An in-depth investigation across different product sales

Keith Conti Borda
Institute of Information & Communication Technology
University College
MCAST, Paola PLA 9032
keith.conti.a100360@mcast.edu.mt

***Abstract*—Investigating across the different sales of various products in different areas by different individuals. The aim of this project is to analyze the different buying behaviours, to try and identify certain buying patterns for different products, buyers and also supply stores from the data set which has been provided to us. The data set consists of a combination of categorized products, orders, stores and clients, from which we have to extract the data and insert into our customized cube. The proposed solution includes a custom OLAP cube which has been implemented to cater for the data stored in the original data set after being cleansed for duplicates. This project's outcome has resulted in a positive one, since all the migrations of data have been executed successfully. A Windows based system was used for this project evaluation, together with SQL Server Management Studio, Microsoft Excel and the Oracle Data Modeller to design the ERD's.**

***Index Terms*—analysis, investigation, dataset, data cleansing, operations**

I. Introduction

With regards to this whole project, alot of effort and energy has been put in by our tutor, in order to provide us with the dataset, and all the necessary material in in order for us to be able to work and implement the required operations for this scenario. The aim of this entire project was to analyze the different buying behaviours, to try and identify certain buying patterns for different products, buyers and also supply stores from the data set which has been provided to us by our tutor. Having this data stored in the dataset, we were required to compile our own SQL Scripts in order to query the dataset, to collect data and information we need in order to satisfy the requested reports.By having done this, we were able to identify who bought which product,from which category, from which store in which region, during what year, month, quarter and even day. All the data provided to use by our tutor inside the olap database, was filtered thoroughly, and then was migrated into our own custom OLAP cube, whereby the data may be displayed in a similar manner, however cleansed.

II. Literature Review

It has become a vital and extremely challenging for organizations to sustain their position within a competitive market and also to understand the customer. Most importantly, technological improvements have paved a way in order to process queries rapidly and efficiently, thereby minimizing the waiting time. Data mining utilities have become the most important tools in order to analyze and filter the huge amounts of data to make the correct decisions. The objective of this paper is to analyze big data together with the consumer's behaviours and also to make the correct decisions leading to competitive rivals.

The database technology we know of today has been changed drastically since the 1980s. Research and development activities on new and powerful database systems has advanced much more than it used to be, thereby employing the advanced data model. The extreme growth of computer systems, together with the hardware and software technology in the past three decades has led to a large supply of powerful, and and cost effective computer systems, data collection equipment and also storage media. Thanks to all this, a huge number of databases and data repositories were made available for transaction management information retrieval and data analysis. The online analytical processing (OLAP) withholds analysis techniques with functionalities such as summarizing,consolidation and aggregation together with the view of data from different angles. These OLAP tools have been used for in depth analysis such as classifications of data, clustering and the data changes over time. The term 'Data mining' refers to the extract of knowledge from huge amounts of data. This is a simple step in the knowledge process to extract knowledge store inside databases.

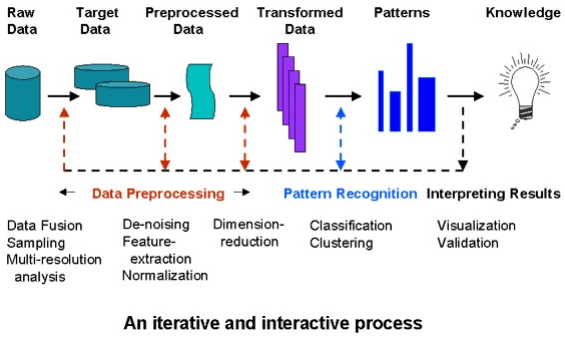

Fig. 1. DM Diagram

The forecasting and sales analyses of certain products and services makes a difference in the profit and loss per accounting period for the organization which then determines

the success or failure of the organization itself. Therefore, the reliable prediction of sales is of utmost importance. This article presents the sales forecasting approach by the integration of fuzzy systems and data clustering to construct a sales forecasting integrated system. All clusters are entered into an independent GFS model with the ability of rule base extraction and database tuning. The performance of an extracted expert system is compared against the previous sales forecasting methods using mean absolute percentage error (MAPE) and root mean square error (RMSE). These results show that the proposed approach outperforms the other previous approaches.

Market sales is the only factor than measures the outcome of marketing efforts and therefore the sales forecasting are extremely important when it comes to marketing planning. Within the competitive market each business faces, sales forecasting is a very helpful tool, because an organization may use this to its advantage to have a rough idea of what sales to expect based on current of past figures. A reliable sales predication will improve the organizations sales because it entitles production, sales, marketing, and finance sections to effectively develop programs to continue enhancing the organizations size. Examples of these software programs are: Sales planning, budgeting, promotion and advertising plans.

Market basket analysis is also known as association-rule mining. This terminology is a very useful method of discovering customer purchasing patterns data from stores transactional databases. Since the information obtained from the analysis may be utilized to form marketing, sales, service, and operation strategies, it increased the research interest. However, the existing methods may fail to discover important purchasing patterns in a multi-store environment, because of an assumption that products under consideration are on shelf all the time across all stores. Within this paper, a new method is being proposed to overcome this weakness. This evaluation shows that the proposed method is efficient, and has an advantage over the traditional method when stores are varied in size, product mix changes rapidly over time, and larger numbers of stores and periods are considered.

Thanks to all the advances in information technology we have today, organizations may effectively gather and store transactional and demographic data on individual customers at very reasonable costs. The most popular challenge for all corporations is that of how to extract customer data information from their vast customer databases to gain a competitive advantage. This Market Basket analysis is a method of how to discover customer purchasing patterns by extracting data from their transactional databases. The result of this method may be rather biased since a product may be on shelf before its first transaction or also after the last transaction. This product may also be put on shelf and taken off shelf many times during this market basket data collection period. This is the first problem within this research. The second problem is that of finding the common association patterns in subsets of stores. To overcome these problems, an algorithm was developed in order to automatically extract association rules in a multi store environment, however the rules also contain the information on store together with the time where the rules hold.

Since the time and store location factors will be considered, the rule generation is much more complicated than the algorithm mentioned above. Within this paper, the simulation results which show the proposed method that is computationally efficient and has an advantage over the traditional association method, when stores under observation are large in size and have large mixes of products that are altered rapidly over a period of time.

III. Research Methodology

The CRISP DM terminology stands for cross-industry process for data mining. This methodology offers an approach that is structured to plan a data mining project. The crisp-dm methodology is powerful, practical, flexible and useful when utilized to its advantage to solve business issues. The CRISP-DM model is shown hereunder. [1]

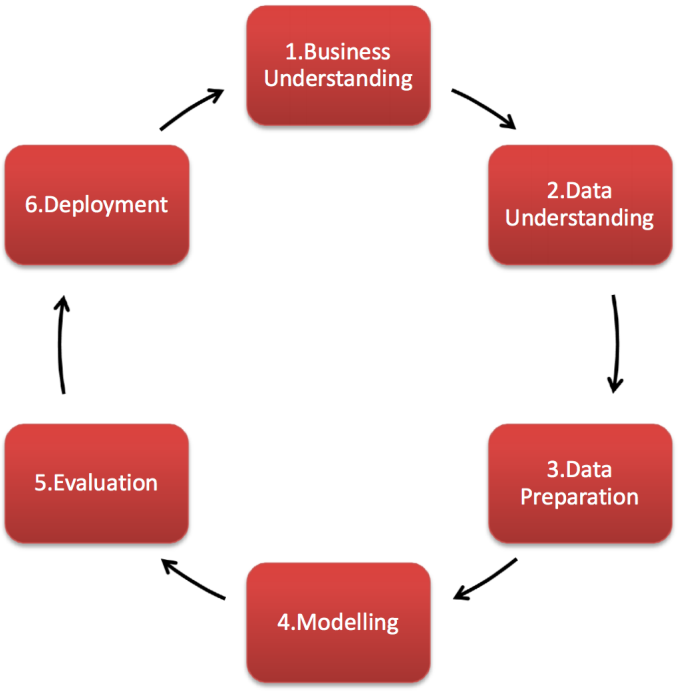

Fig. 2. crispDM

This model is made up of a sequence of stages, which may be performed in different order and will often be required to backtrack to previous tasks and repeat certain actions. This model does not capture all the routes through the data mining process. The 6 stages that make up the crisp-dm are as follows:
Business Understanding
Within the first stage of the crisp-dm, one has to understand what has to be accomplished from a business perspective,

and after this happens the knowledge is converted into a data mining problem definition. The goal of this stage is to uncover important factors which may influence the outcome of the project. Ignoring this step may drop all the effort made thereby producing the right answers to the wrong questions. Data Understanding

The second stage of the crisp-dm starts with an initial data collection and proceeds with activities in order to familiarize along with the data, to identify the problems with data quality, find first insights in the data, or even to detect interesting subsets to form hypothesis for hidden information. Data Preparation

The third stage of the crisp-dm is the data preparation. This stage covers all activities to build the finalized dataset. These preparation tasks are performed multiple times, and not in any prescribed order. These tasks mentioned include tables, records and also attribute sections together with transformation and the cleaning of data for the modelling tools. Modelling

The fourth stage of the crisp-dm is that of modelling. There are various modelling techniques that are chosen. Typically, there are many techniques for the same data mining problem type. Some of the techniques have specific requirements when it comes to the form of the data, thereby we would have to step back to the data preparation phase. Evaluation

Within the fifth stage of the crisp-dm methodology, the model built will appear to have a very high quality, from a data analysis perspective. Before advancing to the final deployment of the model, it is very important to have an even deeper evaluation on the model itself, and after that review the steps used to build the model to be sure it achieves the business objectives specified. At the end of this phase, a decision on the use of the data mining results must be reached. [2]

Deployment

Finally, the sixth stage of the crisp-dm methodology. The creation of the model is not the end of the project. Even though the purpose of the model improves the knowledge of the data, the knowledge gained must be organized and shown in a certain way that would be beneficial to the customer. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a data mining process. In many cases, it will be the customer, not the analyst, carries out the deployment steps. Even if the analyst deploys the model it is important for the customer to have a good understanding of the actions which will need to be carried out in order to actually make use of these created models.

In order to be able to design the 3D OLAP cube, we had to first create our own ERD from the original OLTP ERD. In my case, my erd consisted of 5 tables including the Order Fact table. Once this was complete, I then designed the 3D OLAP as per below diagram.

The OLTP ERD provided to us by our tutor shows all the tables that may be found inside the database, including all their rows and data types, primary keys, foreign keys in addition with the way they are linked to each other. This gives us a

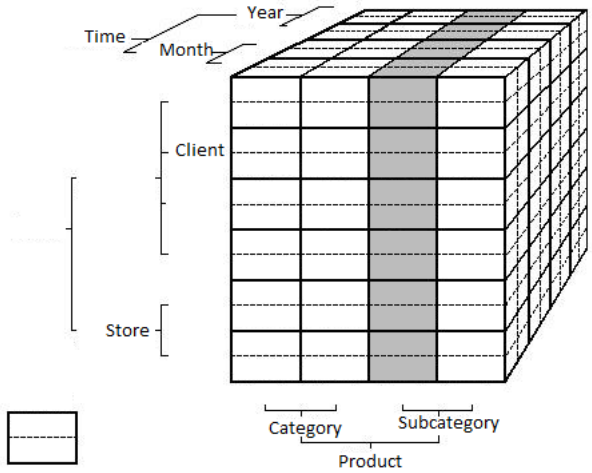

Fig. 3. OLAPTemplate

better understanding of how we can aggregate the data from more than one table to another. The original OLTP ERD may be seen hereunder;

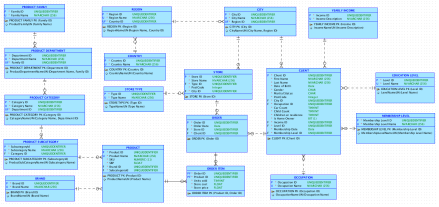

Fig. 4. Original OLTP ERD

What we had to do was the following; We created our own ERD, by choosing a dimension of our choice. In my case, my fact table was the Orders fact, which is connected to another 4 four tables being Client, Store, Time and Product. Obviously, the tables I chose were taken from the original ERD (OLAP), together with their respective columns and data types too. This entitles us to query and display the data from another dimension; thereby we can say what type of client bought which product, from which store in which region, during what year, month or even quarter and also on which day. The ERD I have created may be found hereunder;

With regards to the preliminary analysis and data cleaning, a script was built in order to gather all those duplicate records inside the OrderItem table inside the database.

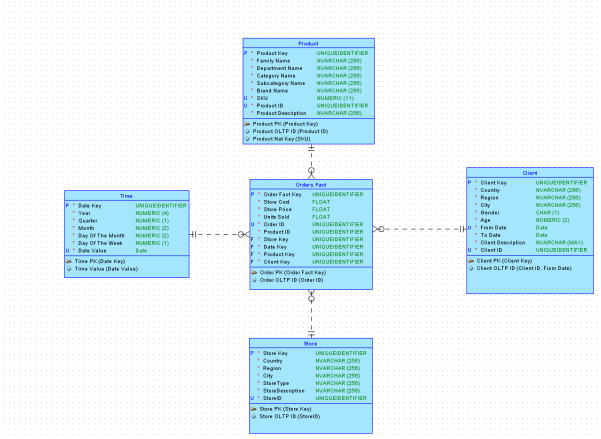

Fig. 5. Order Fact ERD Dimension

```sql
-- Preliminary Analysis to check for duplicates
SELECT      productId
            , orderId
            , COUNT(*) AS 'Count'
FROM        [oltp].[orderItem]
GROUP BY    productId, orderId
HAVING COUNT (*) > 1
```

Fig. 6. Code For Duplicates

The duplicates found by executing the above SQL query were as follows:

Once these duplicates were analyzed, another SQL Query was built in order to view only those duplicate records inside the OrderItem table in the database.

By making use of a Common Table Expression (CTE), the duplicate records inside the database were removed, thereby only having unique values in the table without any duplicated values.

ETL is used to migrate data from one database to another, to form data marts and data warehouses and also to convert databases from one format or type to another. ETL is an acronym which stands for Extract, Transform and load. These three are database functions that are combined into one tool and are used to extract information from the database, and insert it into another database or table.

Extract the process of reading the data stores inside a table in the database.

Transform the process of converting the gathered data from extraction into the form it needs to be stored in so that it may be placed inside another table or database. The transformation happens by using the lookup tables or by combining the data with the other data.

Load the final process of the ETL. The data is written into the target table or database.

In our scenario, we made use of the ETL procedure in order to migrate the stored data given to us originally into the OLAP 3D cube we designed ourselves.

After the ETL implementation and execution was complete, the the data was migrated from the original OLTP and inserted into the OLAP cube created by us. In order to confirm that

BI Assignment Scri...RDA-PC\keith (52))* | SQLQuery1.sql - KC...RDA-PC\keith (53))

```sql
-- Preliminary Analysis
SELECT      productId
            , orderId
            , COUNT(*) AS 'Count'
FROM        [oltp].[orderItem]
GROUP BY    productId, orderId
```

100 %

Results | Messages

	productId	orderId	Count
1	F9FB11EC-D16D-411F-8879-16AA50450ED8	63013224-35E4-4915-9856-127B20B9BD71	2
2	063618B1-92CB-468E-92BA-6E569328B6F6	2AC02AAA-7197-41DF-B078-B906797E3662	2
3	85075FAF-3B7D-4517-B709-9E01EC5B76F3	29AA82F8-2E19-44E3-B135-09A3779DB9DC	2
4	7A696B78-0327-4259-B073-694401D6D2AE	CE5E42E6-9526-4AA2-8ED1-701B87234C32	2
5	1A6987F8-827B-478F-9E6F-C6590C2F73FD	F92DD804-B5ED-4A64-8775-72A36CAF7042	2
6	12A31A12-1589-46D2-B5D0-D14DD220F1F9	2840FA53-6BF6-4CF6-B356-74C978304F28	2
7	0372DEB8-1349-41BD-9B4E-4AEB2DFC647D	D660FD49-E115-4882-A5F8-2580C10AF03D	2
8	DFCC8ECB-3232-4795-B1D3-83039F81CCCA	A274D3F9-D19C-4C88-B3F5-E23816771C98	2
9	6C2C66EF-3767-47A5-8491-731854F46F3D	C913A722-7B59-43C1-B285-4C8E562A64A2	2
10	AFE593B6-C34C-4883-BDB3-74BD0ADB459C	36EBCFCE-A151-4541-9482-3D09CE73C50D	2
11	F134C2C9-40A9-4E65-BE4A-FD816B294011	F76137AD-C4BA-44B2-B2BE-30607A6BC2...	2
12	5DE810A8-1C47-4709-9648-EE2C02C46C6B	D960BBF0-E821-44BF-8F4D-4E1628029DD6	2
13	F67AE0B4-F347-4AF2-AD08-5F189D58AD66	E70FB22A-2F04-49BD-8918-27DE0FD95311	2
14	C853E313-735C-49DA-B130-5469409AF48F	89CF3614-E9E4-43E7-9AB7-6A0555158EB1	2
15	A801FC9F-33AC-4F48-B664-BDD15B62B835	3595F548-66DC-446E-A32D-1527141CB810	2
16	AF0A20C2-6BCB-4B0E-B8B9-868E0F2BB98B	60AD0C80-379F-47FC-9FD6-272DFA35E16F	2
17	A0EB6C66-88EE-4063-ABDC-3758B8AE15C0	F3A0A4BB-C5AF-4E4D-A473-1A29B374A8...	2
18	B738816F-A8BA-4133-B86C-7A163FE53958	261D2BFE-CE2D-457E-892A-F27C46CCBE...	2
19	7CF81669-6371-4257-BC9D-48E0CEEA893D	2AC02AAA-7197-41DF-B078-B906797E3662	2
20	C4D7A503-D101-4345-8705-D18BC34DF8CC	2DBC3FA9-05D4-4160-9E08-8D515F787993	2
21	3942546C-46ED-456E-BB9A-DDA815232FDA	86D4A407-89C0-4B1D-8FD8-6AD4B3C600...	2
22	FCE56342-A499-4423-A5CC-8B5A895FE8BD	08F7E18D-99C4-4B5A-8413-E995A6D6C14A	2
23	3742F6F0-104C-4A50-AADC-4D7FDCCB15EE	93DC94C3-7472-4ACB-B778-9A1F63B9CA59	2
24	E1D88438-1681-43F9-9289-C4C9FBD5F4BC	E3E5B8E1-5440-44A8-8E00-BD12AE8849A6	2
25	31A9F722-DCBC-4BDE-830E-6DA3573E6A6C	0641C6A7-1D3E-44D1-AB01-23AAF96AFD...	2
26	42948296-C571-46DF-B00E-C091F082C5FF	BE2B60DF-1EDE-481F-A7A4-64429337DB...	2
27	96B1EECE-B86B-4AA7-9E0F-08DBD293CE6C	4A84B719-6DD9-4445-AB79-CA04C7352A25	2
28	6799F289-AFE9-4DCB-AAE8-13E2746358D7	D15DD6E2-077C-4ED5-ACDE-AFD04BCE0...	2
29	E9388E1D-991B-474F-B540-AFF08133DCE5	296CE7A2-4643-4515-BD86-C964EAF9D95F	2
30	AF7F9487-FEBA-4026-8D81-08A9D63DC61E	76449F5B-26BA-4D45-B99F-499C8ED3C5FA	2
31	03DBB068-C68E-4F25-AC13-3725A6F60DBE	665595BE-1435-4AE3-A6D0-8E91AC329EA7	2
32	B23174F7-1CAB-4AD3-8D1C-2AAF3DA589A1	F10E38BB-4934-4058-998D-859F30146652	2
33	AC395A39-56B2-4EFC-B20A-07C60C6F4EE7	15D1F4B1-19D4-4325-AD0C-1A9520627302	2
34	B7742D02-30C9-404D-AC11-93320DA15179	E8FE1D8E-AC48-4769-906F-E79F4ACB0D8F	2
35	6CB0BD9E-9CF2-4575-B4ED-9CDF243A724C	5761C54F-4038-43E4-871D-B10CA16F3A22	2
36	4B285086-CDD6-45B0-81CB-3BA4399602D2	FC47CF0C-F9FF-42D4-8B68-05673CA3491D	2

Fig. 7. Duplicates

```sql
-- CTE To Delete the duplicates found in the previous statement
WITH    CTE
AS      (SELECT productId, orderId, ROW_NUMBER() OVER (PARTITION BY orderId ORDER BY ord
        FROM    [oltp].[orderItem])
DELETE  FROM CTE
WHERE   rowNumber > 1

-- Drop Tables
DROP TABLE [ProductFact].[Product];
DROP TABLE [ProductFact].[Client];
DROP TABLE [ProductFact].[Store];
DROP TABLE [ProductFact].[Time];
DROP TABLE [ProductFact].[OrderFact];
GO
```

%

Messages

```
(196795 row(s) affected)
```

Fig. 8. Common Table Expression

the data was stored successfully inside the new OLAP table wihtin the database, we executed simple select statements to display the data that had just been stored inside the database.

IV. Conclusion

Thanks to this interesting project, us students / researchers have learnt alot about how the stored data may be extracted, analysed, transformed, migrated and stored inside a custom OLAP 3D cube by making use of ETLs. We were well equipped with study material in order to be able to work on

Fig. 9. ETL

this project and this may motivate us to keep on extending this study further.

Fig. 10. MCAST Logo

REFERENCES

[1] "What is the crisp-dm methodology," 2017. [Online]. Available: http://www.sv-europe.com/crisp-dm-methodology/dataunderstanding

[2] "Cross industry standard process for data mining," 2017. [Online]. Available: https://en.wikipedia.org/wiki/Cross$_Industry_Standard_Process_for_Data_Mining$