Bachelor's thesis

Degree Programme in Information Technology

Information Technology

2019

Jenna Vivi Maria Kosonen

DATA WAREHOUSE USAGE IN AN INSURANCE COMPANY ENVIRONMENT

TURKU AMK
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR'S THESIS | ABSTRACT

TURKU UNIVERSITY OF APPLIED SCIENCES

Information Technology

2019 | 23

Jenna Vivi Maria Kosonen

DATA WAREHOUSE USAGE IN AN INSURANCE COMPANY ENVIRONMENT

A data warehouse is the main source of data within a company. All the relevant inputs to leads and trends can be built up on data warehouses data collections. It is important to keep the data relevant and updated to be used in daily bases.

The main goal of the thesis was to expand the readers' knowledge of data warehouse and describe the different methods used to bring in the data that is needed. Additional goals of the thesis were to show how the data is loaded in to the database and to explain what implications GDPR has for a data warehouse.

The goals of this thesis were achieved by having a communication between the IT department and the other departments of an insurance company to know what data needed to be stored for data analysis within the insurance company.

The thesis explains the Extract, Transfer and Load method which is the main method of extracting, transferring and loading data into the database. Historical data loadings are usually the only way that data is loaded in to the core of databases, so that there is a timestamp when an older version was relevant.

A great amount personal data can be found within the data warehouse of an insurance company, therefore data protection needs to be taken into account. According to the General Data Protection Regulation, GDPR, customers have right to access any information that a company holds on them, and the right to know why that data is being processed, how long it is stored for, and who can see it.

KEYWORDS:

Data warehouse, GDPR, ETL, historical loadings

CONTENTS

FIGURES

LIST OF ABBREVIATIONS

DS	DataStage
DWH	Data Warehouse
ETL	Extract Transfer Load
OLAP	Online Analytical Processing

1 INTRODUCTION

A data warehouse (DWH) is a system that is used for reporting and data analysis. Data collection is important in an insurance environment because based on the data that is collected, the business can create leads to gather more customers. Since mainly the business side creates and needs the information about customers, a data warehouse needs to work closely with the business to deliver what is needed.

The aim of this project was to describe the methods and explain the tools to expand the readers' knowledge of DWH in insurance environment. Working with the data in an insurance company is considered as backend work, therefore, a good co-operation with business and data providers is crucial. After collecting the data from various sources, the business user can decide how they would like the data to be displaced within their reports or how they would like it to be stored within the DWH. Business intelligence systems provide the ability to answer critical questions by turning the massive amount of data into a format that is current, actionable, and easy to understand. (Guerra & Andrews, 2013, 2) There are several methods that the data can be displayed, for example, a report made with a reporting tool or a view made in the database with only the variables displayed that the business has asked for. The business uses the data to analyse the current and long-term trends, alert instantly to opportunities and problems and give continuous feedback. (Guerra & Andrews, 2013, 2)

A data warehouse has a great amount of personal data in the database and this needs to be kept in mind when testing the data. There should not be any individual lookups to a person or any lookup to identify a specific customer. Since the General Data Protection Act took effect in May 2018 (https://gdpr-info.eu/) , all the old data should be anonymized. The source system deletes all the old data and then data warehouse has processing rules for data anonymization.

The thesis is structured as follows: Chapter 2 describes the methodologies how the thesis was achieved. Chapter 3 and Chapter 4 introduce the theory of data warehouse and explain in more detail what the different tools and methods used in DWH and GDPR. All the practical parts are in Chapter 5, that is, the explanation of how data is processed and GDPR is implemented.

2 METHODOLOGY

DWH is not well known within the IT industry or the business but is one of the most important parts of a company. A business needs a great amount data to be able to build trustworthy reports and leads to acquire more customers. The thesis was carried out in collaboration with a business user who needed more customer data in DWH to be able to report on how people respond when they receive a call.

To be able to provide this information to the end user, the data needed to be extracted from the source system. The data was in text format, therefore it needed to put through a transformer to change the data types. Because the end user wanted to have a historical loading, timestamps needed to be added in to the data so that they line up with customers historical information. After all the changes, the data was transformed to data could be loaded in to the DWH and made in to a report.

Because all of this was customer data, there needed to be system to separate the fields that are GDPR-related and according to DWH rules, they will be anonymized when the customer has been deleted from the source system.

3 DATA WAREHOUSE

A data warehouse is relational database designed for reporting and analysis. It is a collection of data from one or more sources which are then organized so that the data can easily be used, managed and updated. A data warehouse usually includes historical long-rage data from transactional data.

A data warehouse environment may include an extraction, transaction, online analytical processing (OLAP), and data mining. As a comparison to a database, a data warehouse contains infinite applications and infinite databases. A data warehouse is usually separated from the front-end applications which allows it to be scalable. A data warehouse with OLAP is mostly designed to facilitate reporting and analysis, not for a quick hinting transactional need. A data warehouse is designed to handle large analytical queries.

There are four distinct characteristics in data warehouse as William Inmon (Inmon, 1994, 31) has described:

- Subject-Oriented
- Integrated
- Nonvolatile
- Time-variant

1. **Subject-Oriented** means that the data warehouse is built around the organizational needs. (Inmon, 1994, 31) The date that is stored in the database is what the business requires and what the analytics need.
2. **Integrated** data means that data coming from various sources is converted, reformatted, re-sequenced, summarized and with the same physical corporate image. There should be a consistency in encoding, naming conventions, physical attributes, measurement of attributes. Integration assures that all the databases across the data warehouse have a common naming method in the tables so that the tables are easier to join.
3. **Nonvolatile** means that once the data is added from a source system, it should not be changed. If there is a change in the data, there should be a new row added which would tell that the old one has been updated and not deleted.

4. **Time-variant** means that the historical data is kept in the data warehouse. This is, for example, to discover trends in business, analytics need a large amount of data. Time-variant in a data warehouse focuses on changes over time.

As Kimball as described what a data warehouse is, there is a back room and front room. (Kimball, 2017, 15) The back room in this case means that all the development work for the data received is done, for example, ETL, cleaning and making sure that the data structure is as the business or analytics want it to be. The front end is where the end users are. They present what is important from a DWH, investigate the causes, try what-if situations and track the decisions and changes back to the DWH. With DWH developers and the business, there is a close connection since the DWH needs to listen to the business needs and bring data available for them to use.

3.1 ETL and DataStage

ETL stands for extract, transform and load. ETL is a method used in a DWH to reduce costs to be reliable and automated the load jobs can run every day, usually at night. To be extensible when the organization grows the DWH grows, to be transparent so that everyone in the organization know that when the information is available and how to access it. DWH also aims to be compliant, fast and relevant. (Kimball, 2017, 5)

The extract in ETL means that the requirements and documentation of the data being uploaded have been completed, the source has been created with the requirements from the business. After the source has been established, the isolation of changes needs to be thought through. It is important to consider: 1. What the end user expects to get out of the data that the DWH provides. 2. How the data is stored, is it loaded updates daily or more rarely. 3. Weather the data has a historical data collection, or will it only be updated with current data. 4. The GDPR aspect of the data, if it has some personal information stored, how the data deletion will be handled and how is allowed to see the data.

After these steps, the data can be extracted from the source and loaded in to the DWH staging area. The staging area is where the data is extracted straight from the source without any transformation or format changes.

After the data has been loaded in to a staging table, the data can be cleaned by using the transformation in the ETL. The transformation step is the most time consuming and

requires the most work from DWH side. The data needs to be structured correctly so that the data qualifies to be entered in to the target table. The data will be controlled and checked that there are no duplicates in the process.

The last step is to load the date in to its target table. In this step, the attributes are changed so that they are as required, the unique identifiers also known as surrogate keys are given to the data, so it can be easily traced. Finally, the fact table is inserted and updated.

There are several ETL tools that can do ETL work, but this thesis will only concentrate on a tool called DataStage because the company that commissioned the thesis had this tool as their main ETL tool. DataStage (DS) is an IBM tool that has a graphical notation to construct data integration solutions. By using DS, the data can be read from any kind of a source in to the application (Figure 1).

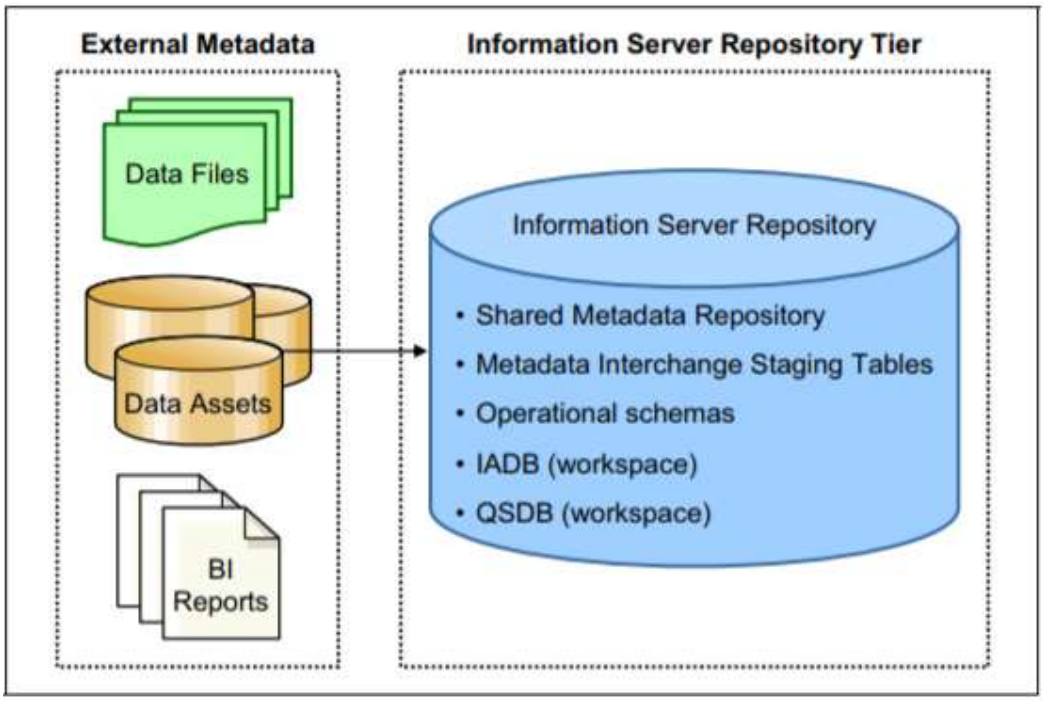

Figure 1. Data can be read from all sources in to the DataStage

When external data is loaded with DS in to a staging, the staging area provides an interim load area for external metadata, where this new metadata can be compared with existing

metadata (before loading) to assess the impact and act on the new metadata, and keep a record of the loaded metadata. (IBM, 2012, 8)

DataStage has a rich interface which provides straightforward methods to design jobs that extract, transform and load.

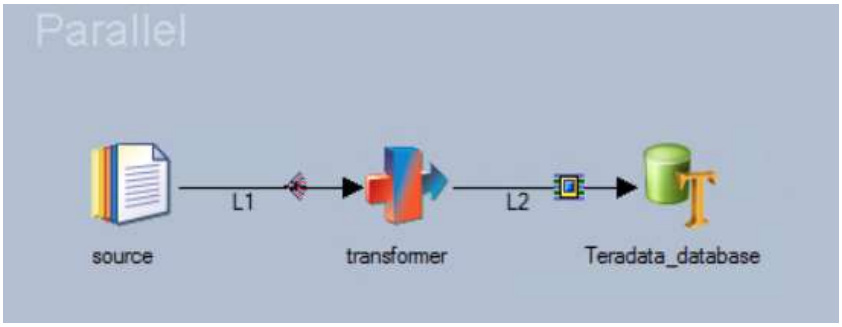

Figure 2. Example of DataStage ETL job.

As Figure 2 shows, there is a source file which then goes through the transformer and is then loaded in to a database, in this case Teradata. There are many ways to carry out this process is possible and the one shown in Figure 2 is the simplest. With DS, there is a possibility to read several files from different sources and add them together in a target table or staging area as Figure 3 shows.

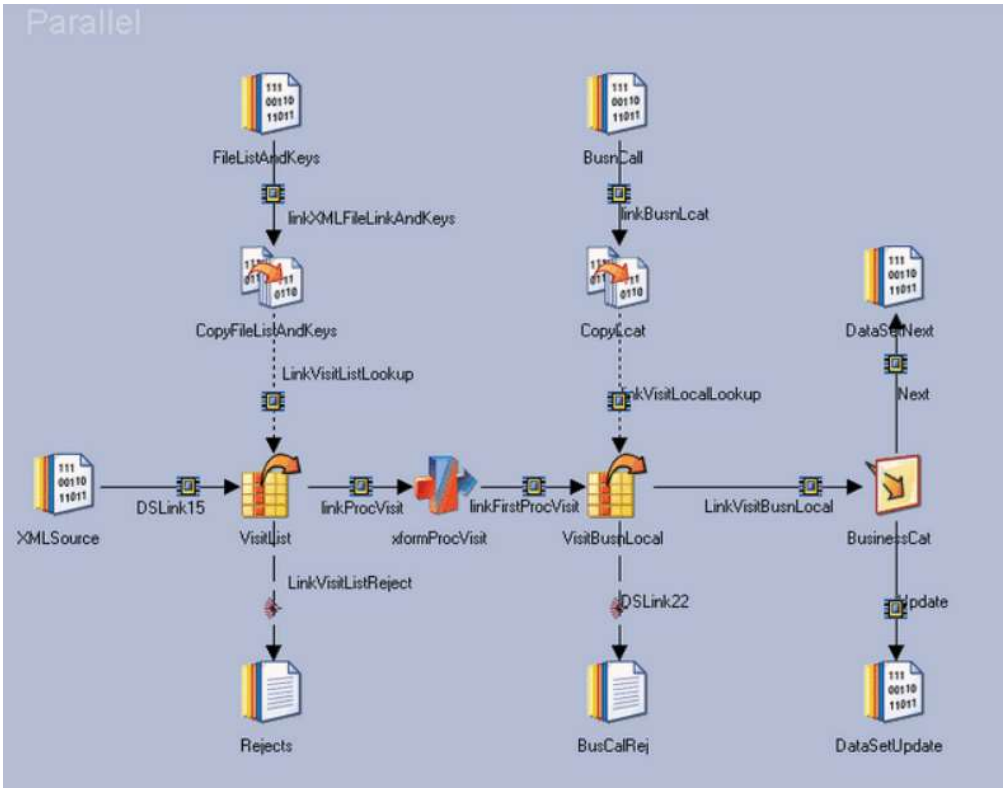

Figure 3. Adding several sources in DataStage.

3.2 Teradata as the database for DWH

A database is a collection of related data (Elmasri & Navathe, 4, 2017). It may be generated manually, or it can be computerized.

There are several kinds of databases that can be used within the data warehouse and that can relate to DataStage. Teradata is a fully scalable relational database management system and it is widely used to manage large data warehousing operations. Teradata acts as a single data source that can run many numbers of requests from multiple client applications.

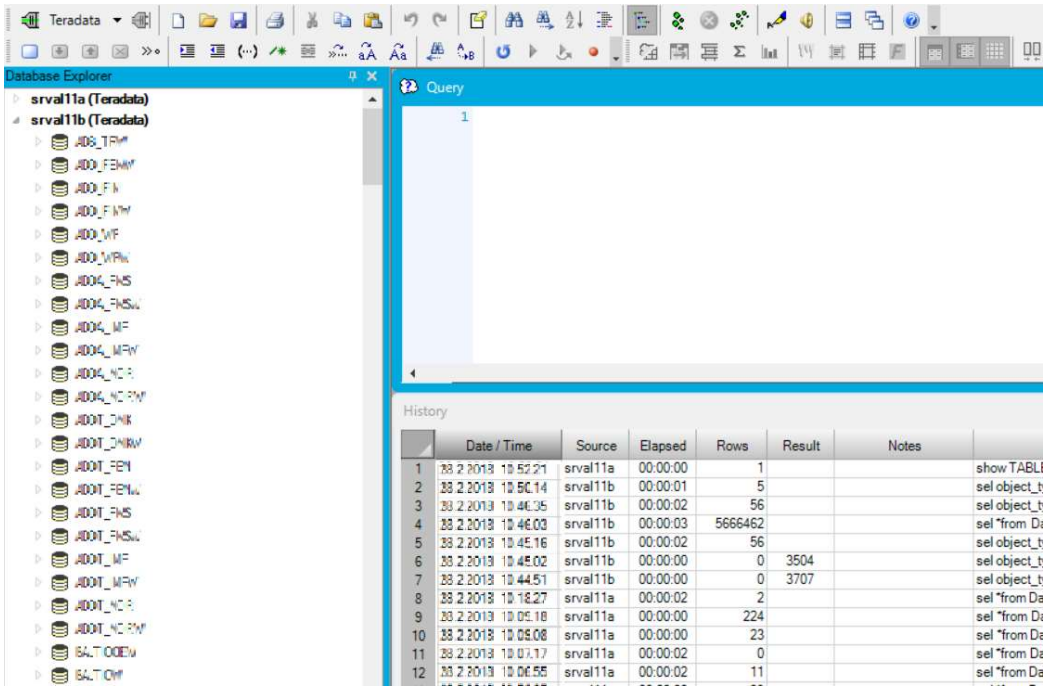

Figure 4. Teradata connection in SQL Assistant.

When sending requests in to Teradata, there are applications that need to be used to connect with the data. Teradata uses a basic Teradata query (BTEQ), it allows SQL statements and BTEQ commands. BTEQ can be used for importing, exporting and reporting purposes.

Within the database, there can be several different environments, for example, one for the development environment, one for testing, and one for production. The developers work on the development side and once the data goes to business testing, it can be transferred in to the test environment where the testers, usually the ones that have asked for the information, can test if the table is relevant.

4 DATA PRIVACY

The data privacy regulation is, also known as General Data Processing Regulation GDPR (https://gdpr-info.eu/). GDPR became one of the most discussed topics of 2018, because it applies to companies which process personal data. This means that people would have more control over their personal data how it is being used. ‘Controllers' and 'processors' of data need to abide by the GDPR. A data controller states how and why the personal data is being processed, while a processor is the party doing the actual processing of the data. (Curtis, 2018, ITPRO) So the controllers are usually the organizations business side and then the processers are the IT department who is doing all the processing what the business requires. When the data is being processed, it needs to be lawful, transparent, and for specific purpose.

Since GDPR is about deleting personal data, people have the right to access any information that a company holds on them, and the right to know why that data is being processed, how long it is stored for, and who can see it. (Curtis, 2018, ITPRO) The deleting process must be transparent to everyone and people need to be able to know what kind of data companies collect from them, what do they do with the data and how they process it. People have also the right to demand companies to delete their personal data, if it is no longer necessary for the same purpose as it was when collected. This is known as “right to be forgotten”. In addition, if a person wants to move their data from an organization to another, the controller must be able to store all the data in a file, for example a CSV file, and send that to the other organization free of charge.

5 ETL IN DATA WAREHOUSE

An IT department in an insurance company works closely with the business side to create the needed data. The data that end user requests could be either to update a table with new data or then create a completely new resource for them to use. When updating tables and adding new columns, the IT department needs to make sure that the old data is kept as it is and the new field is added, possibly with initial data. Creating a completely new table and job is easier, since the data can be loaded in to a new table.

The latest request that was given by the business was to create a job that would read customer experience data to Teradata, from which the business could gather information.

For the process, the ETL structure was used. First, the file was read and transformed in to the data type that could be stored in to Teradata.

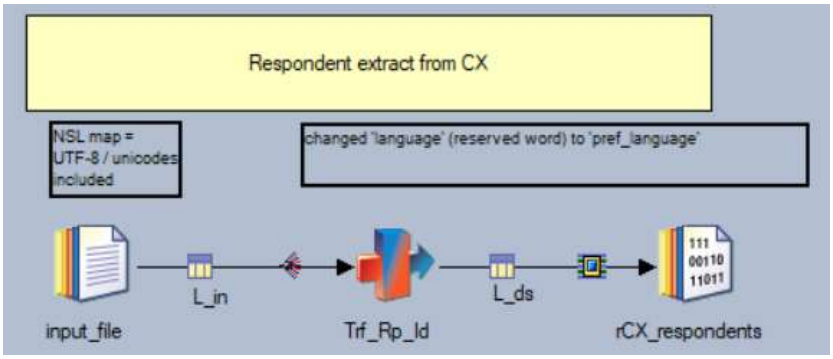

Figure 5. First step in ETL process.

Within the transfer some of the data types were changed because the whole date was read in as varchar.

L_in

	Column name	Key	SQL type
1	respid	☑	VarChar
2	AnonFlag	☐	VarChar
3	Brand_id	☐	VarChar
4	CaseManagerData	☐	VarChar
5	Channel	☐	VarChar
6	Cntry	☐	VarChar
7	CreatedDate	☐	VarChar
8	CuBeClass	☐	VarChar
9	Customer_number	☐	VarChar
10	CustomerClass	☐	VarChar
11	Date1	☐	VarChar

L_ds

	Column name	Key	SQL type
1	respid	☑	Integer
2	AnonFlag	☐	Integer
3	Brand_id	☐	VarChar
4	CaseManagerData	☐	VarChar
5	Channel	☐	VarChar
6	Cntry	☐	Char
7	CreatedDate	☐	Timestamp
8	CuBeClass	☐	VarChar
9	Customer_number	☐	VarChar
10	CustomerClass	☐	VarChar
11	Date1	☐	Date

Figure 6. Transformation for the data type.

When changing the data type from varchar to data, the data needs to be reformatted within the transformer.

```
If L_in.Date1 = "" Then SetNull()
Else If L_in.Date1[2,1] = "." And L_in.Date1 [4,1] = "."  Then L_in.Date1 [5,4] : "-" : "0" :L_in.Date1 [3,1] : "-" :  "0" :L_in.Date1 [1,1]
Else If L_in.Date1 [2,1] <> "." And L_in.Date1 [5,1] = "."  Then L_in.Date1 [6,4] : "-": "0" : L_in.Date1 [4,1] : "-" :  L_in.Date1 [1,2]
Else If L_in.Date1 [2,1] = "." And L_in.Date1 [5,1] = "."  Then L_in.Date1 [6,4] : "-": L_in.Date1 [3,2] : "-": "0" :  L_in.Date1 [1,1]
Else If L_in.Date1[7,4] = "9999" Then L_in.Date1 [7,4] : "-" : L_in.Date1 [4,2] : "-" : L_in.Date1 [1,2]
Else L_in.Date1 [7,4] : "-" : L_in.Date1 [4,2] : "-" : L_in.Date1 [1,2]
```

Figure 7. Data fixed from VarChar to Date format.

When the data is completely modified to the right format in the transformation, it can be loaded in to a reference table in Teradata.

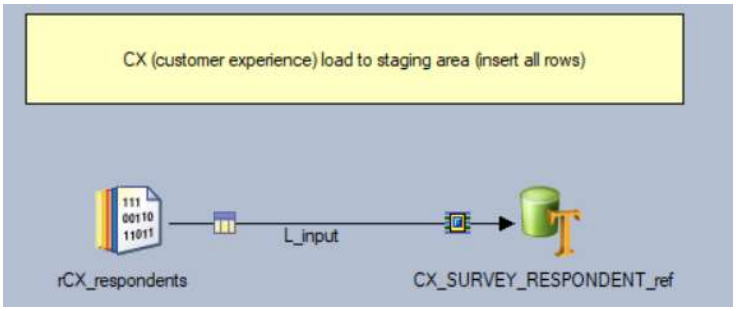

Figure 8. Data loaded in to a reference table

The reason why the data is loaded in to a reference table is because if the data is loaded in to the same table daily, there would be several different entries valid at the same time. Here is where the history data comes in. When the source system sends

the valid data, the DWH needs to compare the new data with the old one. If there are any changes in the data, the old row needs to be ended and the new row needs to start from that point on.

	Txn_Fm_Tms	Txn_To_Tms
1	5.5.2011 09.20.03	25.5.2016 20.52.15
2	25.5.2016 20.52.15	23.6.2016 12.45.16
3	25.5.2016 20.52.15	23.6.2016 12.45.16
4	23.6.2016 12.45.16	11.8.2017 03.55.48
5	23.6.2016 12.45.16	11.8.2017 03.55.48
6	11.8.2017 03.55.48	31.12.9999 23.59.59

Figure 9. Validity periods.

Figure 9 is good example of the different validities, as the first time this data appeared was on the 5th of May 2011. There was a change in this data on the 25th of May 2016 as the old row was ended and the new row was added. The last change for this data happened on the 11th of August 2017 and this data is still valid. The way how the latest or the one that is valid can be separated from the others is with the Txn_To_Tms = ’31.12.9999 23:59:59’. The date ’31.12.9999 23:59:59’ is taken to be the separator because there is no such a date in the calendar, so it is easy to separate what is valid. If this date had disappeared from the source system completely there would not have been Txn_To_Tms = ’31.12.9999 23:59:59’, we would only see the date row as in Figure 10.

	Txn_Fm_Tms	Txn_To_Tms
1	5.5.2011 09.20.03	25.5.2016 20.52.15
2	25.5.2016 20.52.15	23.6.2016 12.45.16
3	25.5.2016 20.52.15	23.6.2016 12.45.16
4	23.6.2016 12.45.16	11.8.2017 03.55.48
5	23.6.2016 12.45.16	11.8.2017 03.55.48

Figure 10. When data is no longer valid

All this logic for the valid data is added within the BTEQ job as a part of the ETL.

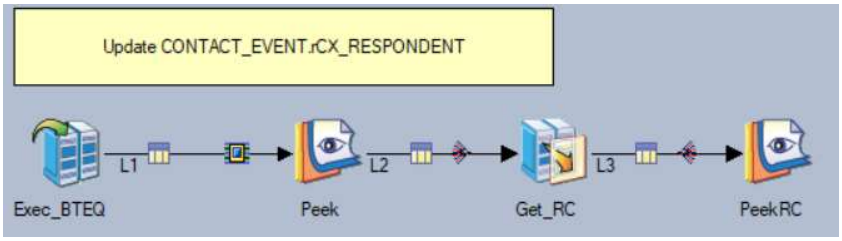

Figure 11. BTEQ job for data validity.

Within the BTEQ job, there is usually an update in to the target table to set the Txn_To_Tms equal to the reference table Txn_Fm_Tms where the target tables Txn_To_Tms equals to the high date. After the data is updated, there is an insert if the data field is new. The data from the reference table is inserted in to the target table with the Txn_To_Tms as high date. This loop is repeated each time the file is read through the DS job.

In addition to this, as the date for the data privacy is approaching, the source system had created a flag to indicate if this data was anonymized. There was a field called AnonFlag, when the field AnonFlag was equal to 2, then the data was anonymized in the source system. The data anonymization took place in the same BTEQ job as the update. If the reference table has the Anonflag = '2', then all the data about that specific customer is being anonymized in the target table.

```
set  email = '§',
     Name = '§',
     FirstName = '§',
     LastName = '§',
     customer_number = '§'
```

1	5	§	§	§	§	§	29.3.2018 09.15.05	29.3.2018 09.15.05
2	5	§	§	§	§	§	29.3.2018 10.44.38	31.12.9999 23.59.59
3	6	§	§	§	§	§	29.3.2018 09.15.05	29.3.2018 09.15.05
4	6	§	§	§	§	§	29.3.2018 10.44.38	31.12.9999 23.59.59
5	7	§	§	§	§	§	29.3.2018 09.15.05	29.3.2018 09.15.05
6	7	§	§	§	§	§	29.3.2018 10.44.38	31.12.9999 23.59.59

Figure 12. Data anonymized.

When all the requirements from the end users have been fulfilled and the data is loaded in to the test environment, the data is ready for acceptance test. There is a view created

on top of the table for the business to access and accept it. If the data is accepted, the DS job can be taken in to production and all the tables can be created in to the production server. After data is moved to production, there can be an initial load of the data from the source to have all the data in the table from periods before. A scheduling is ordered so that the job runs automatically at the given time to have the data up to date after the move to production. Usually the whole DWH is being loaded every working night so that the end user has fresh data for the next day.

5.1 Data privacy in an insurance company

There is a huge amount of personal data stored in an insurance company and there are different rules that apply to what data can be stored and for how long. Since there is data stored over the years, there must be a system that deletes or anonymizes this data. The source systems, such as mainframe, should delete the personal data and let the downstream systems to know what they have deleted. Whatever has been deleted from the source system should be anonymized in the DWH side. For this, there has been created a solution so that the source systems send a blacklist and a whitelist to the DWH team.

A blacklist stands for all the data that has been deleted from the source system side and what should be anonymized from the DWH side. The blacklist would be delivered once a month and the DWH should have a system to anonymize this data from the database. A whitelist is the opposite, i.e., all the data that the source system has, and this should be delivered once a year. These two lists ensure that in the database there are only cases that are current. because All the data that is outside of the whitelist should be anonymized.

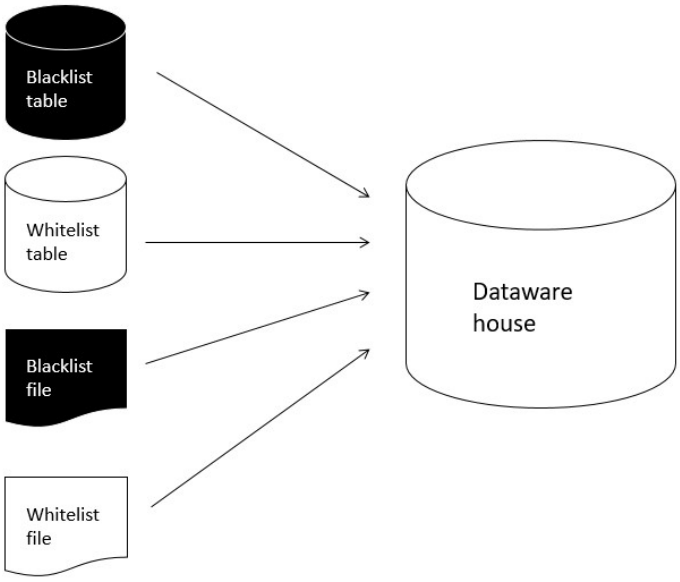

Figure 13. Source systems sends blacklists and whitelists to DWH.

The blacklists and whitelists can be sent as DB2 tables or they can be sent as files. These are read in to the database with a DS job and then anonymized based on whether all the source systems have also deleted the personal data.

When data is deleted from the source system, it is not deleted in the DWH side, but it is anonymized. Within the company studied in this thesis, the anonymization symbol has been decided to be §. When data has been anonymized, the § symbol can be seen in the table for the personal data.

Agrm_Sts_Cd	Agrm_Sts_Dtl_Cd	Agrm_Sts_Rsn_Cd	Sts_Dt	Ext_Refr
OFFER	N/A	000	4.12.2015	§
OFFER	N/A	000	4.12.2015	§
ACTIVE	N/A		12.6.1992	§
CANCEL	N/A	004	20.1.1997	§
ACTIVE	N/A		1.11.1996	§
OFFER	N/A	000	20.5.2013	§
CANCEL	N/A	004	2.11.1997	§
OFFER	N/A	000	4.12.2015	§

Figure 14. Example of anonymized data.

The personal data, such as the social security number, name, address, birthdate and others that can be linked to a person, are anonymized. However, to be able to link the

data of an unknown person, there is a uniquely distinguished entity to connect them all together.

6 CONCLUSION

The aim of this thesis was to expand the knowledge of data warehousing in an insurance company. This thesis explained the methods and tools are and provided examples of how these work in a data warehouse. The thesis also explained the data warehouse process clearly from the start to finish and from the point of business how the data can be manipulated after it has been loaded in to the database. The important co-operation between the business and the IT department has been brought up and the importance of effective communication and understanding of each other.

An ETL tool was introduced and it use was explained as part of a data loading to a database. It is important that there are good processing rules for the source system data and that the tables and data are easily added in to the database. It is the business/company that determines processing rules for the data, how the data should be loaded and how the company would like to orientate the data in to reports or views.

This thesis also addresses the issue of data privacy which was very current at the time of writing this thesis. Data privacy is about anonymization or deletion of personal data. Since data warehouse stores all the data and the history information, there need to be processes to anonymize the old data that has been deleted from the source system. These processes include the compilation of blacklist and whitelist provided by the source systems. The source system would send either a whitelist of people that are still in the system and all the others could be deleted around this source. Alternatively, the source system could send a blacklist which would contain all the deleted information. As a deletion or anonymization job, the whitelist is sent once a year and the blacklist is sent every month.

REFERENCES

General Data Protection Regulation (GDPR) – Official Legal Text: https://gdpr-info.eu/ [Accessed: 21.8.2019]

Joseph Guerra, SVP, CTO & Chief Architect David Andrews. (2013). Why You Need a Data Warehouse. [online]. Available at: magnitude.com/wp-content/uploads/2014/01/2013-03-Why-You-Need-a-Data-Warehouse.pdf [Accessed: 25.5.2019]

W.H. and Hackathorn, R.D. (1994). Using the data warehouse Inmon.

The Kimball Approach | quest for knowledge. (2015). Data Warehouse ETL. [online]. Available at : https://www.q4k.com/training/data-warehousing/data-warehouse-etl-kimball-approach [Accessed: 30.5.2019].

Elmasri, Ramez. and Navathe, Shamkant B. (2017). Fundamentals of database systems.

What is GDPR? Everything you need to know, from requirements to fines http://www.itpro.co.uk/it-legislation/27814/what-is-gdpr-everything-you-need-to-know [Accessed: 21.8.2019]