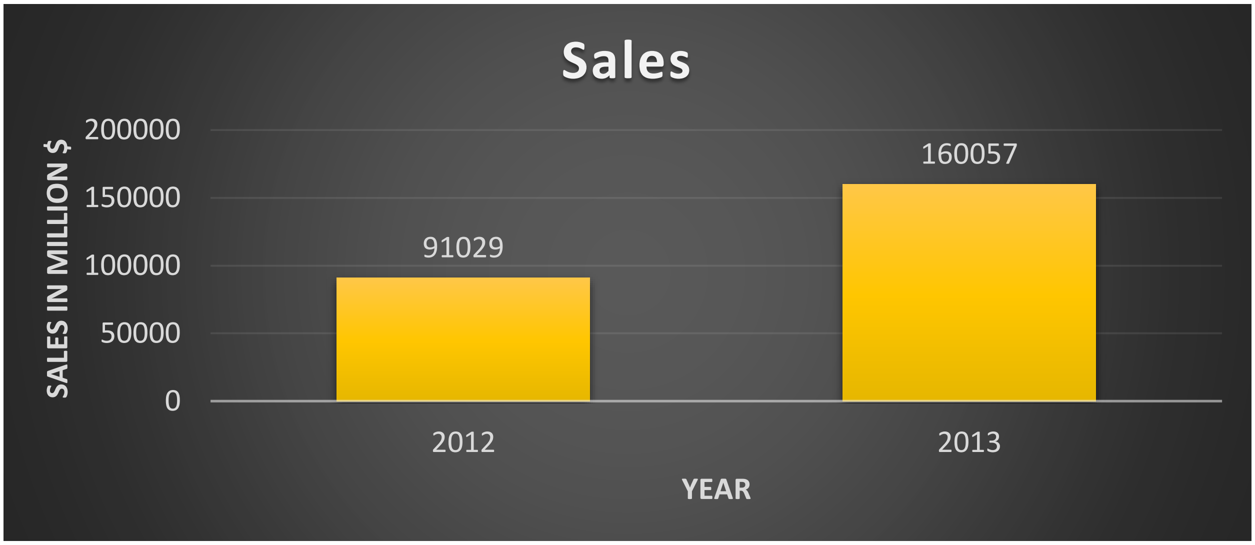
2) Difference in the sales of a brand when a discount was applied





We can say that the sales of the 4 brands have increased dramatically when offers were promoted on the brands.

3) Change in the net sales year wise after the promotion was applied:



We can conclude that, 2012 has less sales compared to the year 2013, where promotions were offered to the customer in the year 2013.

4) List of customers who did not avail offers

id	offer
100017875	1200581
100033247	1200581
100043455	1197502
100051423	1197502
100053952	1204821
100072053	1197502
100084808	1197502
100166617	1197502
100172712	1200581
1001966139	1197502

Bibliography

The Data Warehouse Toolkit – A definitive guide to dimensional modelling, By – Ralph Kimball and Margy Ross

http://www.kaggle.com/c/acquire-valued-shoppers-challenge