

Assignment 1

by Martin Pugsley

Submission date: 25-Mar-2022 02:13PM (UTC+0000)

Submission ID: 174564104

File name: 18449_Martin_Pugsley_Assignment_1_398298_1811788438.docx (735.25K)

Word count: 2703

Character count: 13224

Name: Martin Pugsley

Student ID: 1907620

Module Name: Data Warehousing and Data Mining (ACCA6026)

Data Warehousing

Table of Contents

Data Warehousing	1
1. The Data Warehouse	3
1.1 Difference Between OLTP and a Data Warehouse	3
1.2 The Data Warehouse Architecture	4
1.3 Data Marts	7
1.4 Normalised and Dimensional Structures	7
2. The Rags n Riches Star Schema	9
2.1 Internal and External Data Sources	9
2.2 Data Warehouse Star Schema	10
2.3 Appropriate Data Marts	11
2.4 ETL Process	12

1. The Data Warehouse

1.1 Difference Between OLTP and a Data Warehouse

OLTP and Data Warehouses are both popular methods of storing data to gain valuable insight, but both are used differently and according to business requirements.

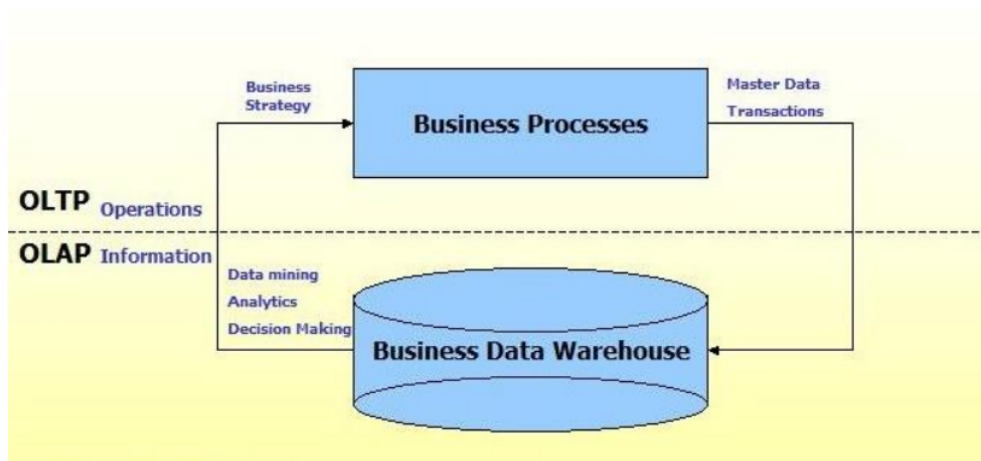
OLTP databases are primarily used when there is a need for real-time data processing to maintain the accuracy of records. For example, when a customer places an order with Rags'n'Riches, clerks will interact with the OLTP database to establish if the item is in stock before completing the order and then edit the items rental status to complete the transaction, ensuring no 'double-booking' of items occurs. Both query and recorded editing happen in real-time, which is a very important operation for ensuring the accuracy of the items rental status, but when it comes to data mining and historical analysis, OLTP's are not the suitable choice, their primary objective is to process data not analyse it, rather a Data Warehouse is considered the better option for analysis tasks.

The role of the Data Warehouse is to consolidate large amounts of data, both current and historic. They are theme-focused, meaning data is more subject orientated (Sales, Marketing, Customers etc) rather than business operational, this results in a greater emphasis on Business-Intelligence, providing a simpler process to analyse data and perform ad-hoc tasks, whilst excluding un-useful data around a certain subject. However, Data Warehouses consist of much more than just a storage facility, it is an entire data management system, which includes:

- A relational database.
- ETL processes to capture and transform the shape of data from various systems, including data-entry forms, into a standard view, which is then uploaded and stored into relevant tables in a timely fashion.
- Tools to create statistical reports, analysis, visualisations and can include more advanced applications to aid in data science applications.

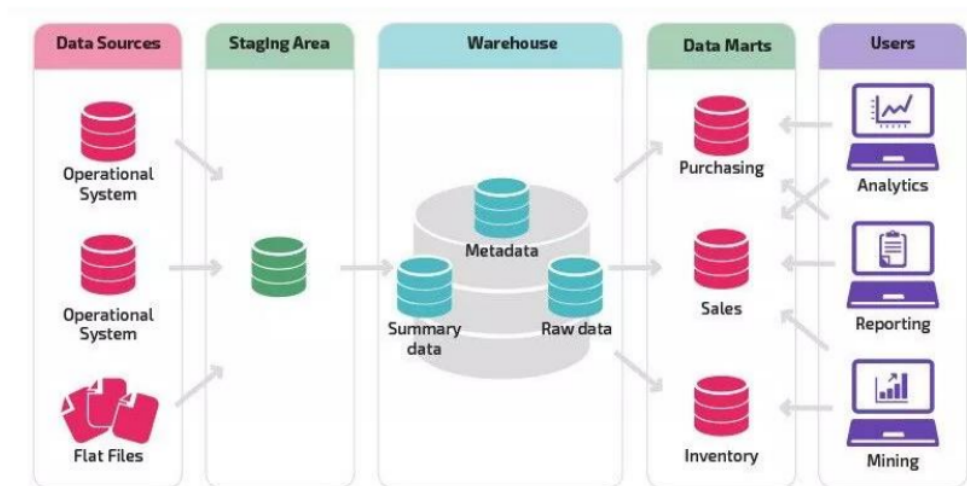
Due to the amount of historic data which can be stored, analysts can quickly discover trends, for example, the seasonality of costume type rentals or most popular rental items over time, so the company can stock up on more items or release items which are unpopular. This is possible because data is non-volatile, meaning data already stored is not removed when new data is loaded in.

The OLTP database is still an important tool and can still be used in operations, but it will work as a 'data source' which feeds the Data Warehouse so analysts can use historic data as part of their research. Another important difference is with the OLTP database, specific requirements on the data which is captured have been created, but when changes are required, it is a complex process to ensure any edits do not disrupt the daily operations, but with Data Warehouses, the process is much simpler as potentially it may be one data source or one storage database that needs to be modified and this can be done behind the scenes, so daily operations can continue.



1.2 The Data Warehouse Architecture

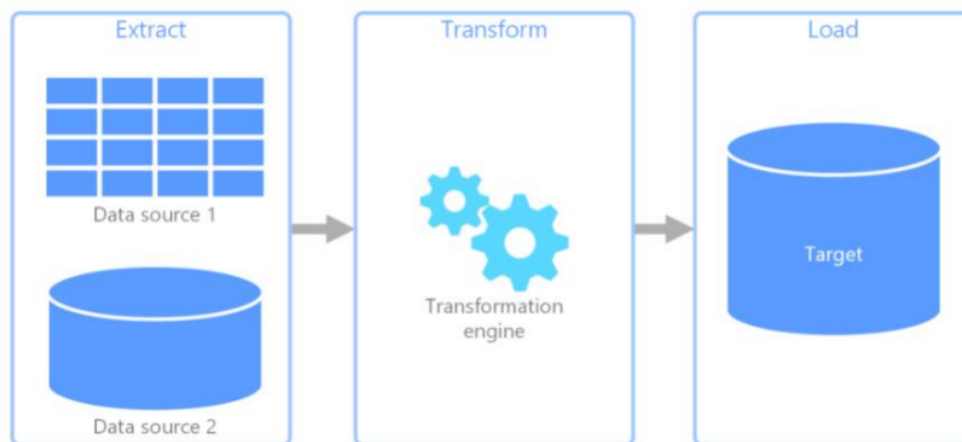
The architecture of a Data Warehouse defines the arrangement of data, which combines information from the various data sources into a unified schema. There are three main types; One, Two and probably the most common type the Three-tier architecture, which produces an organised flow of data, from its raw format to valuable output or insight, which is represented in the below graphic.



The flow of data first starts at the 'Data-Sources' level, which can comprise from various sources such as business data-entry applications to manual flat files like Excel, Access, text files or even emails. The user will use the software packages as part of their daily tasks and save any inputted

data as normal, and as previously mentioned, the Data Warehouse is much more than a storage system, it has an ETL (Extract, Transform, Load) tool, and uses an intermediate storage area, known as the 'Staging Area' to house and process the raw data until the ETL tool has cleansed and organised the captured data into the defined data structure within the databases, located in the Warehouse.

At face value, the ETL process may seem simple, but it can be quite complex, as not only is the raw data from each data source shaped into a suitable format, but it can also consist of filtering, cleansing, removing duplications, authenticating, and validating the data too. It can also carry out calculations at a data level so an analyst wouldn't have to do it themselves, such as calculating VAT on purchases or currency exchange rates, but probably the most important feature of the ETL process is to ensure the quality of data is maintained by checking data formats match the database criteria and to expose anomalies. As an experienced Data Analyst may say is that 80% of Data Analysis tasks consist of collecting and cleansing the required data, but with an ETL process in place a lot of the hard work is done before the data is stored, which will speed up the reporting process.

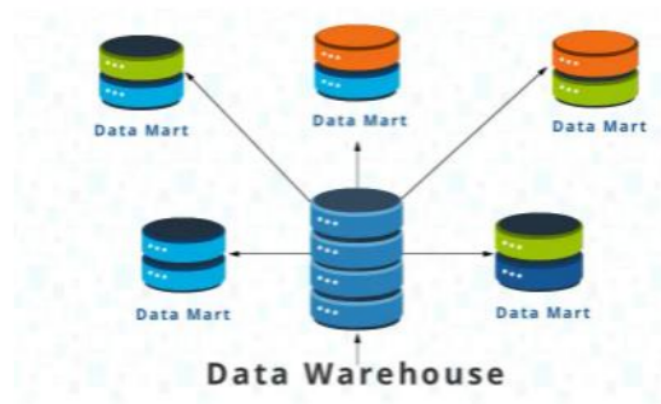


It is only after the successful transformation of the raw data that the information is loaded into the targeted data storage areas and can be used in reporting and analysis. The ETL process can be time-consuming, but it is an automated process which most organisations usually run within business out-of-hours, so it doesn't affect the daily operations of the business or any system performances.

It is at the 'Warehouse' level where all the transformed data is stored but again data can be stored as different views, for example every single piece of information is stored in separate defined tables, which would be used in data mining tasks, but there can also be aggregated or summary tables of the raw data, which are pre-defined by the business to encapsulate the most common answers to things the business need to know quickly or often. These are included because to process a lot of raw data can be very time-consuming, having the summarised tables would mean the data is to hand in a defined view and can be accessed quickly.

Data is not only limited to what has happened within a business transaction, the Data Warehouse will also store the metadata, which in simple terms means 'data about the data', which is a complex area, but to associate it with the Rags'n'Riches business it would include data about who inputted the record, when, where, using what data source etc, which is very important in the field of traceability and GDPR, knowing who has done what will help with maintaining the integrity of the data and knowing what data source was used may help in determining the trust in accuracy of the data, for example more trust may be placed in a number calculated by a system than manually input by a user.

After the 'Warehouse' level there is the 'Data Mart' level, which is part of the storage component. There can be as many Data Marts as required but they will store information of a particular function within the business and will contain subsets of the data stored in the Warehouse. These can be known as 'Views' and are designed for specific users, with a need to see certain data fields. A Data Mart can have fields from different databases, but they eliminate the need for the user to join tables together especially if the task is needed to be performed often.



Finally, the last level of the Data Warehouse architecture is the 'User' level. This is where analysts will use the tools, which may be provided by the Data Warehouse or not, to report insights and findings. Again there are few stages to the 'User' level as this may consist of BAU (Business-as-Usual) reports which can be automated as they are required often, ad-hoc queries which the user will have to select and join the appropriate data fields from the Warehouse or use a Data Mart to discover answers, and data mining tasks, which can be more experimental and can evolve around finding new things the business did not know, with the purpose of applying findings to daily operations to improve the overall business.

1.3 Data Marts

It is understood there are various parties, which need to see specific, relevant data to aid in decision making processes; a Data Mart would be a suitable method for this.

A Data Mart differs from a Warehouse, where the Warehouse consolidates and stores all the historic business data from many areas in one location, a Data Mart will be designed with a certain business function in mind and will join selected columns from multiple tables, within the Warehouse into one table or view. It does this by using the relational keys within the tables and exist to eliminate the need for having to constantly join tables or extract data frequently.

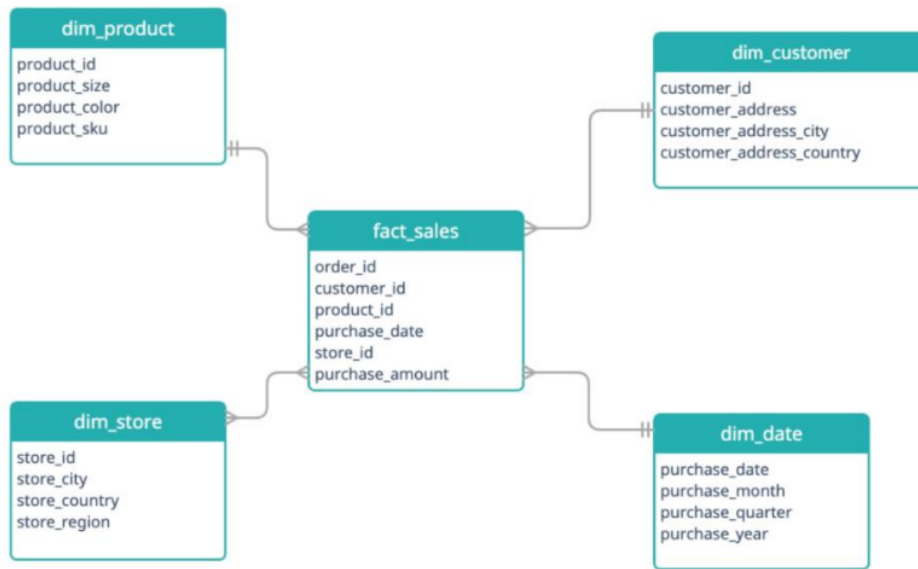
An example of how a Data Mart can be used within the finance section is to join numerical data from various tables, i.e. Sales, Refunds, Operational Costs to calculate overall profits or losses but can also join external sources such as VAT rates to apply to calculations. Also, the case study states the finance team run a report to display which items are proving popular and profitable, so there is a need to include columns such as item types from the Orders information in the same Data Mart. This would then give the finance team the one data view needed to answer their typical questions.

Security is applied to the Data Mart, as it is to the entire Data Warehouse so only authorised users would have access to it and see the view. Data Marts can be quite versatile and to make things easier for the user, they can be made downloadable, so the user would only have to select the start and end dates to retrieve a data extract, or a direct connection can be made via the user's analysis software package. Another option is to make them into 'data-cubes' so the user could create summarised views, just as they would through a normal pivot table in Excel.

1.4 Normalised and Dimensional Structures

The OLTP database which Rags'n'Riches currently use will structure the data in a normalised fashion. Traditionally data is stored in tables, which can be connected by using primary and foreign keys (A unique identifier for a data entity), for example if there are 2 tables, one with Customer details and one with Order details, it would be likely the Order table would have the unique CustomerID as a foreign key as a part of the data entity, so it can be used to traceback who the order belongs to, from the Customer table.

The main advantage of a normalised data structure is there is little data redundancy, data is stored once, there isn't duplication, but with dimensional designs data duplication is quite large but it is purposely designed this way to simplify analysis tasks. In this method, the FACT tables hold the numerical data and dimensional tables hold the various information about the ways we would like to analyse the data. For example, in the below image the fact_sales table holds the information about a particular order and the surrounding tables hold the dimensions which relate to the various id's, linked to an order, so if we wanted to find out how many orders from Swansea were made, it is possible to link back the customer_id from the fact_sales table, to the dim_customer table and find the customers who made orders from Swansea.



By designing data structures using dimensional modelling, we can continuously improve analysis techniques by adding additional dimensional tables when they are required, it is a much easier process than using the OLTP system as it can be done out-of-hours to not disrupt operations and no data is deleted or edited, just added to the new dimensional tables.

2. The Rags n Riches Star Schema

It has been decided the Data Warehouse option will be used within the Rags'n'Riches business going forward. The following chapters will depict the first draft of a star schema around how the architecture can be structured, started with Internal and External Data Sources, and finishing with the Users.

2.1 Internal and External Data Sources

The schema starts off at the Data Source level and consists of both Internal and External sources. Internal relates to data which originates from within the company, whereas external consists of data not owned by the company.

The internal sources start with the **OLTP**, which is used to check if stock is available when an order is placed, ETL processes will move updates of the OLTP to the 'Staging Area' of the Data Warehouse, which will then eventually land in the databases within the Warehouse.

Stock Movement Forms are used frequently, so it may be best to invest, if not already, in creating a be-spoke software package to allow the data capture process to be easy for users.

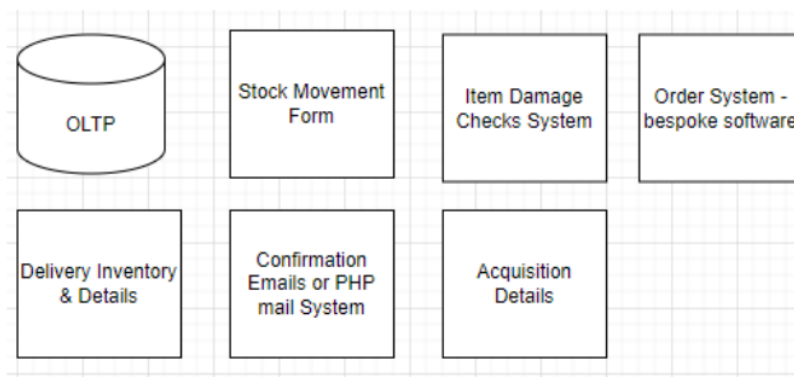
The same applies for **Damage Checking**, this task is carried out every time an item is shipping or returned so a be-spoke software package to help with the capturing would aid the user.

Ordering system would be a primary data source as this will send the relevant order details to each of the tables.

Delivery and Inventory Details would be another internal data source, once all items have been gathered, they are then shipped to the customer.

Email Confirmations to inform customers about orders and invoices.

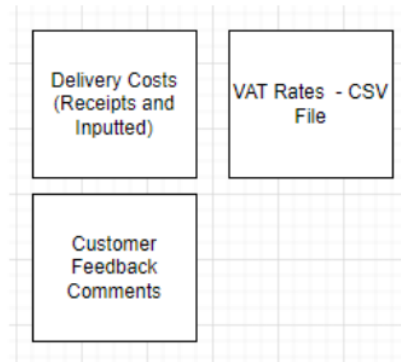
Acquisition forms are used to note new items of interest, so once again a be-spoke software may be the best option for this.



The external data sources start with **Delivery Costs** as the company is conscious of its delivery costs, so receipt values must be captured when using fuel cards and vehicle mileage too, to validate costings.

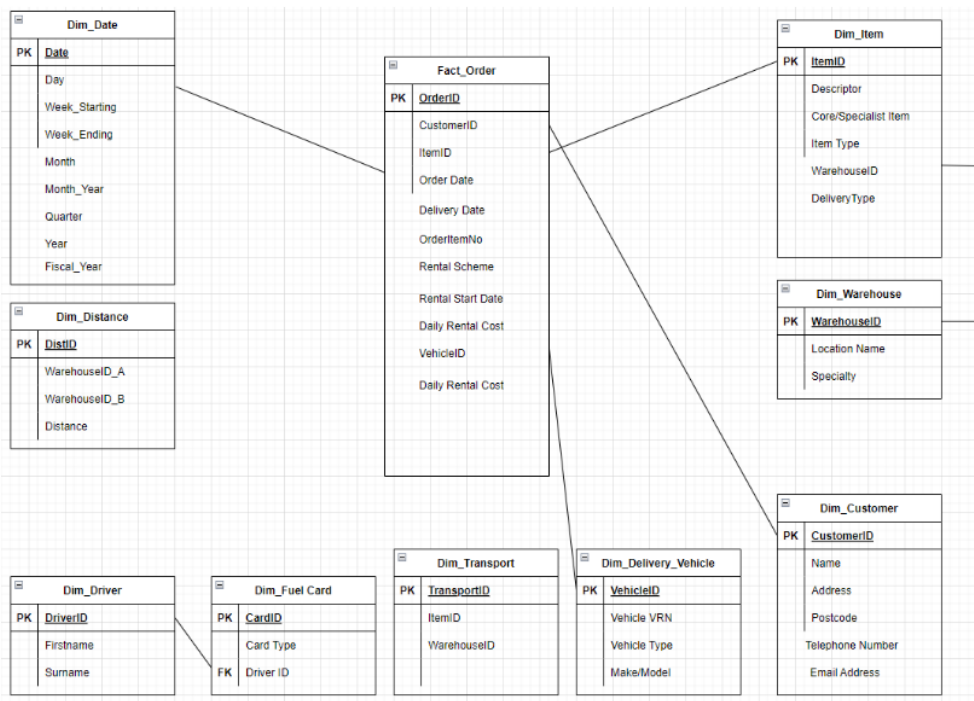
VAT Rates can be captured within a CSV file, no need to do anything fancy with these as they change less frequent than other items. A user can simply change the tax rate and apply a 'Start_From' date so when applying VAT to calculations the correct percentages are used for the desired period being analysed.

Customer Feedback has also been included as an external data source as it is vital for businesses to capture customer experiences to see what went well and what didn't to apply going forward and to also see if the quality of items is fit for purpose.



2.2 Data Warehouse Star Schema

Below is a very basic un-finished star schema of the databases in the warehouse. This is a first draft and can be adapted over time but will give an overview of how dimensional modelling can be used to answer analysis tasks around a typical order.



The Fact_Order would hold all information about the order and the surrounding databases have the lower level of detail around a dimension in the FACT table.

The Date table could contain many metrics around a specific date and can be used in combination with any date field (order date or delivery date etc). There would also be databases containing customer details, warehouse details throughout the country, item and delivery details.

2.3 Appropriate Data Marts

Already discovered in part one, finance, operations have a need for a Data Mart, but the transport, Acquisitions, Communications and Warehouse teams also need Data Marts.

Finance will need one to report anything cost related, communications will need to know daily who they must send reminder emails and invoices to.

The Transport department will need a data mart to, so they know what items to move from warehouse to warehouse and acquisitions will need a data mart to see what new items have been recommended so they can do some research.

Each Warehouse may also need a Data mart to prepare items for shipment to customers or other warehouses.

2.4 ETL Process

Assignment 1

GRADEMARK REPORT

FINAL GRADE

34/100

GENERAL COMMENTS

Instructor

martin pugsley		
1 engagement with literature		
comment		available
no evidence of engaement with literature	20	
2 Knowledge and Understanding		
comment		available
knowledge demonstrated and appropriate images used to reenforce the discussion - but no references used or credit given to image sources.... At level 6 this is considered the bare minimum		
3. Cognitive and Intellectual skills		
comment		available
discussion of internal and external data sources present but somewhat misinterperated- think more about sources relevent to to the organisation than simple facts and figures (e.g. known trends, local events) etc	10	
4. Practical Application Skills		
comment		available
2.4 is a 'dead-heading' I assume the candidate ran out of time		
minimul design and discussion included on design for the datawarehousing (elements that were covered are legitimate but very basic)	40	
5. Transferable Skills for life and Professional pract		
comment		available
layout sound - no major SPaG errors, however, major sections missing.	10	
Neutral language used to allow the reader to understand the content		
Total:	34	

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12
