Design and Development of Data Warehouse for Analysis of Supermarket Data using BI Tools

Abstract — This study represents the design and development of data warehouse for supermarket data. The analysis provides the sales and profit generated by the supermarket based on the data such as product, customer and time. In this paper, analysis of profit generated with the sales is provided along with various structures to support the reasoning. Dataset of customer contains considerable information from various cities of Canada. Different products based on their shelf life and stock available is considered. The data is modelled in different dimensions. The data needs to be cleaned before applying to the analysis tools. This paper uses Pentaho for analysis and results are generated by modifying the mdx queries using different charts available. Such a model is necessary for making key decisions and strategic planning for a supermarket. The results provide with the representation of data with respect to four measures considered, namely, actual and original sales, profit and discount with respect to various dimensions.

Keywords: Data Warehouse, Extraction, Transform, Pentaho, Mondrian, Supermarket

1. Introduction

The expansion of supermarket chain at an alarming rate has resulted in a progressively competitive market. Computer terminals are mainly set up to process big chunk of everyday data, and this data is significantly increasing day by day. Typically the storage capacity of traditional DBMS system is in the range of millions and is a major bottleneck for supermarket with huge amount of heterogeneous and decentralized data to store and process as the resources are waste. Even with the availability of some data warehouse products such as Oracle Warehouse Builder, IBM Visual Warehouse, and NCR designed for Walmart stores, it becomes difficult for many growing supermarkets to buy these and access without any language barrier or lack of skill. Customization of data warehouse system is required for making decision as per the customer needs, buying trends, operation efficiency, favourable customer experience, loyalty and retention of customers along with anticipating demand for efficient inventory management, cash management and increasing overall profitability. These insights provide decision makers with a competitive edge over other and can help in business growth. Data warehouse mainly enables to make informed decisions rapidly with key initiatives and thus saving time. It also enhances operational efficiency by analysing data from various sources and with BI solution. For retailers it is essential to demand forecasts as the scale of operation is not constant for them for the entire year. In addition, data warehouse in retail can provide significant analytics which can be helpful in customizing offer and marketing. Pricing, marketing plans and even stock prices are highly dependent on the generated sales and developing a warehouse for the retail sales typically requires an in depth understanding of the type of business and effect of sales to pricing strategy and discounts.

Mainly the retailers are worried about efficient use of space, labour and cost. Profit is generally calculated from sales volumes and margin profit for each product sold. By analysing this information the retailers can pile up the highest selling products or offer promotional discounts. A data warehouse for retail sales should be the basis for profitability analysis. Large retailers are motivated to create better warehouses for tracking sales due to the competitive market.

2. Literature Survey

The paper [1], provide analysis of various functions in retail organization and imply that business intelligence is of utmost importance. In addition it proposes that BI can significantly improve various functionalities in retail sector such as finance, HR management etc. Authors in this paper [2] provides analyses the increasing quantity of data and provides this analysed information of past and integrated data to admins at senior positions to make use of available information and thus to help in decision making. In the paper [3], authors present the entire procedure for designing and development of data warehouse architecture and focus on main points to consider while data warehouse building that comprises of characteristics of data store to modelling techniques and the main principles that has to be taken into consideration for efficient DW architecture. Creation of multiple data marts with multidimensional data model along with the design of the data cleaning and transforming process for populating the data marts from data sources is provided by authors in the paper [4]. In addition, authors also have incorporate data quality checking on the data source and data detection rules to filter out unmatched data schema and data range from being stored in the data warehouse for analysis. OLAP system was designed in this paper [5] by using web technique, through which knowledge of decision making was displayed on web interfaces. This system provides effective support for manager's decision making with respect to supermarket sale. This paper [6] proposes a methodology for requirement analysis in the perspective of data warehouse systems denoted by mainly integrated requirement analysis for designing data warehouse (IRADAH), provided as a method to accomplish the unusual aspects integral to such systems. Authors provide an outline of BI as the key technology, and the establishment & application of Business Intelligence System in retail industry [7]. Authors in the paper [8] proposes that retail records of data resource are beneficial for effective decision support in the data warehouse promoting sales in a supermarket. In [9], a business intelligence system is proposed which used the data in the supermarket corporation affair database, and then transformed it into information followed by subliming the information into knowledge. The system provides efficiently the support for making decision of governor in supermarket supplier management. In the paper [10], a new technology for analysing namely data mining, is used in supermarket administration, by which management of supermarket is transformed to knowledge management. This paper [11] focuses on general DM technology and its application in operation of supermarket and discusses the specific use of DM in the process of customer relationship management.

3. Data Warehouse Architecture

Construction of data warehouse starts with the relationship between multiple layers of data warehouse architecture. Data warehouse architecture for retail industry helps us to understand the basic overview which can be optimized for supermarkets. Data is obtained from multiple sources in the form of flat files, databases or sheets etc. This mainly contains data about finance, marketing, demand planning, forecasting, logistics, RnD and sales data. This data is extracted from the sources and various data cleaning operations are implemented on this data. These consist of filtering, cleaning, joining, splitting and sorting. This transformed data is then fed to data storage and aggregation layer which consists of data warehouse and several data marts. It contains operational data metadata and data marts for various domains such as product portfolio, logistic, finance and overall goals and strategies in the business. The data is then presented to the decision makers in the form of operational reports, analytics, dashboards, alerts and scorecards. Figure 1 provides with the overview of the data warehouse architecture in retail industry.

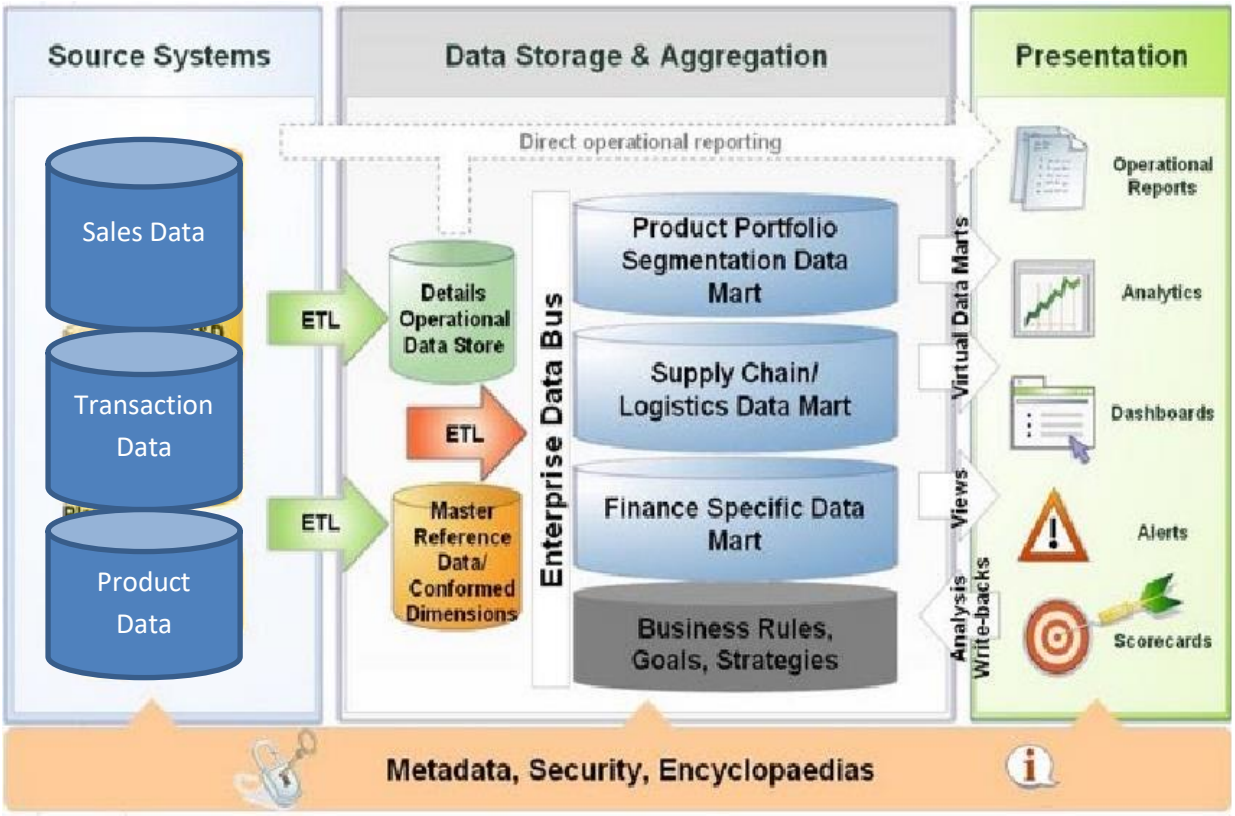

Fig 1: Data warehouse architecture overview for retail industry

Supermarket industry mainly focusses on increasing sales and demanding forecasts. Detailed data warehouse architecture for supermarket industry is shown in figure 2.

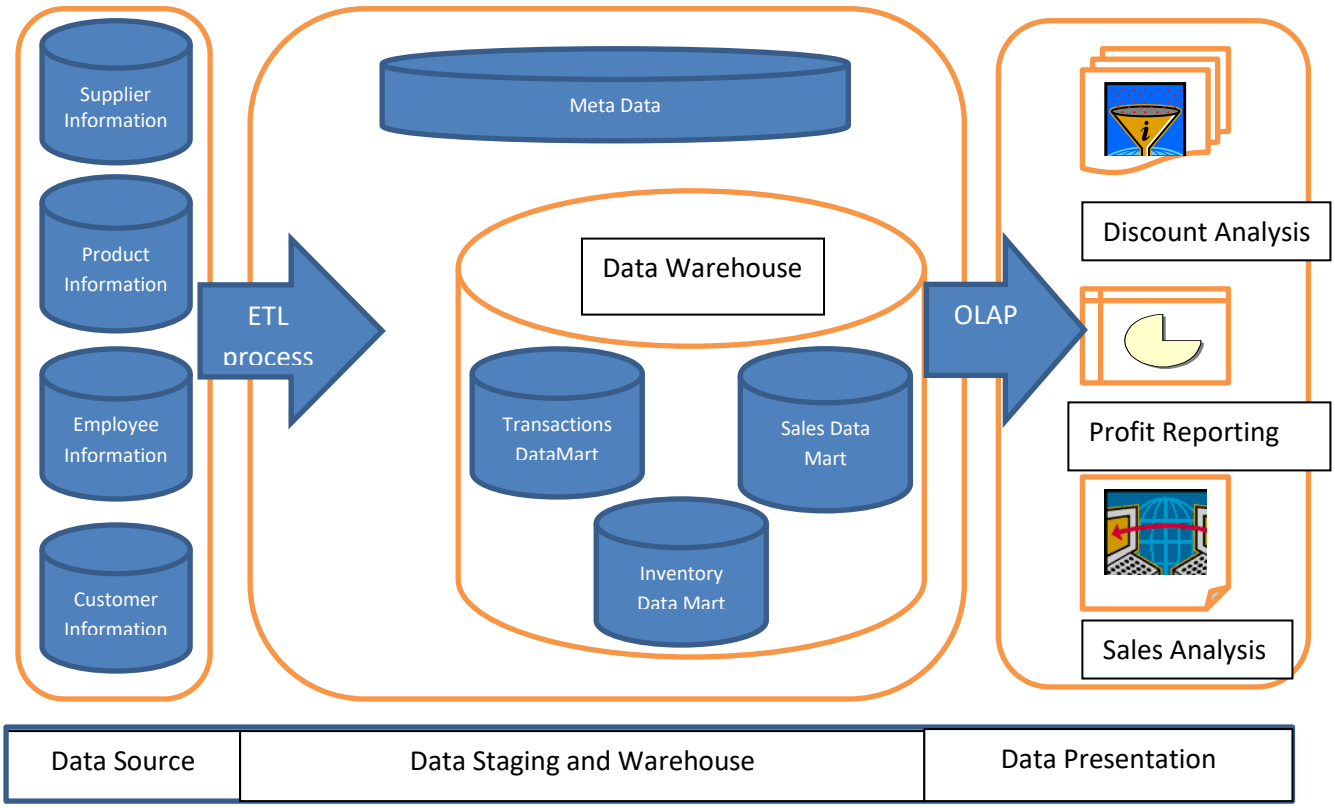

Fig 2: Data Warehouse Architecture for Supermarket

The top-down approach is used to design the data warehouse. The first step is to organize the data sources. Multiple heterogeneous data sources are used. The data maybe present in existing operational systems, flat files, and other external sources. For this paper data sources include data from supplier, customer, employee and product. The next step is data staging and building data warehouse. The Extract Transform Load (ETL) process takes place to help integrate the data into the data warehouse. It involves extracting the data from the data source; this might be a database, flat file, web service or other sources. After extraction, the data is cleaned, this consists of detection and removal of invalid, duplicate or inconsistent data to help improve the quality and utility of the data before it is transferred to the data warehouse. In the next phase, the data is transformed. Here, the data is transformed into format that can be stored and processed by the data warehouse; this requires standardizing a data type for the particular attribute being transformed. Depending on the number of sources, degree of heterogeneity and number of errors in the data, multiple data cleaning and transformation steps will be required. In the final phase of ETL, the extracted and transformed data is loaded into the Data Warehouse; i.e. the fact and dimension tables. At this stage the data warehouse is ready for use for analysis and report generation. The data sources for this case study of supermarket are flat csv files having fact table as sales and dimension tables of product, customer, and time dimension.

4. Schema Design

Tables and the relationships between them can be represented using schemas. Various types of schemas such as star schema, snow flake schema, fact constellation schema are available.

The schema for a particular warehouse is chosen according to the source data model and user requirements. The most simple of these schemas is star schema which has comparatively lesser tables and well defined join paths. Compared to normalized schema that is used for OLTP, this schema provides faster response time and query execution time. A star schema is a physical entity in the database with one fact table in the middle surrounded by the dimension table around. Data in fact table can be treated as permanent and read only whereas the dimension table data can change over a period of time.

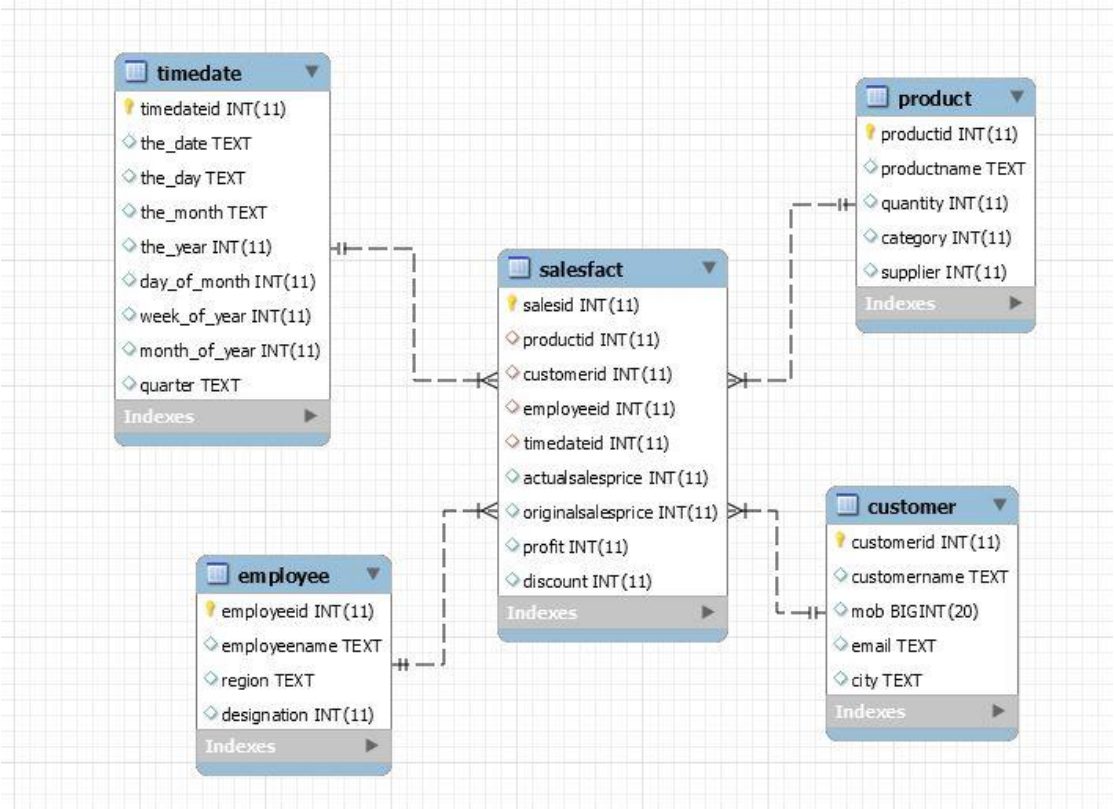

Fig 3: Star Schema for supermarket data warehouse

The OLAP uses the denormalized structures for schema design for faster processing. This helps in faster execution plan. Denormalization means that the normalization is not done on the dimension tables enabling familiar views to the end user and reducing the complications of breaking the table. A graphical representation of a star schema for sales is shown in figure 3. Initial data warehouse design for the supermarket sale case study contains the product, customer, employee, time dimensions, and one fact table using facts like the sales, and the profit. It forms a basis which can be used by any retail industry with appropriately adding the required dimension. Each dimension can eventually grow in size as per the requirements.

5. Development of Data Warehouse

The first step after the design in the development of data warehouse is ETL. It stands for Extract, Transform and Load. At this stage, data is gathered from various sources and then the cleaning process takes place. This data includes data of product information that consists

of product name, quantity, category if perishable or not perishable and based on shelf life of products. Also the type of product depends on the supplier as shown in table 1.

Table 1: Product Dimension Data

productid	productna	quantity	category	supplier
1	Milk	115	1	1
2	Curd	172	1	1
3	Coke	163	1	1
4	Rice	170	2	1
5	Flour	132	2	1
6	Spices	145	2	1

Table 2: Customer Dimension Data

customeri	customername	mob	email	city
1	AlexJames	8711738367	demo@gmail.com	Toronto
2	RichardDavid	8124138884	demo@gmail.com	Ottawa
3	JamesBrad	8985457157	demo@gmail.com	Vancouve
4	MikeLew	9600477089	demo@gmail.com	Calgary
5	TimDek	9983312816	demo@gmail.com	Toronto
6	DemiJake	8660453158	demo@gmail.com	Vancouve

Table 3: Time Dimension Data

timedatei	the_date	the_day	the_mont	the_year	day_of_m	week_of_	month_of	quarter
1	07-01-1998 00:00	Wednesd	January	1998	7	4	1	Q1
2	08-01-1998 00:00	Thursday	January	1998	8	4	1	Q1
3	09-01-1998 00:00	Friday	January	1998	9	4	1	Q1
4	10-01-1998 00:00	Saturday	January	1998	10	4	1	Q1
5	11-01-1998 00:00	Sunday	January	1998	11	5	1	Q1
6	13-01-1998 00:00	Tuesday	January	1998	13	5	1	Q1
7	14-01-1998 00:00	Wednesd	January	1998	14	5	1	Q1

The next dimension contains the customer data that includes name, email, mobile number and the city as represented in table 2. Time dimension consists of the quarters, year, day, month and time. This is shown in table 3. The fact table of sales consists of various foreign key from all dimension tables as shown in table 4 along with the different measures. The Employee dimension from the design phase has been omitted further in the development in this paper due to complexity. This data is in the form of CSV files. The fact and dimensional data is imported in supermarket database. Pentaho workbench tool is used for Mondrian Schema. A Mondrian schema basically consists of a model comprising cube with different hierarchies and levels along with the specified measures. Figure 4 shows the representation of the Mondrian Schema.

Table 4: Sales Fact Table

salesid	productid	customeri	employee	timedatei	actualsale	originalsa	profit	discount
1	2	7	4	21	61	63	2	40
2	15	7	1	20	79	81	2	9
3	28	1	2	8	151	153	2	25
4	14	13	4	21	103	105	2	26
5	25	13	5	3	46	48	2	14
6	27	4	4	5	193	195	2	17
7	23	4	4	12	60	62	2	1
8	29	13	3	20	163	165	2	6
9	2	11	2	24	130	132	2	6

Mondrian Schema is verified using MDX Query. MDX stands for Multidimensional Expressions. It can be used to query multidimensional data. All the manipulation to obtain the different analysis representation is using MDX queries. For obtaining results on supermarket data the following MDX query is considered.

MDX Query: Select {[Measures].[Original Sales], [Measures].[Actual Sales], [Measures].[Discount], [Measures].[Profit]} on columns, {[Customer].[All Customers], [Product].[All Products], [TimeDate].[All TimeDates]} ON rows from
Sales

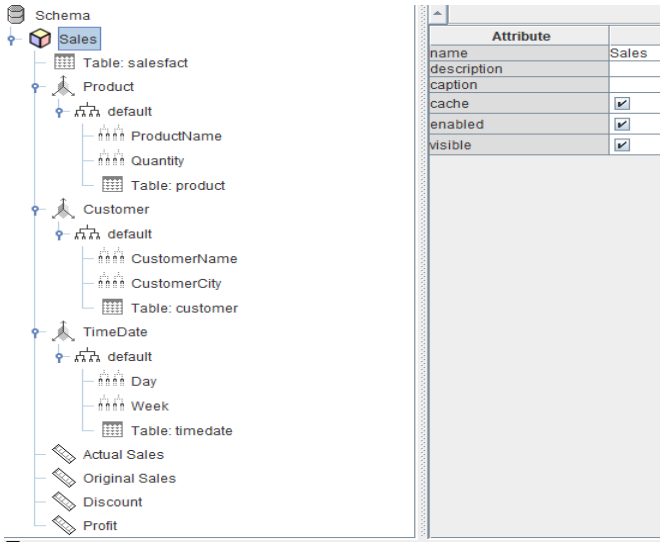

Fig 4: Supermarket Mondrian Schema

This schema is stored as XML file which is published on Pentaho server for analysis using JPivot.

6. Experimental Results

For the experimental results one thousand two hundred samples have been considered in the sales data mart i.e. part of supermarket data warehouse. For the purpose of the experiments, the pseudo data is generated in a scientific manner to resemble supermarket data. Pentaho

Data Integration and JPivot view on Pentaho BI tool is used for transforming and showing supermarket information in tabular as well as chart format. Various representation is shown in the below figures for the data.

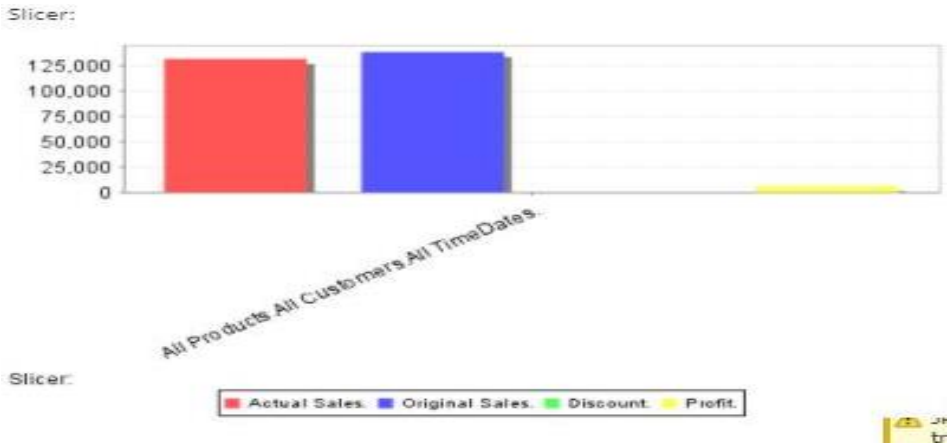

Fig 5: Bar Graph for all Measures on all dimensions

Figure 5 shows all the four measures considered with respect to all the dimensions in the cube and represented in the form of a bar graph.

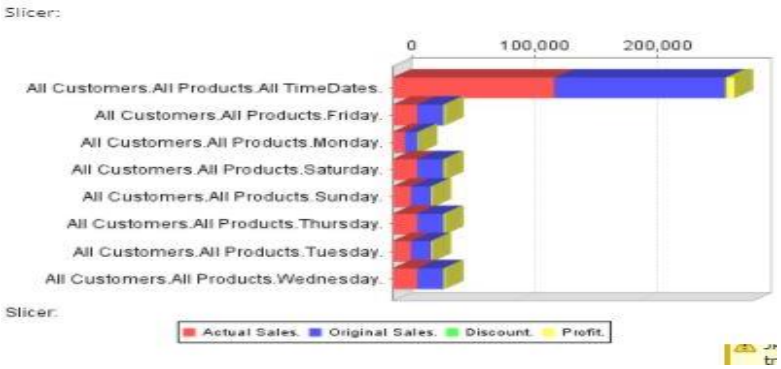

Fig 6: Stacked horizontal bar for all days of the week

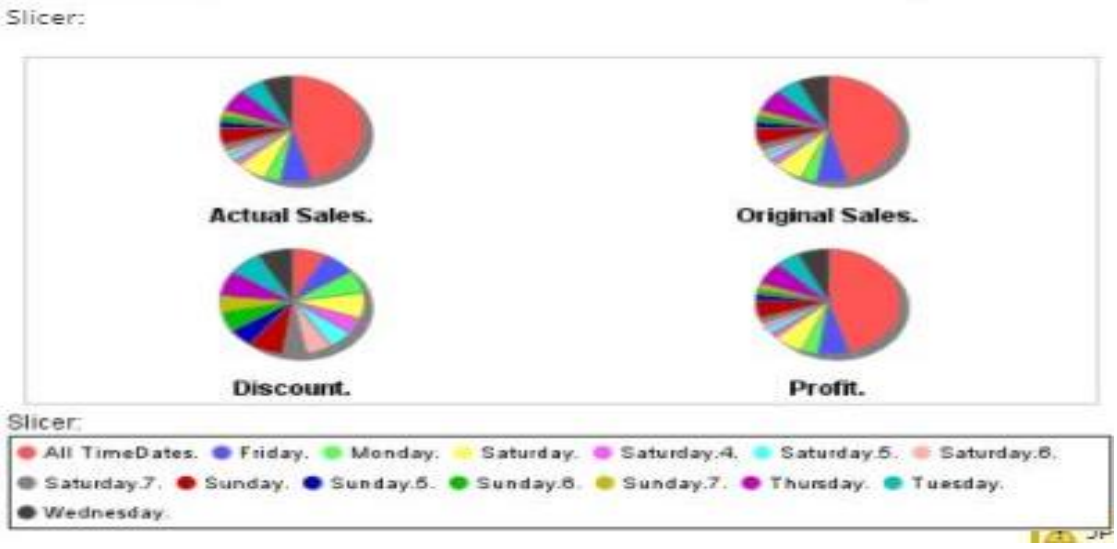

Fig 7: Pie chart of four different measures based on Time Dimension

Figure 6 displays horizontally stacked bar graph of sales, discount and product with respect to all seven days of the week. Pie chart is represented in figure 7 displaying the analysis of all the measures for each day of the week and two days, i.e. Saturday and Sunday are further categorized as the week in the year. Thus representing the analysis based on the time dimension.

Figure 8 shows the pie chart for profit and discount based on the seven days of the week. Slicing and dicing operations can be performed by modifying the MDX queries. One such operation on the product for Sunday and 4th week is shown in below figure 9.

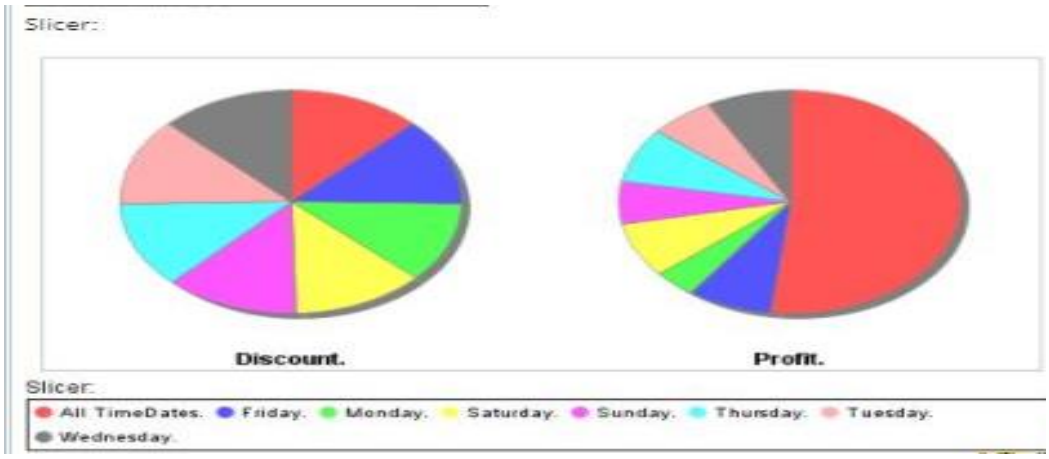

Fig 8: Pie chart for discount and profit for seven days of the week

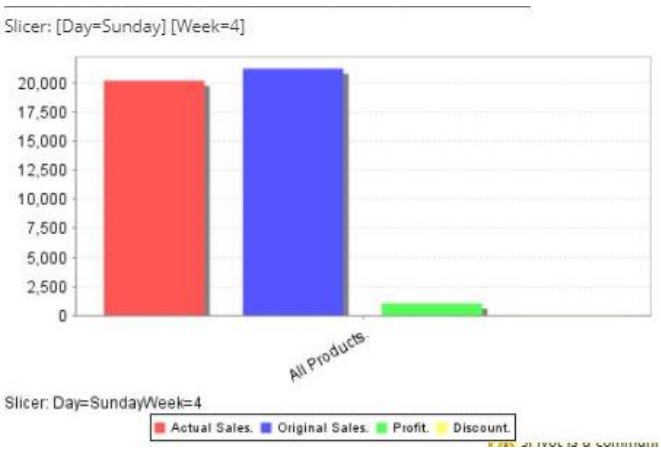

Fig 9: Slicing operation on the cube

7. Conclusion and Future Work

Data warehouse and Pentaho tools helps to overcome the grouping of data in current data source. Additionally the results generated can be viewed through multiple different graphs and charts which are easier to manage. Three dimensions are considered and all the graphs are generated for the measure based on these dimensions. Various operations are performed on the cube using Multidimensional MDX Query. Pentaho Business Intelligence server is used to represent the data processed in Pentaho workbench. This helps the managers to make strategic decisions for the efficient and smooth running of their businesses. This helps the

industry to maintain competitive edge over others and to increase overall profitability. Similarly the design can be extended to other factors affecting the sales and efficiency of supermarket including supply chain management and customer satisfaction can be analysed by adding more cubes and dimensions to the existing warehouse.

References

1. Haigang Li, "Applications of data warehousing and data mining in the retail industry," Proceedings of ICSSSM '05. 2005 International Conference on Services Systems and Services Management, 2005., Chongquing, China, 2005, pp. 1047-1050 Vol. 2, doi: 10.1109/ICSSSM.2005.1500153.

2. H. Gu and C. Lin, "Application of data warehouse techniques in retail trade," 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, Sichuan, 2012, pp. 2594-2597, doi: 10.1109/FSKD.2012.6233966.

3. M. Rifaie, K. Kianmehr, R. Alhajj and M. J. Ridley, "Data warehouse architecture and design," 2008 IEEE International Conference on Information Reuse and Integration, Las Vegas, NV, USA, 2008, pp. 58-63, doi: 10.1109/IRI.2008.4583005.

4. B. K. Seah and Nor Ezam Selan, "Design and implementation of data warehouse with data model using survey-based services data," Fourth edition of the International Conference on the Innovative Computing Technology (INTECH 2014), Luton, 2014, pp. 58-64, doi: 10.1109/INTECH.2014.6927748.

5. X. Hong, L. Zai-wen and M. Hai-yang, "Study and Realization of Supermarket BI System Based on Data Warehouse and Web Technique," 2008 International Conference on Computer Science and Software Engineering, Hubei, 2008, pp. 482-485, doi: 10.1109/CSSE.2008.877.

6. Munawar, N. Salim and R. Ibrahim, "Towards data warehouse quality through integrated requirements analysis," 2011 International Conference on Advanced Computer Science and Information Systems, Jakarta, 2011, pp. 259-264.

7. Tong Gang, Cui Kai and Song Bei, "The research & application of Business Intelligence system in retail industry," 2008 IEEE International Conference on Automation and Logistics, Qingdao, 2008, pp. 87-91, doi: 10.1109/ICAL.2008.4636125.

8. Xie Wu and Huimin Zhang, "Design and implementation of data warehouse of minor chain supermarkets," 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems, Xiamen, 2010, pp. 828-830, doi: 10.1109/ICICISYS.2010.5658331.

9. H. Xue, P. Guo, H. Zhang and B. Kang, "Study and Realization of Supplier Business Intelligence System for Chain Supermarket," 2009 International Conference on Computational Intelligence and Software Engineering, Wuhan, 2009, pp. 1-4, doi: 10.1109/CISE.2009.5366538.

10. Y. Yue, T. Zhang and X. Xu, "Analysis on the effect of data-mining to supermarket," 2010 5th International Conference on Computer Science & Education, Hefei, 2010, pp. 1852-1854, doi: 10.1109/ICCSE.2010.5593809.

11. Z. Aiguo, J. Lanling and S. Ping, "Application of Data Mining in supermarket," 2011 International Conference on Consumer Electronics, Communications and Networks (CECNet), XianNing, 2011, pp. 1082-1085, doi: 10.1109/CECNET.2011.5769080.