# OLAP Data Analysis and Cleaning

Kurt Montanaro
Institute of Information & Communication Technology
University College
MCAST, Paola PLA 9032
kurt.montanaro.a100362@mcast.edu.mt

***Abstract*—Nowadays, data plays a really important role in one's daily activities. This is continuously being acquired through various research methods such as empirical research, better known as direct observation by the acquirer, or the most common method used nowadays known as data mining. This data is then stored as raw data in large data warehouses to be used for multiple purposes. With the aid of Business Intelligence, this data could be further on monitored, analyzed , predicted and cleaned. This reduces the amount of redundant and/or duplicate data, retaining the high quality needed to carry out analysis and allot data to specific sectors. With the use of some Business Intelligence techniques, this project managed to optimize and analyze data on a given data set to be provide certain feasible forecasts for the market in question.**

***Index Terms*—OLAP - Online Analytical Processing; OLTP - Online Transaction Processing; SQL - Structured Query Language; ETL - Extract,Transform,Load; ERD - Entity relationship diagram; CRISP-DM - Cross Industry Standard Process for Data Mining SCD - Slowly Changing Dimension**

## I. Introduction

Database applications help gathering and storing data for their specific applications, but since it's a large bundle of data which is not yet optimized, it will not be feasible to use the traditional relational database. Although these are commonly used in a database industry, they still do not prove feasible for data warehouse applications, as they cannot handle such complex structures. This need was met by means of the Online Analytical Processing (OLAP) concept, which is a virtual cube of data, that accomplishes good speed of processing and analyzing large data sets. Modern day Business intelligence is dependent on such OLAP cubes.

The main aim of this project was to analyze , clean and make use of a large data set containing information on sales done across stores in various locations , dispersed in years 1997 and 1998, by means of OLAP operations. The Cube was first to be implemented as to represent a multidimensional Entity relationship diagram (ERD), better referred to as an Online Transaction Processing (OLTP) cube, structured from the given data set. By simplifying each dimension to a representing table, all connected to the Fact table, an OLAP cube was created, designed to cater for data to be used for the upcoming processes. The migration of data from the OLTP to the OLAP cube was processed by means of an Extract, Transform, Load (ETL) process implemented on a database management tool, that address each table in the OLAP and places the values in the affiliated attributes, with the rules , order and constraints. Prior to this process, Data cleaning had to be carried out as to respect implemented constraints and remove any redundant data from the given OLTP. After all these processes were done, the data mining phase started, where excerpts of data was taken out from the fresh OLAP cube tables in order to report and analyze patterns in data which give out interesting outcomes such as forecast information or determining decision trees. The research served as a great introduction to large data handling and analysis, a sector which interests many, of which outcome is profitable to those who make use of it responsibly.

## II. Literature Review

This section carries out a survey on multiple techniques used to analyze the data extracted from the given OLTP, that have been found useful during the data analysis process as documented in Section III as well as various other ways of how data be analyzed in further detail. Market basket analysis is a common measure being used in the documents being surveyed. Key points are pointed out, helping the process of the applied research in the process.

In paper reference [1] , Data Mining is given importance in the fact it should more interpreted as Knowledge Mining, which defines the knowledge of which parts to mine and their future applications are to be taken into account before mining. Knowledge, in the words of Raorane, is a process that consists of the coming steps: i) Data Cleaning, removing redundant data and noise; ii) Data Integration, referring to combining multiple data sources where possible; iii) Data selection, the process of choosing the data relevant to the process; iv) Data transformation, where data is transformed into forms for further mining by summary or aggregation; v) Data mining, intelligent processes deployed to extract data patterns; vi) Pattern evolution, referring to knowledge of interesting patterns; and vii) Knowledge presentation, the process of presenting the output mined knowledge to an average user.

Giering [2] uses Customer Demographics, at Store Level, in order to gather retail sales prediction and Item Recommendations. This is done by sectioning his research in various sectors which build up the final report expected. First off Giering studies data and goals, by acquiring sales figures, breaking them over a set of N non-overlapping customer types. That being said, another study has to be conducted as to identify every possible customer type. A statistical overview of the data is carried out as to break down the large data set into smaller models or subsets which are well equipped with data for further use. Further on this goes through mathematical

formulation as to gather estimates of values that was primarily missing from the original data set. after plotting the output results against the previously designed customer models one should already get a quick idea of such data aiding sales prediction at store level.

Liang [3]states that market basket analysis is a useful method when it comes to exploring data on customer purchases patterns, when it comes to sampling data from store's databases. It is also stated that such data can be bought for reasonable cost, but unfortunately not all corporations involved know how to extract such complex data for analysis. Liang stresses that thanks to market basket analysis this is possible, and will lead to profitable outcomes once fully analysed accordingly. Along the process, it is pointed out that Association-rule mining is a useful method of exploring client purchase patterns by evaluating associations form a store's transnational database. On the other side it does not prove feasible towards multi-store environment, as then it would generalise patterns throughout all stores, which is a bad practice. In order to cater for multi-store environment, a store-chain method is used as to evaluate purchase patterns even in mixed purchasing strategies involved.

## III. Research Methodology

The Data Mining project was implemented based on the CRISP-DM methodology. CRISP-DM defines a structured plan of events that have to take place before starting off with the actual Mining and analyses project.

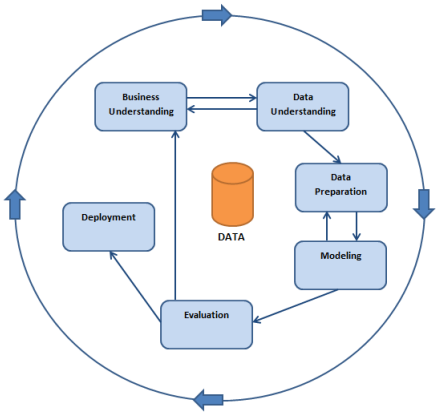



Fig. 1. CRISP-DM diagram showing all the stages covered in the process.

Primarily I had to plan a set objectives both from a client's and my perspective, produce a plan for the main goals and the data mining itself and set up and assess whether the project has been successful towards a business perspective. This is also known as the Business Understanding Phase in the CRISP-DM methodology. The following stage allows me to access the initial data and familiarize myself with it, as to identify quality problems, which are of an issue when it comes to large amounts of data, set an insight to the data and detect on which sub set I can form a hypothesis on for any possible hidden outcomes. This stage is known as the Data understanding stage.The given data structure is multidimensional, which meant gathering data into groups in order to get a first hand simplified view of the OLTP setup.

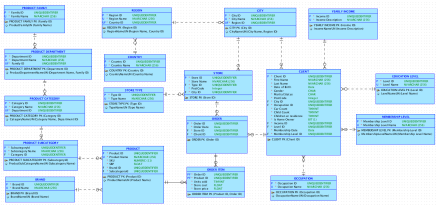


Fig. 2. Data understanding of the given OLTP ERD.

After understanding the data, the Data Preparation stage is where all activities regarding data migration to the final data set is planned out. In my case I had to revert to the previous stage multiple times as to get a good outcome of this stage, which as a final product I had set up segments of the data as tables, attributes and keys along with required restrictions. Prior to the model design itself, the data was cleaned from any redundant or duplicate data which would interfere the process later on.With the use of a CTE, a temporary table was created as to partition data with a composite key which is present more than once in the OLTP data, and remove these in the process. By doing this, I have eliminated redundancy of data which would interfere with the deployment phase. As shown in Figure 3, in comparison to Figure 2, the view of the given data set is much more simplified and readable, portraying the Order Item as the fact table in the OLAP cube designed, along with the Client, Product and Store dimensions. In difference to the OLTP, the OLAP would include a Time Dimension in order to properly portray time in analysis. In this scenario one might notice that the Slowly Changing Dimension in the OLAP design is the Client dimension, by portraying From Date and To Date attributes as to keep track of the effictive date in the data.

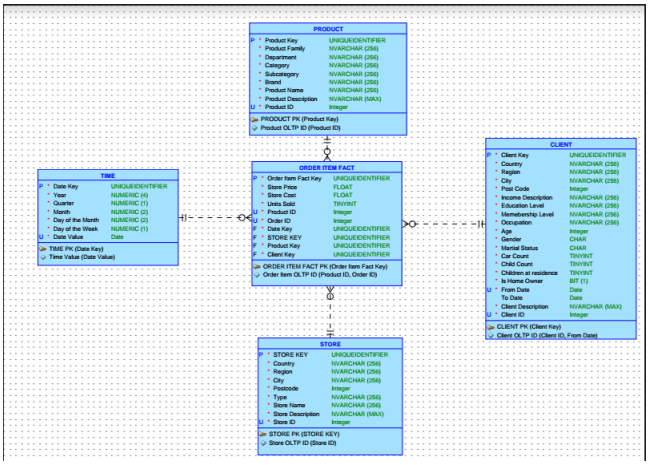



Fig. 3. An OLAP ERD setup for the fore coming OLAP operations.

As previously referred to, the Modelling phase has been

carried out using the OLAP cube approach, both in code and design, continuously referring to the Data Preparation stage to get a good quality outcome. This phase consisted of getting a raw image of the simplified data in previous stages and give it an application for analysis and data mining. Figure 4 shows a 3D OLAP Cube of the previous statement.

The following stage is where an an analysis is carried out on the the previously implemented model. This is reviewed by each step of execution required to create it, as to check if all business objectives are met. I met some issues in this stage which led me to refer and amend on the first phase (Business Understanding) until I got a clear view of what the data mining results had to be, hence in CRISP-DM methodology, this phase is referred to the Evaluation phase. An ETL was carried out in order to migrate all data to the implemented OLAP cube, which in the process pointed out all flaws and suggestions before migrating the data to the dimensions. The Fact table, which would be the last of the dimensions to be given data, would not proceed to be implemented unless all flaws from the OLTP data are handled accordingly. First time I had carried out the ETL process, one of the main flaws pointed out was that some items were redundant/duplicate, for which I had to refer to the first stage as to determine which values I wanted to keep, for which I moved on to the Data Preparation phase with which data I had to prefer over, and carry out a data cleaning process. The final stage, Deployment Phase is where the model done is put into practice. Although it is known that the model is provide knowledge on the given data set, it would need to be organized in a way that it caters to a general learning curve, meaning anyone including the client can use it. All stages had to be documented before the final report took place. Upon fully maintaining the project ,the final report and presentation were ready to be outputted. The bulk of the process was covered in the ETL code, for which now it meant that the Data mined from the OLTP, can be used for whatever analyses planned earlier on. This phase provided multiple data outcomes.

## IV. Data Gathered

This section covers all data gathered after all processes done in Section III where one can a see simplified view of data, which was once a large chunk of raw data, now can be used for any reporting necessary. The Order Item Fact dimension was designed to cover Client, product , store and order data, all in relation to each data. I decided to focus the main goal as to identify the whole product dimension against orders done per quarter of each year in question. As the data is extracted from only two years, these being 1997 and 1998, I was only limited to portray a maximum of 2 years, 24 months or in better OLAP terms, represented as 8 quarters. The Product dimension consists of the following valuable data: Product Family, Department, Category , Sub category, Brand and usual Product details, of which each can be segmented against amount of orders, grouped by the order of significance or popularity in the stated period. In order to acquire this data an OLAP cube was designed and implemented with the previously mentioned dimensions as for simplification of approach towards the meaningless data situated in the OLTP. First off the data had to go through cleansing of redundancy as to avoid unwanted factors later on in the deployment process. After the required tables and procedures were coded and implemented, an ETL has been implemented as to migrate data in the newly structured tables. The procedure in question handled client registration by means of a cursor. It can be a common mistake that if the ETL is not structured properly, the data may either not be migrated or transferred incorrectly, therefore it was made sure that each table was executed in its order, leaving the fact table the last one as it contains constraints and foreign keys. Once the ETL proved successful and all OLAP dimensions were populated accordingly, reports were carried out on the dimensions by means of OLAP Operations which are Slice, Dice, Drill-down and Roll-up reports.With these operations each product detail was analyzed against order items, calculating each attribute ratio. By means of pivot I managed to portray a forecast of product families and the amount in orders they are included in, per quarter of the year.

| Data Gathered | Description |
|---|---|
| Data | Data Cleaning and Data migration |
| Cube | Design an OLAP cube as to simply Data gathering |
| Analysis | Analysis of data after previous stage |
| Reports | Carry out reports with given data |

TABLE I
INITIAL DATA GATHERING PLAN

## V. Data Analysis

After the Data Gathering stage was successfully executed, I have managed to output fruitful reports with data that is readable and adequate for formulations in order to determine any research outcomes. As stated in Section IV, my reports, as illustrated in Figure.4, are about all dimensions in product against number of orders, in a specific locations, per specific quarter of the year.

Primarily I gathered the total amount of data in each attribute in the Product dimension as to determine the upcoming result on them, in terms of ratio values. This proved useful as to determine the popularity of the data in question according to its application. Each attribute was then reported the amount of data linked to the amount of orders they are linked with. First attribute to be reported was the product family against amount of orders, showing a significant gap between the Family 'Food' with the second in amount, being 'Non Consumable', having a gap of 49% difference in ratio. In further detail this was also reported by means of a pivot where both years in question were segmented in quarters by means of the time dimension, showing amount of orders of each family per period stated. With 'Food' being the highest and 'Drink' being the lowest, it was still fascinating enough to see that the 'Food' family had a much more steep down fall towards the last quarter that that of the drinks sector. In terms of forecast all families were predicted for a downfall in the upcoming two quarters.This is clearly illustrated in one of the reports as illustrated in Figure

5. All this is based on one country, that being USA. With further reports I have seen that as for most popular orders were for products which were Produce (Department), of Category Vegetables', subcategory 'Fresh Vegetables' and most popular brand being 'Hermanos'.

## VI. Conclusion

In this project we applied a Data Mining approach towards Store Sales scenario. Whilst I had the opportunity to analyze real world data, all the process of acquiring this data for use was done using the CRISP-DM methodology. This has proved a good effect on the outcome of the project, making it easier to understand data of such a large amount. Even though data acquired can be useless nowadays as for its age, one can appreciate that certain patterns of data outcomes still reflect up till today and are useful for multiple purposes. The project achieved clean and readable data, from the OLTP given, with the use of OLAP cube models and operations. OLAP proved efficient as primarily it is perfect for a multidimensional scenario such as the one I was provided with, reducing CPU costs as less Joins and aggregations are required, which when handling large amounts of data, would have a really long query response time which would prove impractical.Reports generated are also a good outcome of this project, showing readable data in action, ready to be applied for good use. In this project I had encountered some shortcomings, which are due incorrect planning on my end. It would have been much more appreciated if the outcomes were used for a decision tree, which shows the data's effectiveness and accuracy. Whilst this was possible with the given data, Decision trees are time consuming, which proved a challenge in multiple cases. Limitations in regards to this project was that data was of a two year time period, which did not allow much space for it be experimented on in terms of Forecasting reports. It was implemented on Transact SQL platform, which requires a learning for some people to execute these processes for their own study use, also limited to certain platforms for it to run on.

I would suggest using CRISP-DM for any future Data Mining projects, modelling data in OLAP cubes for easier understanding of data as these two factors allow access to most data warehouses, even if the data is near impossible to interpret from OLTP. This research will serve as a basis for further projects I will carryout on recent matters which are evolving in Malta.

## Appendix A
## Supporting Material


## Acknowledgement

Biggest thanks to the lecturer for providing adequate aid on our school platforms in order for us to maintain high quality in our deliverables.

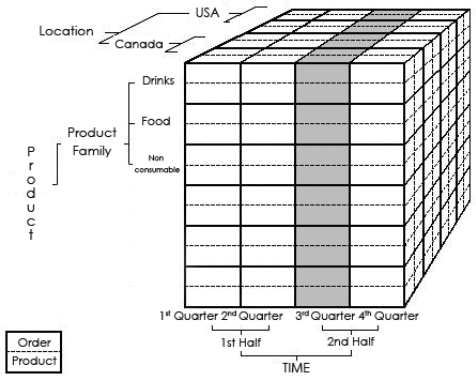
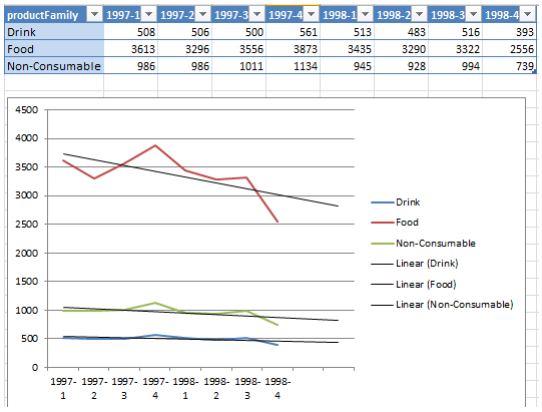




Fig. 4. A 3D Modelled cube of the OLAP analysis.



| productFamily | 1997-1 | 1997-2 | 1997-3 | 1997-4 | 1998-1 | 1998-2 | 1998-3 | 1998-4 |
|---|---|---|---|---|---|---|---|---|
| Drink | 508 | 506 | 500 | 561 | 513 | 483 | 516 | 393 |
| Food | 3613 | 3296 | 3556 | 3873 | 3435 | 3290 | 3322 | 2556 |
| Non-Consumable | 986 | 986 | 1011 | 1134 | 945 | 928 | 994 | 739 |



Fig. 5. Pivot Report done on Product Family in orders per quarter of the year.

## References


[1] Raorane A.A., *LaTeX: Association Rule Extracting Knowledge Using Market Basket Analysis* , Vivekanand College, Kolhapur, 1st edition, 2012.

[2] Michael Giering , *LaTeX:Retail Sales Prediction and Item Recommendations Using Customer Demographics at Store Level* , N/A, 2nd issue, Vol.10, N/A.

[3] Yen-Liang Chen., *LaTeX: Market Basket Analysis in a multiple store environment*, Department of Information Management, National Central Univeristy, Chung Li , 320 Taiwan ROC , 1st edition, 2004.

[4] V.A. (Vincent) Kool., *LaTeX: OLAP Cubes* , Faculty of Science - Computing Science , 1st edition, 2012.

[5] Chapman, Pete, (NCR)., *LaTeX: CRISP -DM 1.0.*, Chicago, IL: SPSS, , 1st edition, 2000.