10/23/2017

Collins, Josh

Liverpool John Moores University

Coursework One

6048COMP Innovations in Software Development

Weighting: 30%



Contents

[Figures 2](#_Toc498020681)

[Abstract 3](#_Toc498020682)

[1. Introduction 4](#_Toc498020683)

[1.1 History of Databases 4](#_Toc498020684)

[1.2 Business Intelligence 5](#_Toc498020685)

[1.3 Data Warehousing and Data Marts 5](#_Toc498020686)

[1.4 OLTP 6](#_Toc498020687)

[1.5 OLAP 6](#_Toc498020688)

[2. ETL 8](#_Toc498020689)

[2.1 Introduction 8](#_Toc498020690)

[2.2 Extraction 8](#_Toc498020691)

[2.2.1 Overview 8](#_Toc498020692)

[2.2.2 Methods of Extraction 8](#_Toc498020693)

[2.2.3 Online and Offline Extraction 9](#_Toc498020694)

[2.2.4 Data Extraction Sources 9](#_Toc498020695)

[2.3 Transform 10](#_Toc498020696)

[2.3.1 Overview 10](#_Toc498020697)

[2.3.2 Data Cleansing 10](#_Toc498020698)

[2.3.3 Correction and Normalization 10](#_Toc498020699)

[2.3.4 Business Rules 11](#_Toc498020700)

[2.3.5 Conclusion 11](#_Toc498020701)

[2.4 Loading 12](#_Toc498020702)

[2.4.1 Overview 12](#_Toc498020703)

[2.4.2 Star Schema and Snowflake Schema 12](#_Toc498020704)

[2.4.3 Fact and Dimension Tables 12](#_Toc498020705)

[2.4.4 Aggregations 12](#_Toc498020706)

[3. Case Study 14](#_Toc498020707)

[Conclusion 16](#_Toc498020708)

[References 17](#_Toc498020709)

# Figures

[Figure 1 Relational Database example 3](#_Toc498017099)

[Figure 2 Visual representation of the cube and how it can be dissected 6](file:///C:\Users\cmpjcol1\Downloads\726681CW1.docx#_Toc498017100)

[Figure 3 Four different methods of data extraction (Rainardi, 2008) 7](#_Toc498017101)

[Figure 4 Structured .txt file example 8](#_Toc498017102)

[Figure 5 Semi-structured xml file example 8](#_Toc498017103)

[Figure 6 Process of normalization, standardization, and correction in the transformation process (Golfarelli & Rizzi, 2009) 10](file:///C:\Users\cmpjcol1\Downloads\726681CW1.docx#_Toc498017104)

[Figure 7 Process of how aggregations may be used in a business environment (Golfarelli & Rizzi, 2009) 12](#_Toc498017105)

[Figure 8 Example of data flow using SSIS tools in Microsoft SQL Server Integration Services 13](#_Toc498017106)

[Figure 9 Fuzzy Grouping & Fuzzy Lookup in SSIS tools 13](file:///C:\Users\cmpjcol1\Downloads\726681CW1.docx#_Toc498017107)

[Figure 10 Multicast in SSIS tools 13](file:///C:\Users\cmpjcol1\Downloads\726681CW1.docx#_Toc498017108)

[Figure 11 Lookup in SSIS tools 14](file:///C:\Users\cmpjcol1\Downloads\726681CW1.docx#_Toc498017109)

[Figure 12 Merge & Merge Join in SSIS 14](file:///C:\Users\cmpjcol1\Downloads\726681CW1.docx#_Toc498017110)

[Figure 13 Pivot & Unpivot in SSIS tools 14](file:///C:\Users\cmpjcol1\Downloads\726681CW1.docx#_Toc498017111)

[Figure 14 Representation of an Excel PivotTable 14](#_Toc498017112)

# Abstract

In this research report, I will be investigating the uses of OLTP, OLAP, and an in-depth analysis of the ETL process including uses and purposes. In addition, I will be following up with a case study using Microsoft SQL Server Integration Services (SSIS) with examples.

## 1. Introduction

## 1.1 History of Databases

In the days before databases had been developed, organizations used file based systems with huge indexes to store and find information. This still has a few benefits of using in terms of security, one lockable filing cabinet in a secure office room of an organization. From the Victorian ages up to the release of electronic databases, punch cards were used to hold information by putting holes through as a simple true or false.

Databases have been continuously developing every decade ever since the introduction of navigational DBMS in the mid-1960s. Quickly, databases were soon to be used in many commercial environments for business as they became rather cost effective. During the 1960s, the two most popular data models were CODASYL and IMS. IMS was IBMs own version of a DBMS which was originally developed for the Apollo program, but had been taken and developed further to become IMS.

In the early 1970s, Edgar Codd of IBM produced the relational DBMS because he was unhappy with the navigational model of CODASYL. The relational DBMS brought the idea of using normalized tables and linking them together on certain columns, primary key and foreign key. In addition, this allowed relationships between tables that include: one to one, one to many, and many to many relationships. This idea of relationships in database was a huge benefit for businesses as the performance of databases were essentially optimized.

For example, if a plain database were to contain invoice information, it could be normalized and split into several tables. Only relevant information would be contained in each table as instead of all information in one table. This was also a security issue if any sensitive information were to be contained with other information. Figure 1 below shows a simple entity relationship between a customer, item and invoice table that. Item and customer have a relationship with the invoice so only relevant information is shown in the invoice, also these both use the one to many relationships. This clearly identifies that that one customer can have many invoices, and one item can have many invoices as there is many of the same item in stock.

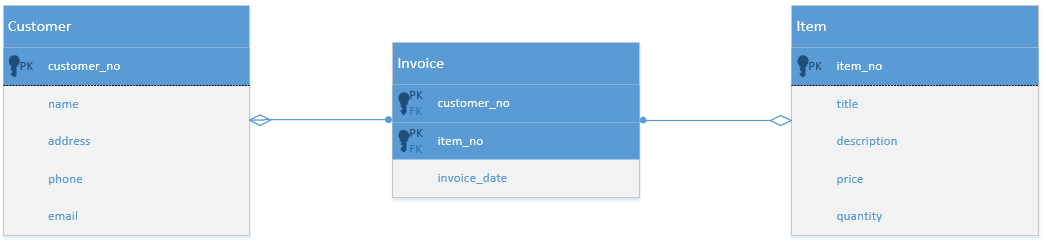


Figure Relational Database example

In the late 1970s, Edgar Codd’s concepts influenced IBM to produce a prototype which later became SQL. SQL now one of the most popular programming languages today. Oracle V1, Oracle V2 and PostgreSQL were also later developed following the release of SQL. PostgreSQL is still used in the industry today and used by a company Jaguar Land Rover outsource production line web systems to, Infinity Technology.

The 1980s era saw the rapid growth of SQL, Structured Query Language which became the industry standard query language. Many different database systems now heavily used SQL for their database systems for the complexity and functionality benefits it gave companies with their data.

Object Oriented Database Management Systems, OODBMS, prototypes were released in the 1990s which introduced several benefits such as enhanced modelling capabilities, extensibility, they were more powerful than relational database management systems, and more. The idea of OODBMS was to treat data as objects in a more graphical format, more user friendly in some ways.

After the millennium, database applications had a decline in growth in the industry, but continued to grow. XML, JSON, NoSQL and NewSQL come about in the 2000s which all become popular and used by organizations. MongoDB is a popular open source NoSQL system; NoSQL systems are often fast operating databases. Citius, VoltDB and Google F1/Spanner are all examples of NewSQL databases. Although new database systems have been releases, SQL is still the standard query language many years later.

## 1.2 Business Intelligence

The key aim for any business intelligence activity is to improve business performance. This may include analysing customer profits, evaluating sales records across different categories or regions, and optimising supplier performance. Microsoft SQL Server and other related software such as Oracle SQL Developer offer business intelligence applications. These applications include analysis services, data mining, alerts, and reporting. This is beneficial for an organization by optimizing the business performance by finding sales patterns and relationships when using data mining applications. A key use for business intelligence is extracting the facts from a database and putting them into a separate table to turn the facts into useful business information. To do this the database administrator must run ad hoc queries that may include INNER JOIN and OUTER JOIN for others to analyse the findings from the data.

Using relational databases for business intelligence can improve the decision making in a company by analysing the information found. For example, gathering staff information on time management over a quarter or year, this would show which staff require disciplinary action and others that may deserve praise.

Business intelligence involves using large amounts of data to track trends, make customer suggestions, show order history etc. This is visible when going onto your Amazon account and you are still able to see orders you placed five years ago. On a larger scale for all customers, to find specific information you must run queries using many rows, this can slow a system. However, few users have access to business intelligence which would be beneficial.

## 1.3 Data Warehousing and Data Marts

Data warehousing is an important concept of business intelligence in databases and is often used for reporting and data analysis. The purpose of using this is to store all corporate data in a database system, large organizations such as Google, Yahoo, Amazon, and more all make use of this. There are many benefits data warehousing brings to business such as it improves data quality by using data assembly, data cleaning, and quality assurance.

Data warehouses may be made up of different architectures, the two-layer architecture and three-layer architecture. The two-layer architecture is made up of two layers but consists of four data flow stages. The source layer consists of the operational and external data, the data warehouse layer contains all meta-data and data marts. The three-layer architecture is similar to the two-layer, however there is more detail in between the source and data warehouse later, with a reconciled layer being introduced. In addition, the three-layer architecture is consists of six data flow stages.

Using a large data warehouse, executing queries can be extremely timely. For example, for a company such as Toyota, if the director wished to execute a query from the top end of the database as he has all privileges. Running a query at this level could take hours and days because the amount of data. This data would be likely to be in Terabytes and Petabytes.

Data marts come into place to allow low level users to access a subset of the database, low level users could be store managers or business manages for example. This is beneficial because it adds security so that low level users are unable to access any unauthorized information that an employee further up in the hierarchy would have access to. Other benefits to using data marts is narrow application area, easier data maintenance, and cost effective with less development time and low costs. In contrast, for a standard data ware house could take two years to develop and costs £2 million. In comparison, a data mart would cost around £200,000 and take around five months to develop. Data marts are suited to smaller organisations with the time scale and cost management.

## 1.4 OLTP

OLTP (Online Transaction Processing) is the process of creating information for a database table, using standard functions that include INSERT, UPDATE, and DELETE. The OLTP stage will also use CREATE TABLE, CREATE TRIGGER, and CREATE PROCEDURE. OLTP is the source of data which is extracted to the ETL stage of the ETL process which is finally loaded to OLAP. OLTP is much simpler in comparison to OLAP because it is simple gathering the operational data and is the first stage of the ETL process. OLTP is typically a fast process because it uses simple SQL statements to create and update tables, and insert data. It also uses information gathered from JSON and XML for cross platform compatibility to process into a database. However, depending on the size of the databases, this process could be much slower for an organisation with large groups of data to process.

## 1.5 OLAP

OLAP (Online Analytical Processing) is the final process of the ETL process that handles historical data, uses analytical services, and uses complex queries that include aggregations. The benefit of using OLAP is to aid organisations with planning, problem solving and improving business decision making. The data used in OLAP gets normalised in the transform stage of ETL. OLAP can be represented in a few methods, the cube being a popular method to store information. In SQL Server, the cube is known as analysis services. OLAP can be used in three different architectures Relational OLAP (ROLAP), Multi-Dimensional OLAP (MOLAP), and Hybrid OLAP (HOLAP).

The cube is a multi-dimensional representation of a database and can link information cleverly on each side of a cube. In the figure below, you can see that the cube makes use of height, width, and depth to store information on each side of the cube. Each block in the cube contains data, it is also possible to create a cube inside a cube, and another cube inside that depending on the complexity of the data. For example, the information containing days, this could include additional information inside that breaks the information down into hours. Understandably working with this on a large scale can be timely trying to analyse information.

Figure 2 Visual representation of the cube and how it can be dissected

Product

Date

Customer

Name

Email

Address

In conclusion, OLTP and OLAP are two completely different concepts but used in the ETL process. OLTP may use the SQL Server Database Engine and it primarily focuses the inserting and updating of data, this is typically in 3rd Normal Form before it is then extracted to ETL. The ETL stage normalizes the data after extraction and then loads it. After this stage the data is moved into OLAP, which uses multi-dimensional models such as the cube to represent the data. The user is then able to perform analytics such as SQL Server Analysis Services.

# 2. ETL

## 2.1 Introduction

ETL is a vital process used for constructing a data warehouse. The function of the ETL process is to extract, transform, and load data into the data warehouse and once complete provides benefits to businesses to perform analysis using OLAP on clean data that may have been normalized. The ETL process takes place when the data warehouse is first being developed, and then continuously in the future whenever the data has changed.

## 2.2 Extraction

### 2.2.1 Overview

Extraction is the first stage that is used in the ETL process, Extract, Transform, and Load. There are a few fundamentals that need to be considered when extracting data to be used in a data warehouse. There is a large amount of data being extracted, which is where OLTP comes into play to ensure data is fetched in small amounts to ensure the system is not overloading with data. This allows the data extraction to perform at a quick and consistent speed to allow the data to be transferred into the data warehouse sooner.

### 2.2.2 Methods of Extraction

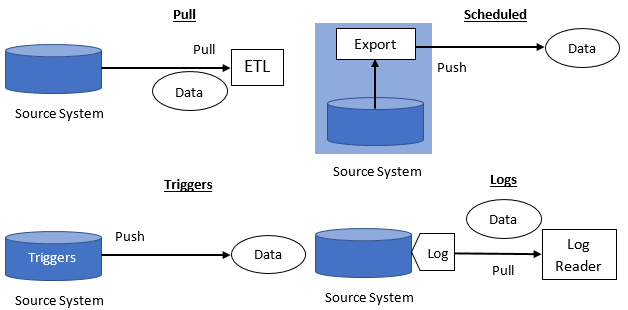
There are four different methods of which Extraction can be categorized into. The most common method is pulling the data out by using queries in the source system database, this requires regular work to continuously extract data. In addition, another method is using triggers, a key benefit of using triggers is being able to store updated rows in a separate table if necessary. This is done by using trigger statements to execute whenever data is updated, inserted or deleted in a table. Scheduled processes are a similar approach to using triggers. As opposed to having a trigger statement to extract data, this is done by the DBA giving a set time to extract data, this is typically best executed at night during low traffic usage in a network. Also, another method is using a log reader. This uses a database log file that will record any modifications to any database tables and stores the output if any changes have been made. These methods are shown in the figure below. 

Figure Four different methods of data extraction (Rainardi, 2008)

### 2.2.3 Online and Offline Extraction

There are two methods of performing the data extraction in ETL, these are online and offline extraction. With online extraction, the data is directly extraction from the source files to then be transformed. Offline does not extract data directly from the source files, instead the data is staged outside the source system. Often the offline extraction method may make use of flat files, dump files and achieve logs to store the data.

### 2.2.4 Data Extraction Sources

The source system is not always a database during the extraction process, there are popular alternatives used such as file systems, e-mails, or others. If the source of the extraction is a file system then the data inside could be either structured, semi structured or unstructured. An example of a file system could be a simple .txt file, a structured .txt file system may look like the figure below.

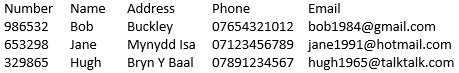


Figure Structured .txt file example

Semis-structured files are very common, more particularly used with XML files. XML uses tags like HTML to show where declarations are made with the data. However, this is more time consuming because there is much more writing to be done. As opposed to listing information in columns, you must write the columns as tags for every row of information. This is shown in the example below.



Figure Semi-structured xml file example

Unstructured data is the most time consuming to extract. This is because it is essentially data used in sentences, such as: “His name is Bob”. On a much larger scale, this would have to be inputted into a file system or XML file to then be ready for extraction.

Finally, extraction is a very important stage of the ETL process. It is the first step of the process and therefore the following stages after this rely on an efficient extraction being performed every time to ensure a faster performance. This would then positively impact the runtime of the transformation process where the cleansing and normalizations are used.

## 2.3 Transform

### 2.3.1 Overview

Transform is a very important part of ETL. When data has been extracted to the server from files, it is essentially raw data which will need to be cleansed, mapped and transformed. This is important from a business aspect because it allows the data to be clean, precise and efficient to work with. Transformation involves many tasks, some basic transformation tasks include: selection, matching, and data cleansing. In addition, there are several important transformation tasks that are often performed. These include: standardising data; conversion of measurement units; removing duplicates; merging fields; and calculated and derived values.

### 2.3.2 Data Cleansing

Data cleansing is performed just before the transformation process, so it improves data quality. This is typically known as removing any ‘dirty data’. Data cleansing involves a few aims to improve the data quality. A very important part is removing any duplicated data. A typical example of data cleansing is mapping NULL values to 0, or “Male” and “Female” to “M” and “F”. This is a small change but allows the data to be read quicker and more useful by not having any NULL values where possible. Duplicate data is not useful as makes the future process slower if containing any duplicate values. Also, this would include removed any values that are not possible. For example, a person’s date of birth may have been recorded as 30/2/2021 which is impossible to have. This value would either be removed or the whole row would have to be removed too.

Furthermore, data cleansing can get confusing when checking for spelling errors and different practices used. An example of this could be various fields with the same supposed name but different approaches: Church Road; Chrch Road; and Church Rd. These errors can be timely to repair. In addition, data cleansing may also include unexpected use of a field. For example, the surname field may be used as address etc. this could result in customers not receiving any products they have purchased.

### 2.3.3 Correction and Normalization

Inspecting, normalization and standardization are important aspects in the transformation phase to improve data quality and accuracy. Normalization and conversion are beneficial because it aids ensuring to make the data uniform. Matching is an important task too, the key benefit of this is matching like-wise values from different sources, such as customer information being recorded in two different areas, which is also duplicate data. An example of using matching and normalization in a business environment could be Vauxhall calculating sales in Europe. The data selected will have to be normalized because some vehicles are names different abroad, although the same car. The manufacturer of the car can be matched simply matching “Vauxhall” with “Opel”. Once the data has been normalized then the data can be loaded properly. The process of cleansing data and transforming uses normalization, standardization and correction to improve the data quality, making it simple, accurate, and efficient. The figure below shows an example of how this process is carried out. For example, using the brief Vauxhall explanation above to store information on a specific car.

Vauxhall Borsa 2015 Red  
1 Litre 3 Doors  
Lton (UK)

make: Vauxhall  
model: Borsa  
year: 2015  
colour: Red  
engineSize: 1  
doors: 3  
city: Lton  
country: UK

make: Vauxhall  
model: Borsa  
year: 2015  
colour: Red  
engineSize: 1  
doors: 3  
city: Lton  
country: United Kingdom

make: Vauxhall  
model: Corsa  
year: 2015  
colour: Red  
engineSize: 1  
doors: 3  
city: Luton  
origin: United Kingdom

Normalization

Correction

Standardization

Figure 6 Process of normalization, standardization, and correction in the transformation process (Golfarelli & Rizzi, 2009)

### 2.3.4 Business Rules

Business rules can be applied during the transformation process of ETL where derived values are created from existing data contained in the databases. This process is completed using calculations, logic and lookups. In addition, business rules can be applied also be applied to prevent the future occurrence of inconsistent values but are logically associated. This can cause problems in a business aspect when later trying to query item numbers, orders numbers, post codes, and more because often characters may be separated by a space or dash. For example, an order number may be stored in three different methods i.e. “8 123 456 7”, “81234567” and “8-123-456-7”. A business rule would need to be applied when storing this data in the database to prevent any future occurrences when querying an order number or any similar data, all three order numbers are logically associated but not the same. Applying business rules when standardizing and normalizing data can modify any similar occurrences into one chosen method.

### 2.3.5 Conclusion

In conclusion, transform is an important part of the ETL process that converts data from the source into the target data warehouse. Good data sources to use in the transformation process will require less transformation, others may require techniques including standardization, normalization and thorough data cleansing to satisfy business requirements set. In addition, poor data sources with inconsistent data may also require removing duplicates, filtering and looking up.

## 2.4 Loading

### 2.4.1 Overview

Loading is the final stage of the ETL process, this where all the data which has been extracted, cleansed and transformed is then loaded into the data warehouse. There are two different steps in which the load can take place, refresh and update. To refresh, this means to erase one or more tables to repopulate with updated data using static extraction. Update is like refresh, but any modifications to data is loaded to the data warehouse as opposed to all previous data being deleted before loading to the data warehouse. Also, using the update step this uses incremental extraction, whereas refresh uses static extraction.

### 2.4.2 Star Schema and Snowflake Schema

Data can be loaded into different schemas during this process, star schema and snowflake schema. The star schema is presented as a star, each point containing a table where all are connected or owned by the middle table. The snowflake schema uses more interconnected tables to make the representation look larger. The key difference between star and snowflake is snowflake schemas are normalized, which is why there are many more tables in comparison to the star. As the snowflake schemas are normalized data, they use less memory on a system therefore can be very beneficial in a business environment. The snowflake schema uses more tables than the star schema, which when querying data, querying the information within can be much more complex because there will be many more JOINs used. In contrast, snowflake schemas are recommended for use in data warehouses because of the key advantage of less storage usage. Also, star schemas are recommended for use in data marts, this is because there is much less data being stored in comparison and would not have much effect in this situation.

### 2.4.3 Fact and Dimension Tables

Fact and dimension tables are an important part of loading in the ETL process. Dimension tables must be created before the fact tables because it is not possible to have facts without the dimensions. The key difference between a fact and dimension table is a dimension table consists of de-normalized data. The fact table is used more often in a business environment as it provides the benefit of easily accessing facts and business information in one table, using primary and foreign keys to produce.

### 2.4.4 Aggregations

Aggregations are used in to improve business intelligence by allowing the user to sum and group information appropriately. This can be used to produce beneficial information to businesses to calculate profit when looking at sales by the product category across different regions in the United Kingdom, but grouping this information by the quarter. The key benefit for using aggregations in a business is having the ability to calculate the sum of data in one table when taken from multiple fact tables, this can then be represented in a multidimensional model such as the cube. Fact tables are generally used to perform SELECT statements to grab data from multiple tables. Depending on the volume of data being used, a simple query executed on a fact table may fetch thousands or millions of rows. For example, if Vauxhall would like to calculate the number of sales in a quarter across the North-West of England. There are 50 Vauxhall dealerships in the region, and 1200 sales every day in each dealership for a quarter. This query would be 50x1200x90=5,400,000. In contrast, this would return more than 5 million rows of data, using aggregations would result in 50 rows, a single row for each dealership. Using a data warehouse with no aggregations in this scenario is not recommended due to the volume of data being used, this would result in performance issues of the system. The figure below shows an example of aggregations used in the above explanation where the region is grouped by ‘Merseyside’ in the North-West.

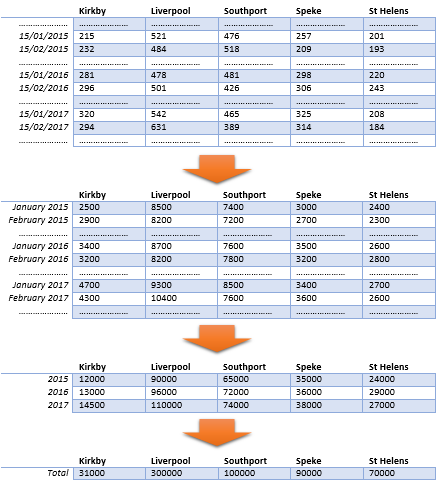


Figure Process of how aggregations may be used in a business environment (Golfarelli & Rizzi, 2009)

The figure above shows an example of how an aggregation may be presented. This example represents selecting sales information that have been grabbed from the different store tables to produce one fact table. This can be grouped in different stages, by month and year to then calculate the total. This is beneficial for staff further up the hierarchy such as a regional manager to see this information.

# 3. Case Study

There are a range of ETL tools available to users and often allow users to be able to create a graphical representation of data flow on the ETL process. This is very useful because it allows the user to add in tools that include merge, pivot, multicast, and many more tools. Microsoft’s ETL tools is Microsoft SQL Server Integration Services (SSIS), Oracle Data Integrator (ODI), DataStage and more. In the figure below this shows an example of how data flow may be represented in SSIS, this example makes use of multicast, lookup, sort, union all, and merge join. (Golfarelli & Rizzi, 2009)

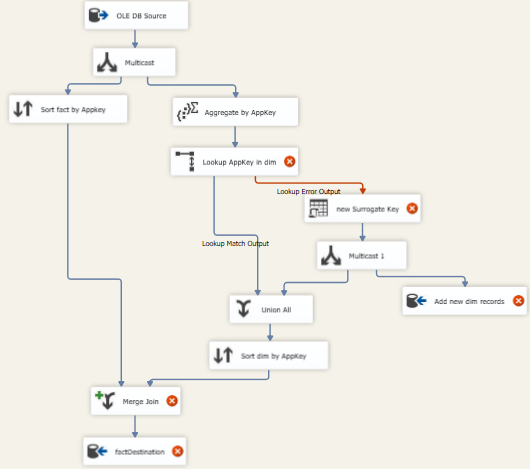


Figure Example of data flow using SSIS tools in Microsoft SQL Server Integration Services

There are a number of useful tools available in the SSIS toolbox when using Microsoft Visual studio. Fuzzy Grouping allows the user to identify similar items within a specific data range. This operates similar to Fuzzy Lookup. For example, using Fuzzy lookup, you are able to select groups of data with similar ranges such as, an item number that may use a hyphen, space or none to break the number up. The Fuzzy Lookup can find data as previously described but also change the data to one specific format using data correction or providing missing values. This can be managed to enter business rules to ensure all item numbers are using the same structure.

Figure 9 Fuzzy Grouping & Fuzzy Lookup in SSIS tools

Multicast is another tool that is very useful in SSIS. The purpose of multicast is to distribute the input to multiple outputs. As shown in the diagram above, the multicast tool is used to distribute the OLE DB Source to sorting and aggregate. It is then also used again to distribute the new surrogate key to add new dim records and union all.

Figure 10 Multicast in SSIS tools

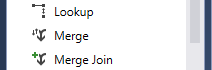
The lookup tool is similar to fuzzy lookup but in some ways not as intelligent. The function of a lookup is to join input data with columns in a specific data set. It widely used when a user wishes to retrieve information and data in a desired table based on specific values.

Figure 11 Lookup in SSIS tools

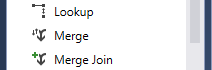
Merge is a more popular tool used in SSIS data flow. This can be used when the user wishes to combine two data ranges into one single range. The merge transformation provides a number of benefits such as, being able to create datasets, merge data from two sources that may include text files or database tables. The merge join function provides a similar function to the standard merge, however uses SQL queries to do so. It often uses LEFT, RIGHT and INNER joins depending on the situation of the merge join the user wishes to execute.

Figure 12 Merge & Merge Join in SSIS

When using Microsoft SQL Server Integration Services, you are able to create PivotTables on select data you may have. The SSIS toolbox gives you the option to Pivot and Unpivot with the data you have chosen. The benefit of using Pivot tables is you are able interactively select which columns you wish to display in the table. For example the figure below shows a PivotTable created in Excel, showing data of computer game sales where I would like to see how much profit, how many purchased and cost to the distributor. In addition, the PivotTable also allows you group large sets of data, automatically choosing year and quarter. Within the quarter it then also groups the games by the production company, whether it is Microsoft, Electronic Arts or 2K Games. You can see with the expanded tab here that it will calculate the sum for each product and category etc. This is very beneficial in a business environment because it gives them the access to turn data into information. A key difference between Excel and SSIS for PivotTables is SSIS requires the user to enter the SQL statement that will include INNER JOINS to produce the table, Excel will product automatically when the data is highlighted.

Figure 13 Pivot & Unpivot in SSIS tools



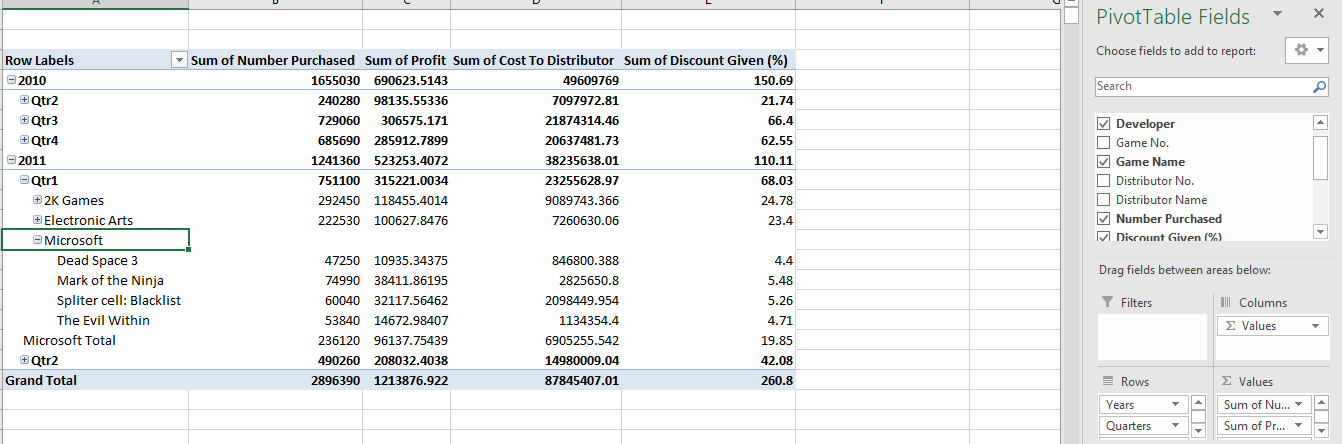


Figure Representation of an Excel PivotTable

# Conclusion

In conclusion, ETL plays an important role throughout the construction and maintenance of a data warehouse or data mart. ETL is used whenever the data warehouse or data mart is first being constructed and then must be performed every time any modifications are made to the data within. OLTP comes before the ETL process and serves the function of writing SQL statements or text files for the files to later be extracted to begin the ETL process.

Extraction is the first stage of the ETL process where the data may be used in various different methods, pull, triggers, log, and scheduled. In addition these methods may use online and offline extraction, each very useful depending on the business situation.

Furthermore, the transformation stage of the ETL process is crucial to ensure data is effectively cleaned and efficiently normalized and standardised to ensure the quality of data in the later stages of the data warehouse. Business rules may also be applied to ensure correction is performed properly to ensure the consistency of data.

In addition, loading is followed by transformation in the ETL process. The loading process may use two different methods, update which uses incremental extraction, or refresh which uses static extraction. Also, the star schema and snowflake schema may be used to present the tables in a database. The key difference between the two is the star schema is primarily used in data marts as it is not very efficient with data storage. The snowflake schema is normalized and uses much more branches to ensure data is stored more efficiently, this is why it is more common to be used in a data warehouse.

After the ETL process, OLAP may be used to perform analysis services on the data, aggregations and multidimensional models such as the cube are popular choices. OLAP provides benefits to businesses by aiding the improvement of planning, decision making and problem solving.

Microsoft SQL Server Integration Services (SSIS) is a popular ETL tool used to populate data warehouse and data marts, cleaning and standardizing data, and applying business intelligence for the ETL transformation process. It is also used to present a graphical representation of data flow in the data warehouse or data mart. Other alternative ETL tools to SSIS include Oracle Data Integrator (ODI), SAS Enterprise Data Integration Server, Informatica PowerCenter, and more.

# References

Ahmed, A. (2012, March 27). *0104 The History of Databases*. Retrieved October 2017, from YouTube: https://www.youtube.com/watch?v=jB3GRAHVa8c

Chapple, M. (2017, June 8). *Facts vs. Dimensions Tables in Database*. Retrieved November 2017, from ThoughtCo.: https://www.thoughtco.com/facts-vs-dimensions-1019646

Data Integration Info. (n.d.). *ETL (Extract-Transform-Load)*. Retrieved November 2017, from Data Integration Info: http://www.dataintegration.info/etl

Datawarehouse4u. (2009). *OLTP vs. OLAP*. Retrieved October 2017, from Datawarehouse4u: http://datawarehouse4u.info/OLTP-vs-OLAP.html

Drkušić, E. (2016, April 28). *Star Schema vs. Snowflake Schema*. Retrieved November 2017, from Vertabelo: http://www.vertabelo.com/blog/technical-articles/data-warehouse-modeling-star-schema-vs-snowflake-schema

ETL Tools. (n.d.). *Data loading ETL process*. Retrieved from ETL Tools: http://www.etltools.org/loading.html

ETL Tools. (n.d.). *Data transformation in the ETL*. Retrieved November 2017, from ETL Tools: http://www.etltools.org/data-transformation.html

ETL Tools. (n.d.). *ETL - Extract Transform Load*. Retrieved November 2017, from ETL Tools: http://www.etltools.org

Fortune, S. (2014, February 27). *A Brief History of Databases*. Retrieved October 2017, from Avant: http://avant.org/project/history-of-databases/

Golfarelli, M., & Rizzi, S. (2009). *Data Warehouse Design.* (W. Rinaldi, & L. Theobald, Eds.) Bologna, Italy: McGraw-Hill. Retrieved November 2017

Hughes, G. (2017, September 27). *6048 Lecture 01*. Retrieved October 2017, from Canvas: https://canvas.ljmu.ac.uk/courses/12276/pages/6048-lecture-01?module\_item\_id=228876/6048 Lecture 01.pptx

Hughes, G. (2017, September 27). *6048 Lecture 02*. Retrieved October 2017, from Canvas: https://canvas.ljmu.ac.uk/courses/12276/pages/6048-lecture-02?module\_item\_id=228891/6048 Lecture 02.pptx

Hughes, G. (2017, October 18). *6048 Lecture 03*. Retrieved October 2017, from Canvas: https://canvas.ljmu.ac.uk/courses/12276/pages/6048-lecture-03?module\_item\_id=240587/6048%20Lecture%2003.pptx

Kandasamy, S. (2014, November 10). *Advanced Database Management System - Tutorials and Notes*. Retrieved October 2017, from ExploreDatabase: http://www.exploredatabase.com/2014/11/advantages-of-oodbms.html

Laudenschlager, D., Hamilton, B., Peterson, T., & Guyer, C. (2017, March 14). *Fuzzy Lookup Transformation*. Retrieved November 2017, from Microsoft: https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/fuzzy-lookup-transformation

Laudenschlager, D., Hamilton, B., Peterson, T., & Guyer, C. (2017, March 14). *Lookup Transformation*. Retrieved November 2017, from Microsoft: https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/lookup-transformation

Laudenschlager, D., Hamilton, B., Peterson, T., & Guyer, C. (2017, January 3). *Merge Join Transformation*. Retrieved November 2017, from Microsoft: https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/merge-join-transformation

Laudenschlager, D., Hamilton, B., Peterson, T., & Guyer, C. (2017, January 3). *Multicast Transformation*. Retrieved November 2017, from Microsoft: https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/multicast-transformation

Liverpool John Moores University. (2017, September 27). *6048 Lab 01 - SSAS*. Retrieved November 2017, from Canvas: https://canvas.ljmu.ac.uk/courses/12276/files?preview=395948

Oracle. (2017). *Database Data Warehousing Guide*. Retrieved November 2017, from Oracle: https://docs.oracle.com/cd/B19306\_01/server.102/b14223/extract.htm

Prabhat, S. (2011, June 9). *Dimension Table and Fact Table*. Retrieved November 2017, from Difference Between: http://www.differencebetween.net/technology/hardware-technology/dimension-table-and-fact-table/

Quick Base. (2017). *A Timeline of Database History*. Retrieved October 2017, from Quick Base: http://www.quickbase.com/articles/timeline-of-database-history

Rainardi, V. (2008). *Building a Data Warehouse* (1st ed.). (J. Pepper, Ed.) Berkeley, California, United States of America: Apress. Retrieved November 2017

Takkar, V. (2015, September 8). *3 - ETL Tutorial | Extract Transform and Load*. Retrieved November 2017, from YouTube: https://www.youtube.com/watch?v=WZw0OTgCBOY

Tutorials Point. (2017). *DWH - Schemas*. Retrieved November 2017, from Tutorials Point: https://www.tutorialspoint.com/dwh/dwh\_schemas.htm

Wikipedia. (2017, October 13). *Data transformation*. Retrieved November 2017, from Wikipedia: https://en.wikipedia.org/wiki/Data\_transformation#Data\_Transformation\_Process

Wikipedia. (2017, October 24). *Data warehouse*. Retrieved October 2017, from Wikipedia: https://en.wikipedia.org/wiki/Data\_warehouse

Wikipedia. (2017, October 20). *Database*. Retrieved October 2017, from Wikipedia: https://en.wikipedia.org/wiki/Database#History

Wikipedia. (2017, October 28). *Extract, transform, load*. Retrieved October 2017, from Wikipedia: https://en.wikipedia.org/wiki/Extract,\_transform,\_load

Yousuf, F. (2015, June 3). *ETL (Extract, Transform, and Load) Process & Concept*. Retrieved November 2017, from Applied Informatics: http://blog.appliedinformaticsinc.com/etl-extract-transform-and-load-process-concept/