Current approaches on how to acquire and integrate external data into Data Warehouses

(HS-IDA-EA-03-409)

Carl-Fredrik Lauren (b00carla@student.his.se)

Institutionen för datavetenskap
Högskolan i Skövde, Box 408
S-54128 Skövde, SWEDEN

Examensarbete på det dataekonomiska programmet under vårterminen 2003.

Handledare: Mattias Strand

Current approaches on how to acquire and integrate external data into Data Warehouses

Submitted by Carl-Fredrik Lauren to Högskolan Skövde as a dissertation for the degree of B.Sc., in the Department of Computer Science.

2003-06-04

I certify that all material in this dissertation which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

Signed: _____

Current approaches on how to acquire and integrate external data into Data Warehouses

Carl-Fredrik Lauren (b00carla@student.his.se)

Abstract

Integration of internal data is often mentioned in literature as the most demanding task when building or maintaining a DW. There is no literature that outlines the approach for the integration of external data into a DW. The integration of external data has increased during the last years enabling corporations to understand the opportunities in the market and to be able to better plan for future success of the corporation. The aim of this work is to exploratory outline current approaches for acquiring and integrating external data into DW and to give a brief overview of the future trends for external data integration. This aim was researched using an interview study. The results show that how to integrate external data is depending on what the corporations purpose with the external data is. Additional results show that how to integrate external data also depends on how the data is acquired.

Keywords: Data Warehouse, Extern data, Acquisition, Integration

Acknowledgements

There are several people I would like to thank for their different kinds of support during my work on this dissertation. First, I would like to thank my supervisor Mattias Strand for his valuable support and comments. I would also like to thank my family for their help and support and furthermore Anders Svensson for his valuable comments. Last, but certainly not least, Åsa for her never-ending support and bearing with me during the process of writing this dissertation.

Table of contents

1. Introduction

The data flooding problem in the worlds of science, business and government has according to Singh (1998), been growing during the 1990's. The capabilities for assembling and storing all different categories of data have increased in a way that leaves the knowledge of analysing, summarising and extraction behind. According to Devlin (1997), the concept of Data Warehouse evolved out of two needs: the business requirement of a company-wide view of information and the need from IT departments to be able to manage company data in a better way. The traditional method of data analysis is mainly based on humans dealing with the data directly but when it comes to very large sets of data, this method is not adequate. The database technology has made it possible to effectively store large amounts of data in a well-organised way and likewise the ability to source large sets of data.

Devlin (1997) claims that to be able to efficiently analyse and understand the performance of a corporation there is a need to access all the operational data in a structured way. To increase and understand the opportunities in the market and to be able to better plan for future success of the corporation, more than just the internal data is needed. For example is general market data of importance. This and other kinds of data gathered from outside the corporation are often according to e.g. Devlin (1997) and Inmon (1999a), referred to as external data. The aim of a DW is according to Singh (1998), to unite the information locked up in the operational systems with information from other, often external sources of data. Corporations are gradually obtaining more data from outside their own corporation as the corporations desire to work more closely to their business partners, with third parties and by the increased use of Internet. (Collet 2002; Salmeron 2001)

Inmon (1999a) further argues for the importance of integration of internal and external data as it creates an enhanced foundation for decision-support, i.e. enabling the comparison of internal and external data. But there is according to Inmon (1999a) and Strand & Olsson (2003), problems in incorporating external data with internal data. Problems mentioned by the authors are; difficulties to ensure the quality of external data and difficulties in integrating external data. However, Inmon (1999a), explicitly states that the most difficult problem is to integrate the external data with the internal data.

2. Data Warehouse

In this chapter the terms of a Data Warehouse (DW) will be described. The following sections will be presented: history, definition, architecture and ETL in details.

2.1. History

By automating the reporting and data gathering procedure, the Information Technology (IT) department thought according to Sperley (1999), that business users would be content and the labour reduction the new technology resulted in would be beneficial for the corporations. But the automation of this procedure was not as successful as first intended. The lack of business understanding in IT departments due to the slight interaction between IT departments and business users, led to poorly developed systems. The different departments had their own systems developed and these were often developed without a plan or architecture, resulting in different systems containing different data types and functions. This forced corporations into a situation where systems differed in the use of data types, codes and lengths to represent the same instance of information according Sperley (1999). Additionally there was a problem concerning the meaning of the same data in different systems. For example, a field address in a market system could refer to the mailing address of a customer where the accounts payable system referred to the billing address.

The development of information systems by business operational units resulted in the development of computer systems that contained very large amounts of detailed data concerning a certain area of the business. These are known as information silos. Sperley (1999) claims that the problem of having several different information silos containing data from different areas of the corporation was that it was difficult, almost impossible to get an overall picture of the corporation. This originates from the inability to integrate information from one information silo into another. What the corporation needed claims Sperley (1999), was an integrated source of data about the condition of the corporation.

The data stored in the different information silos could be very precious for the corporation if the data could be combined. The IT department had to write extraction programs enabling the extracted data to be compiled and used as a foundation for creating reports for the different operational units of the corporation.

One of the main benefits of the introduction of the Relational Database Technology was the reduction of quantity in redundant data stored. The different operational systems could with the Relational Databases Technology access the same data for different operational needs.

As the increased use of PCs and desktop tools accelerated these tools became available for more people and many business information workers found out that it was possible for them, without the previous help from the IT department, to access, analyse and even store new data with their tools. The data stored came from different sources such as internal sources and different internal reports made by data analyst and from external sources such as financial newspapers as Wall Street Journal or Business Week. By consolidating the internal data with the external data they acquired advantages in the decision-making process. This technique made business consumers of corporate data to do the same mistakes as IT departments had done before, i.e. developing their own unplanned decision-support environment. The

workers in the different departments generated according to Singh (1998) and Sperley (1999), their own applications and databases based on their own structures and filled it with their own data. This developed into a situation where the data, stored in the databases was defined and calculated differently for every different database created. Users often stored external data in their databases but different users selected different sources for the external data. Because of this the decisions made were not based on the same figures through out the corporation. This resulted in disagreements on which report was the reliable one. The foundation of these disagreements was the unplanned or unarchitected topology of the decision-support environment. What the corporation needed was a central storage containing all the relevant data, defined in the same way, able to produce the reports the corporation needed; a DW. (Singh, 1998; Sperley, 1999).

2.2. Definitions of a Data Warehouse

There are several different definitions of a DW to be found in the literature, for example; Devlin (1997), Inmon (1996) and Singh (1998). The definitions differs somewhat depending on what focus the author had when presenting a DW. Singh (1998) for example, focused more on the use of the DW and on the organisation in question. Devlin (1997) and Inmon (1996), on the other hand, have got a more technical focus. In this work, as mentioned in an earlier section, a technical focus is of higher relevance and therefore chosen rather than an organisational focus. Due to the fact that Inmon (1996) has given the definitions referred to by most authors and that Inmon's definitions are in line with the problem area of this work, these definitions will be presented.

According to W.H. Inmon (1996 p33) a data warehouse is defined as *"a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process."*

The DW is furthermore defined by Inmon et al. (2001 p.93.) as *"the data warehouse is an architectural structure that supports the management of data that is:*

- *Subject-oriented*
- *Integrated*
- *Time-variant*
- *Non-volatile*
- *Comprised of both summary and detailed data.*

The data warehouse exists to support management's decisions which, in turn support the strategic planning processes for the corporation."

In this definition, Inmon further developed the definition by adding the line *comprised of both summary and detailed data*. This is probably done because the subject area has matured over the years and the author's knowledge in the area has increased leading to a further development of the old definition. In this work the last definition of Inmon et al. (2001) will be used continuously. This choice is made because the second definition is later altered and is as a result, more up to date. In the following sections the definition chosen for this work by Inmon et al. (2001) will be described further.

Subject-oriented

By subject-oriented Inmon et al. (2001) means that the data stored in a DW should be relevant for the corporation. The data is organised according to the lines of the major entities of the corporation, for example; customers, products and vendors.

Integrated

By integrated Inmon et al. (2001), means that data coming from different sources are integrated into one entity. Furthermore is integration introduced by Inmon et al. (2001) as the physical fusion and cohesiveness of the data as it is stored into the DW. Many aspects of a DW are covered by integration such as key structures, encoding and decoding structures, definitions of data and naming conventions.

The data is not integrated by extracting it from the operational environment and loading it into a DW. Instead, the data is transformed according to the data model of how the key structure is defined and after the transformation is achieved the data can be integrated and loaded into a DW.

There are however, according to the authors, different meanings and understandings of integration. Inmon (1996) describes integration as data that passes from its source environment to the DW environment. According to Gleason (1997) integration is a part of the transformation process where the detailed data is mapped and united into one entity. In this work integration will have the meaning of the whole process from the extraction to the transformation and loading process, in line with Inmon (1996).

Time-variant

By time-variant Inmon et al. (2001) means that any record in a DW is precise and exact relative to some moment in time. The time constituent could be a year, a quarter, a month or a day. DWs usually contain data up to at least 5 to 10 years.

Non-volatile

Non-volatile data means according to Inmon et al. (2001), that data loaded and stored in a DW never changes. There are however, though very rarely, occasions when a DW is updated. An exception is for example, that a DW could be updated because of some serious mistakes in the extraction, transformation and loading process that populated the DW.

Comprised of both summary and detailed data

A DW contains according to Inmon et al. (2001), both detailed and summary data. The detailed data reflects the transactions made daily in a corporation and contains for example data about sales, inventory movement and account activity.

There are two different categories of summary data: the profile record and public summary. The profile record contains detailed, raw data that is summarised from the extraction and transformation process and then stored into the DW. The other kind of summary data: public record, is departmentally calculated data that has a wide corporate outreach.

2.3. Architecture of a Data Warehouse

Based on the Figure by Chaudhuri and Dayal (1997), the areas relevant for integration of external and internal data will be presented (Figure 1). Chaudhuri and Dayal (1997), divides their architecture of a DW into two parts; the front-end and the back-end and the tools used as front-end tools and back-end tools. The front-end tools are

used for querying and data analysis and are not relevant for this work and no further information will be given about these. First, external data and its features will be presented and second, the focus will be moved to the back-end environment where the back-end tools are used for extracting, cleansing and loading data into the DW. The back-end tools will not be described in this work, only the process they perform. Subsequently the data repository and meta data will be introduced and in the following sections the presentation of the relational databases and the data marts will take place. The OLAP server is the only part that will be described from the front-end area due to the focus of this work.

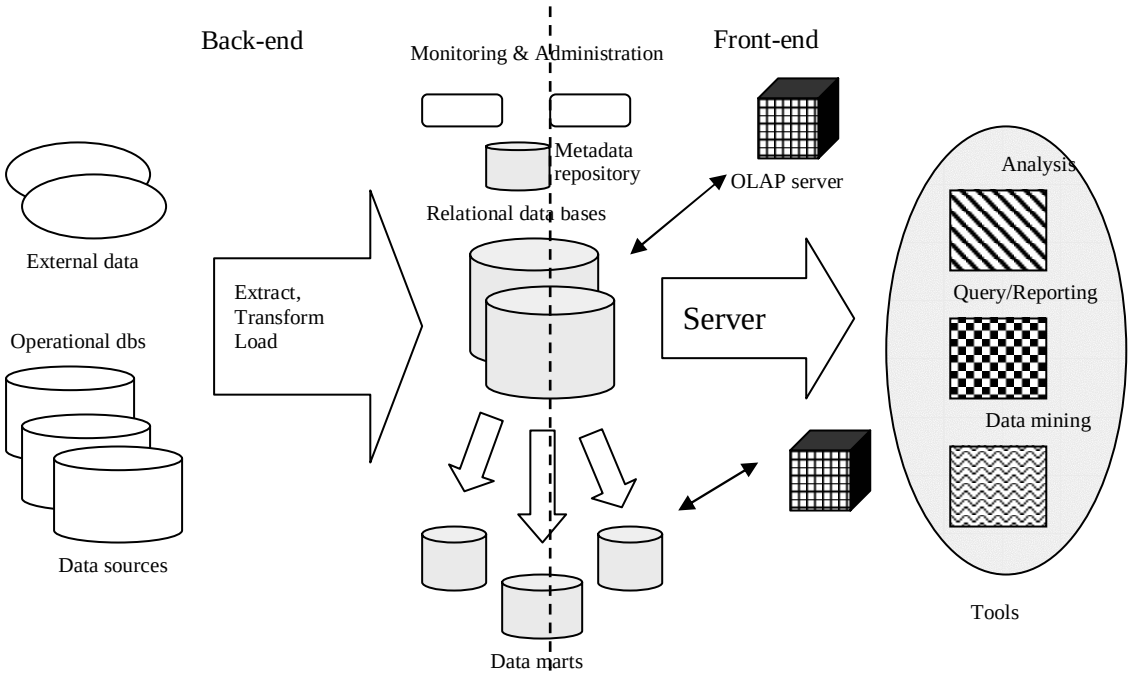

Figure 1. Example of an architecture of a Data Warehouse (Adopted from Chaudhuri & Dayal 1997, p.66).

2.3.1. External data

In this section external data will be defined and described.

Devlin (1997 p.135.) defines external data as *"business data (and its associated meta data), originating from one business, that may be used as part of either the operational or the informational processes of another business"*

External data is according to Inmon et al. (2001), data originating outside a corporation's internal systems. External data is acquired by corporations in different ways. It is characteristically either purchased from or created by, another corporation. Nowadays, the Internet is becoming a more frequently used source for the gathering of external data. Kelly (1997), claims the benefits from analysing external data to be the information about the external environment. This enables the decision makers of the corporations to recognise opportunities, detect threats and to identify synergies in order to respond in time to these events.

External data can according to Inmon et al. (2001), be of very different types and volumes and there are as many types of external data as internal data. The data can

either be detailed or summarised, structured or unstructured. Additionally, external data can according to Strand (2003), be extracted from different sources. The sources could for example be: industry organisations, county councils and municipalities, the Internet, business partners and bi-product data suppliers. The difference in the sources is the quality of data in the aspect of how much the data has to be cleansed before incorporated in the corporation. Kimball (1996) states that syndicated data (data that is bought from syndicate data suppliers such as SCB or ICI) could be of two conditions, i.e. adjusted or unadjusted. These will be further described later in this section.

The external data is different to the internal data in some ways. According to Inmon et al. (2001), the main difference between external and internal data is that internal data can be manipulated. Internal data is relatively flexible and can be changed by adjusting or changing the programs that extract and transforms the data. External data is according to Inmon et al. (2001), not possible to change in that way because the sources of external data are situated outside the corporations. The only option for persons in charge of the acquiring process for external data is to either use the external data as it is or to reject it. There is though according to Kimball (1996), a way to affect the structure of external data. By establishing a contract between a data supplier and a corporation, deciding how and what condition and structure the syndicated data will be delivered in, there is a way to facilitate the integration of external data.

There is however according to Inmon et al. (2001), one exception that enables the integration of unadjusted external data. When external data enters a DW there is a way of modifying the key structure. This is usually performed when matching the external data to the existing key structure of the internal data. The external data usually have a key structure that differs from the internal data and this demands modifications of the external data key structure if the external data should be used in an effective way. The change of the key structure of external data could either be a rather uncomplicated or a complicated task to perform depending on what the situation demands. Inmon et al. (2001) claims that a simple way of changing the key structure of external data is to let the external key go through a simple algorithm to convert it into the key of the internal data.

Kimball (1996) states there is another way of incorporating unadjusted external data. The corporation ought to conform the dimensions of a star, snowflake or starflake schema in order to integrate the external data with the internal data. This is performed by different types of transformations forcing the two data sources to share identical dimensions.

Approaches more demanding then the above mentioned are for example, the usage of algorithms in combination with reference tables or the conversion of external keys manually. The approaches where the external keys are manually changed are not suitable for large amounts of data or when the manual adaptation must be done repetitively.

An issue concerning the incorporation of external data is according Strand, Wrangler and Olsson (2003), the poor quality of external data received from data suppliers. Poor quality of external data was in a study, performed by these authors, mentioned as one of the main issues not to incorporate external data. The study additionally presents the issue of the difficulties in mapping internal keys with external keys as another substantial problem when to incorporate external data.

2.3.2. Extraction, transformation and loading layer (ETL)

The ETL layer is also known according to Marco (2000), as the acquisition layer. The data acquisition layer is according to White (1997), where the development and the execution of the data acquisition applications take place. These acquisition applications (ETL-tools) are used for capturing data from different source systems and further on transforming and loading it into a DW. The development of the data acquisition applications are based on rules that are defined by the DW developer. In the rules, the sources from which the DW's data will be acquired are defined. Other definitions dealt with are the data cleansing and the enhancement done to the data before it is applied into DW databases. This component includes the extraction, transformation and loading of the data. These components will be given a more thorough description in Section 2.4. Further on, the acquisition component generates definitions of the data enhanced which is known as meta data and stores it in the information directory component also known as the meta data repository. Meta data will be presented in the next section.

2.3.2.1 ETL-tools

ETL-tools can according to Gleason (1997), be purchased or developed by the corporation itself. To develop ETL-tools in-house is a time consuming and therefore costly process as with all development of software. First, the requirements must be established, then the effort of programming takes place and finally the software must be tested and maintained. As there is changes in the sources from which the data is extracted there could according to the author, be changes needed for the software as well. The biggest disadvantage with using self developed ETL-tools is however, that they usually do not generate meta data. The meta data supplies information about what transformations have been done to the data. Commercial ETL-tools automatically generates meta data. This is according to Gleason (1997), the biggest advantage with using commercial ETL-tools. Another advantage with purchasing commercial ETL-tools is the elimination of the costly developing process involved when a corporation is developing an ETL-tool on their own. Commercial ETL-tools are however, very expensive and according to Kimball (1998), commercial ETL-tools are not always cost effective for smaller corporations.

2.3.3. Meta data repository and meta data

Meta data repository is according to Chaudhuri and Dayal (1997) and Marco (2000), used to store and manage the different kinds of meta data that is related to a DW, i.e. the physical database tables that contain the meta data. Meta data is known according to Marco (2000) and Sperley (1999), as data of data. Meta data can be described by using following example; in a library a central card catalogue can be used for identifying what books are available in the library and also the physical location of the books. It also contains information of the subject area, author, number of pages, publication date and revision history of each book. The catalogue can be searched to determine which book will satisfy specific needs. Without the central card catalogue, finding a book would be a time consuming and difficult task to perform. Meta data is according to Marco (2000), the central card catalogue in a DW. Meta data helps users locate the information they need for analysis by defining the contents of a DW. There are according Marco (2000), two different kinds of meta data: business and technical meta data. Business meta data supports the end-users of a corporation. Technical meta data is used by the technical users and the IT staff provide them with information about their operational and decision-support systems. This information supplied by

the technical meta data can be used by the IT staff in order to maintain and expand the systems if required.

When data is brought into organisations from an outside source this data is according to Inmon et al. (2001), referred to as external data. (external data will be further presented in Section 2.2.). There is a great need for corporations to capture the meta data describing the incoming data. Important knowledge about the external data is according to Inmon (1996) and Marco (2000), document ID, the source of the external data, date of entry into the system, index words, purge date, physical location reference, length of external data and the classification of the external data. Inmon (1996) additionally claims that the meta data is used to register, access and control the external data in a DW environment.

When data is extracted from different source systems, internal and external, it could be cleansed, transformed and loaded into a DW using commercial ETL-tools. These tools generate according to Marco (2000), meta data about what has been done to the data in the ETL process. An example of generated meta data could be: data transformation rules describing changes of the data in the transformation process.

2.3.4. Data manager component, relational database and online analytical processing

The data manager component is where data, extracted from different data sources and worked on with back-end tools in the Extraction, Transformation and Loading process (ETL) is stored. (Inmon (1996), Marco (2000) and White (1997).

Data in a DW is according to Chaudhuri and Dayal (1997), usually stored in a star schema, a snowflake schema or a starflake schema to represent the multidimensional data model. Inmon et al. (2001) claims that a star schema arranges data in a way that makes it easier to visualise and navigate. Figure 2 and 3 shows the star schema and snowflake schema. A star schema consists of one or more fact tables and a number of dimension tables. A fact table is according to Devlin (1997), made up by the basic transaction-level information that is relevant for a particular application. The fact table often contains several millions, predominantly numeric rows Devlin (1997), Connolly and Begg (2002) and Inmon et al. (2001) claims. Dimension tables contain the data needed to position transactions along a certain dimension. These are, in contrast to the fact tables, relatively small and usually contain descriptive information. Dimension attributes are according to Connolly and Begg (2002), often used as constraints in DW queries. Examples of dimensions in a marketing application could be time period, marketing region and product type. A snow flake schema is according to Connolly and Begg (2002), a variant of the star schema where the dimension tables are allowed to have dimensions. A starflake schema is a mixture of the star schema and the snowflake schema. This is according to Connolly and Begg (2002), the most suitable database schema for a DW because it combines the advantages of the two other. There are different opinions among the authors on how to integrate data into a DW.

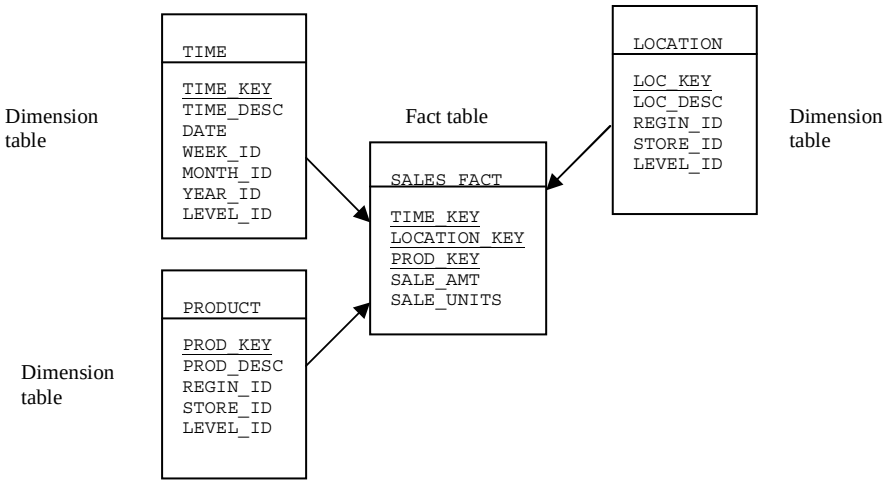

Figure 2. Example of a star schema. (Adopted from Bischoff, 1997, p.197).

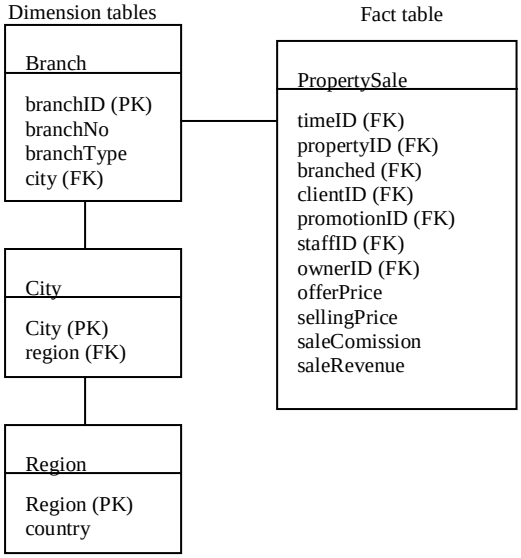

Figure 3. Example of a snowflake schema (Adopted from Connolly and Begg, 2002, p.1081).

On-line analytical processing (OLAP)

The data in a DW is usually modelled multidimensional to facilitate complex analysis and visualisation. An OLAP database is according to Kimball (1998), similar to a multidimensional database. The data stored in the OLAP server, i.e. the OLAP database, is usually stored in a rolled up form in distinction to the DW and the data is furthermore according to Chaudhuri and Dayal (1997), stored in arrays. The OLAP-tools are then used for performing the different analysis operations on the server. Figure 4 presents a cube of multidimensional data which is how the data is stored in the OLAP servers.

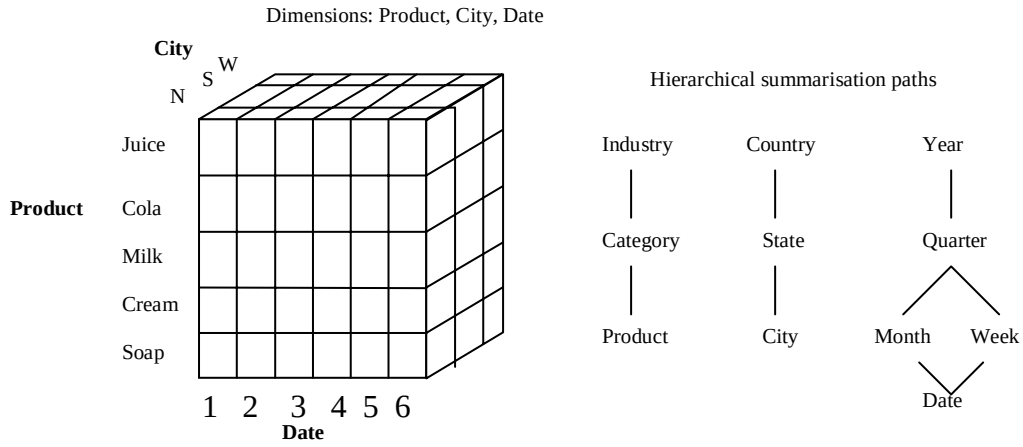

Figure 4. Example of multidimensional data. (Adopted from Chaudhuri and Dayal, 1997, p.68).

OLAP-tools provides for example according to Chaudhuri and Dayal (1997), the ability to do rollup operations (increasing the level of aggregation), drill-down operations (decreasing the aggregation), slice and dice operations (selection and projection) and pivot operations (re-orientating the multidimensional view of data). OLAP-tools are according to Mancuso and Moreno (2002), used by end-users to access the information stored in the OLAP servers for trends and exceptions. OLAP-tools complement the other reporting and analytical tools that can be found in a DW in an important way. OLAP-tools are according to Mancuso and Moreno (2002), one of the most powerful components in a DW. Users equipped with OLAP-tools can perform detailed data analysis and retrieve transactional details when something in the business operations is irregular. Further more, Mancuso and Moreno (2002) claims that when users are able to select analysis tools free of choice and no longer have to understand where to look for the data, users become more self-sufficient and are able to retrieve the information they need when they need it.

2.3.5. Data marts

A data mart can according to Chaudhuri and Dayal (1997), be described as a miniature of a DW containing less data, fewer subject areas and less history. The data in a DW is stored to support the whole corporation and it is not optimised for a specific department. A data mart is according to Inmon (1999b), Inmon et al. (2001) and Sperley (1999), a structure of different subject areas specific organised for departments individual needs. There can be several different data marts in one corporation. Each of the different departments in an corporation such as finance, marketing and sales can have their own data mart. The data marts are based on the different needs of the department using it, which leads to a different substance for each one of the different data marts. The content of data marts are usually summarised and designed for the specific needs of different departments and only contains data relevant for the specific department using it. It exists two different categories of data marts: dependent and independent. A dependent data mart uses a DW as a source while the independent data mart uses operational application environment as a source. This means: the content of the dependent data marts have the same structure and architecture all over the corporation while the structure and architecture of

independent data marts could be very different compared to each other. (Chaudhuri and Dayal, 1997; Connolly and Begg, 2002; Inmon, 1999b; Marco, 2000).

2.4. ETL in details

The Extraction, Transforming and Loading (ETL) process is where data is extracted from different sources, transformed into a predefined structure and thereafter loaded into a DW. These different phases will be presented in following Sections: 2.4.1, 2.4.2 and 2.4.3. The ETL process is also mentioned in the literature as the data replication. The two phases, extraction and loading, are also known as capture and applying. The ETL process described in following sections is in the literature intended for internal data but as there was no literature to be found intending the ETL process for external data, this description is believed to be used for the external data as well. The researcher has searched for information about the ETL process concerning external data in scientific reports, in literature and the researcher has also consulted the supervisor to try to find additional sources of information but the information found was not comprehensive.

2.4.1. Extraction

Extraction is according to Inmon et al. (2001), the process of selecting data from one environment and transporting it to another environment. While most corporations have several different sources there is a problem to capture the source data. The operational data can be stored in sequential and indexed files or in hierarchical, relational and other databases. Besides this there are different formats of physical storage and data representation to take into consideration. There are according to Devlin (1997), two main techniques when extracting data: static capture and incremental capture. Static capture is defined by Devlin (1997 p.179) as *"A method of capturing a time independent or static view of the data in a source data set."* The incremental capture is defined by Devlin (1997 p.179) as *"A method of capturing a record of the changes that take place in a source dataset."*

The static capture is not as common as the more complex incremental capture but could take place for example, the first time a data set from a specific operational system is to be added to a DW. Another occasion when static capture occurs is when a complete history of an operational system data is maintained and the volume of the data set is small. The static capture technique could be said to take a snapshot of the source data at a special point in time. The static capture does not, compared to the incremental capture, take time dependency of the source into consideration. For databases changing over a period of time there is a need to capture the history of these changes. For the issue of capturing the actual changes that have occurred in the source, the incremental capturing type is used.

There are several different ways of representing time dependency in the operational systems and this have developed a number of different ways to perform incremental captures. These different methods or techniques as they according to Devlin (1997) also can be called, can be divided into two types: immediate capture and delayed capture. The techniques related to immediate capture are: static capture, application-assisted capture, triggered capture and log capture. The techniques belonging to the delayed capture are: timestamp-based capture and file comparison. From these teqniques the different output data structure originates, described in Section 2.4.1.3.

2.4.1.1. Immediate capture

The immediate capture technique according to Devlin (1997), captures the changes at the time they occur. This ensures a complete record of changes in transient, semi-periodic and periodic data. The four included techniques are following:

Static capture

The static capture technique's functionality could be described as taking a snapshot of the source data. This could be done using one of the three sub setting dimensions: entity dimension, attribute dimension and occurrence dimension. The data to be captured is usually selected using SQL. Figure 5 presents examples of the different sub setting dimensions.

1. Example of an entity dimension

```
SELECT      *
FROM        CUSTOMER_FILE
```

2. Example of an attribute dimension

```
SELECT      NAME, PHONE_NO, CITY
FROM        CUSTOMER_FILE
```

3. Example of an occurrence dimension

```
SELECT      *
FROM        CUSTOMER_FILE
WHERE       CITY = 'Paris'
```

Figure 5. Examples of sub setting data for capture (Adopted from Devlin, 1997, p.183).

These sub setting techniques can according to Devlin (1997), also be used in the five incremental capture techniques presented below.

Application assisted capture

Application assisted capture is a technique that according to Devlin (1997), is built into the application with the primary task to maintain the operational data. The source application preserves its data for immediate capture of incremental changes in that source. This is done by providing a record of continuous, non-volatile changed data. This technique is difficult to apply to existing systems due to the fact that many operational systems are poorly designed and documented and that the operational systems usually are of a complex nature. This technique is however suitable to apply when designing new operational systems.

Triggered capture

The trigger capture technique could according to Devlin (1997), only be of use if the operational systems are storing the operational data in a database instead of in a file. Triggered capture could be described according to Connolly and Begg (2001) and

Devlin (1997), as the capture of operational data performed by triggers for all database events for which changed data should be captured. The events or the changed records are stored in a separate place for future processing. Since each database-update-transaction that occurs on the operational data triggers a second transaction to capture the change, this leads to a decreased performance of the database. Therefore this technique is only suitable for records where small amounts of changes or events are taking place.

Log capture

When changes occur in the operational data these are according to Connolly and Begg (2001) and Devlin (1997), typically maintained in a log for backup and recovery purposes. This makes it possible to use the log as the source from which changes are captured using the log capture technique.

2.4.1.2. Delayed capture

The delayed capture techniques capture the changes at specified times. This generates a complete record of changes only in periodic data. Delayed incremental capture is according to Devlin (1997), useful for transient data sources only if the business does not need to see all changes occurring in the legacy system mirrored in a DW.

Timestamp based capture

Timestamp based capture is according to Devlin (1997), a technique that is dependent on timestamps in the source data. The source data has to contain at least one field of time stamped information. This field is used as the basis for the record selection. The records to be captured are those that have changed since the last capture. The relevant records have a timestamp later than the time of the prior capture.

File comparison

File comparison is according to Devlin (1997), a technique used for capturing a possible incomplete set of changed data by comparing two versions of the operational data. This is done by comparing for example, a file from yesterday with a file from today and then capture the changes between the two files.

The static capture this is required to be performed at least one time in order to initially populate the DW. To accomplish the feature of history in the DW, one or more of the incremental capture techniques has to be performed. The different techniques of data capture have different strengths and weaknesses and none is therefore an obvious winner. One factor mentioned earlier is the different source systems of corporations where data is represented in different ways and operational data is stored in different environments. The consequence of this is that according to Devlin (1997), a mixture of the different techniques are used combined.

2.4.1.3. Different output structures

The data captured in the capture process, the output data should according to Devlin (1997), be structured and stored in a format that can be easily used by the other phases, transformation and loading, in the ETL process. This feature together with the issue of the documentation of the data content and its characteristics are the two key requirements of the output data. The output data can be structured in different ways depending on whether it is generated on a record level or on a file level. If the output data is generated on a record level, the output data is able to be uploaded from the operational systems to the DW in small batches and even on a record by record basis.

One setback to this technique is however, that the upload needs more thorough controls to ensure that the full data set is uploaded and that no records are out of sequence or missing. The features of data generated from a file level were not mentioned in the literature and is therefore not further described.

If the capture component operates independently, i.e. it is not part of an ETL-tool performing the whole process, the output data can be of self-documenting format. When output data is structured in a self-documenting format, the output data according to Devlin (1997), automatically documents the characteristics and content definitions of the data, i.e. meta data. The meta data of the output data is very important for DWs administrators since it provides them with the information of the source of the data as mentioned in Section 2.3.3.

2.4.2. Transformation

There are according to Gleason (1997), different kinds of transformations where each one of the transformations has its own different characteristic and scenario. One of the problems with data sourced from operational systems, this also includes external data, is that not all data is subjected to the same business rules. Theses differences must be dealt with as the new data is created.

Gleason (1997) defines some of the basic types of transformation as:

Simple transformation

Simple transformation is the basics of all data transformation. In this category the focus is on the manipulation of one field at a time. There is no concern given to the values in a specific field. Examples of simple transformation are: changing the data type of a field or replacing an encoded field value with a decoded.

Cleansing and scrubbing

Cleansing, also known as scrubbing, is the category where the data is formatted in a predefined way and the use of every field is assured to be consistent. One example of when the formatting of fields can be used is the formatting of address information.

Cleansing also verifies that the fields are containing valid values.

Aggregation and summarisation

Aggregation and summarisation is the category where several records of data from the operational environment are combined into less records and stored into a DW environment. Aggregation is sometimes performed in a way that leaves stored data in a DW less detailed than the data in the operational environment. Another case when aggregation is executed is when data marts are created. Data marts contain aggregated or summarised forms of detailed data from the DW.

Integration

According to Inmon (1996), a DW can be described as a system that supports business analysis and decision making by the construction of a database containing data that is subject-orientated, time-variant, non-volatile and integrated. The data is according to Singh (1998), sourced from multiple, often incompatible systems. Before data is loaded into a DW, the data must be changed to fit the key structures of the data already stored in the DW or if the data is being loaded for the first time, it is loaded according to the key structure of the data model. (Singh 1998). When the data has been extracted from the operational systems it is loaded into the transformation and

integration layer. This layer is according to Inmon et al. (2001), mostly made up of applications and this is where the alteration of the data takes place.

The integration is according to Gleason (1997), the process where the data from multiple sources are mapped, field by field, to finally end up as new entities which are loaded into the DW's data structure. Gleason further states that the integration is preceded by different types of transformations in the comprehensive transformation process. In the transformation process the structure of the data and the data values are controlled and when needed, changed to the standard of the DW. In the next section, different techniques for transformation will be described. The feature of integration will be further described in Section 2.4.2.1.

Different techniques for transformation

To be able to match the data extracted from the different sources and to integrate it into one entity according to a corporation's data model and thereby adjust it to the business rules, there are according to Gleason, (1997) and Inmon et al. (2001), several types of transformation that could take place. According to Gleason (1997), the different types of transformation could be divided into two categories: simple transformation and complex transformation.

The different kinds of simple transformation that can take place in the simple-field mapping are according to Gleason (1997):

Data type conversions

Data type conversions are the most common of the simple transformations techniques. Gleason (1997) advocates reformatting the operational data to fit the data model of the DW where the data should be consistent. The data in the operational system makes perhaps sense within the context of its originate environment but not at the enterprise level, which the DW is supporting.

Data Source		Data Warehouse
AMT-BILL DECIMAL (7,2)	→	AMT-BILL DECIMAL(13,2)
D-SALES-AMOUNT DECIMAL (13,2)	→	AMT-SALES DECIMAL (13,2)
ACCOUNT-STAT-CODE CHAR(1)NULLS ALLOWED	→	CD-ACCT-STATUS CHAR(1) NO NULLS

Figure 6. Example of data type conversions (Adopted from Gleason, 1997, p.163).

This transformation can be done using for example DW data transformation tools.

Date/time format conversions/simple reformatting

The operational systems are according to Gleason (1997) usually not designed or constructed in the same way or at the same time which often results in differences of what time and date format are used. This is a problem which is taken care of in the date/time format conversions or in simple field reformatting claims Inmon et al. (2001). The different types of date/time used in the operational environment must be transformed into standard DW format. This can be performed through manual program coding or by the use of the transformation tools.

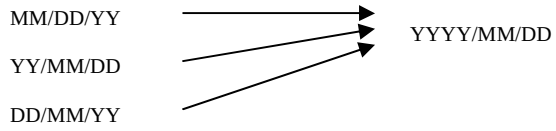

Figure 7. Example of how to reformat data to achieve consistency. (Adopted from Inmon et al. 2001, p.73).

Field decoding/ Encoding structures

In the operational environment the data stored in the operational databases is sometimes stored using coded fields. Male and female can according Gleason (1997) and Inmon et al (2001), sometimes be coded as MALE or FEMALE, M or F and sometimes as 1 or 0. This is perhaps not a problem for the users of the operational systems that know what the codes mean but it could be a serious problem for the end-users of the DW. That is why, according to Gleason (1997), encoded values in operational data and external data should be converted into coded values that easily can be understood before they are stored into a DW.

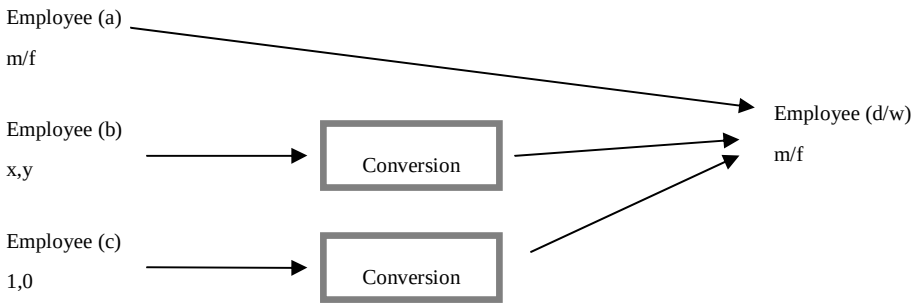

Figure 8. Example of a conversion of encoded values to achieve consistency. (Adopted from Inmon et al. 2001, p.73).

The different kinds of complex transformation are according to Gleason (1997), part of the cleansing and scrubbing process. This process examines the actual contents of fields or groups of fields instead of the storage format. Different types of complex transformation are further described below.

Valid values

There are according to Gleason (1997), different techniques for how to check for valid values in the systems. The easiest way to execute this is to do range checking. Range checking is mostly executed on fields containing numbers or dates. One example could be if a valid value for invoice numbers is between 1000 and 99999, and one invoice is found with the invoice number 777, this invoice is not within the predefined values and will therefore be excluded and investigated further. Another example is if a corporation, founded in year 2000, finds an invoice from 1999 in its system. This invoice would then be further investigated and excluded from the system.

There are according to Gleason (1997), other techniques for checking valid values, i.e. comparing data fields with other data fields or comparing data fields with an enumerated list. The technique where data fields are compared to other data fields is

called: dependency checking. An example of this technique could be the comparison of a purchase order number on one invoice with the purchase order number in the purchase system. Invoices containing purchase order numbers that do not match should be excluded and further investigated. The technique where the data fields are compared with an enumerated list could be exemplified as comparing the values in the data fields with predefined values like “PRO”, “AVE”, and “BEG”. Any value that does not match these values should be excluded.

Complex reformatting
This type of cleansing and scrubbing is used to convert data extracted from different sources with different formatting into one standardised representation. For example, when there is no standardised way in operational systems environment to store addresses this means that the addresses will be represented in many different ways. Complex reformatting uses parsing to determine the essential parts of the addresses. The shortened part “blvd” in the address would for example be converted to the standardised “boulevard”. Figure 9 shows an example of a full address reformatting.

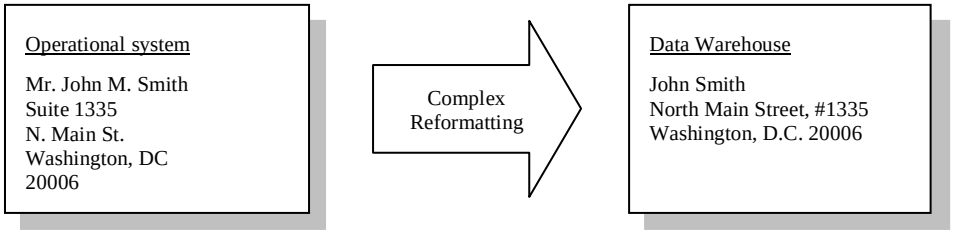

Figure 9. Example of complex reformatting (Adopted from Gleason 1997, p.166).

2.4.2.1 Integration

The most demanding task in combining operational data from different and multiple sources are according to Gleason (1997), to integrate it into one merged data model. The different data sets must be combined into a new entity. The different data sources do not usually have the same business rules, which prevent the data from easily being united. These differences must be accommodated as the new data is created.

Simple integration

Gleason (1997) claims that the most common integration technique is the simple field level mapping and the author further claims that as many as 80-90% of the integrations taking place in a conventional DW are simple field mappings. Simple field mappings are described as moving a data field from the source system and along the way scrub, reformat and perform some sort of simple transformation to the data as it is moved into the DW’s data structure.

Complex integration

10% to 20% of the integrations and the processes of moving data into a DW are more difficult and complex than to just move the data from the source field to the target field. There is a need for additional analysis of the source data to be able to transform it into target data. The most common techniques according to Gleason (1997), will be presented below.

Common identifier problem

One of the most difficult problems with integration is when different sources, with the same business entities, are to be integrated. The problem is to identify the entities that are identical. A customer for example that exists in several different systems can be identified by a unique key in each different system. This unique key is not always the same in the different systems and therefore knowing if it is the same customer in the different systems can be difficult. In a DW, there should only be one unique key to identify a customer. Solving this problem is usually done in a process consisting of two phases. The aim of the first phase is to isolate the entities that can be guaranteed to have a unique key and the purpose of the other phase is to merge these entities into one.

Multiple sources for the target element

Another situation that must be handled is when several different sources for a specific target data are existing. It is common that data from different sources do not correspond and from this the need to decide which data to use in the target data, arises. There are some different techniques to solve this problem. One way is to determine which system is the key system; i.e. the system that is predominant in case of a conflict. Another way is to compare the dates (to decide which one is most up to date) or to compare other related fields to decide which constituent is the dominant one. Additionally, the different values are occasionally summarised and from this sum the average, to be stored in the DW, is calculated.

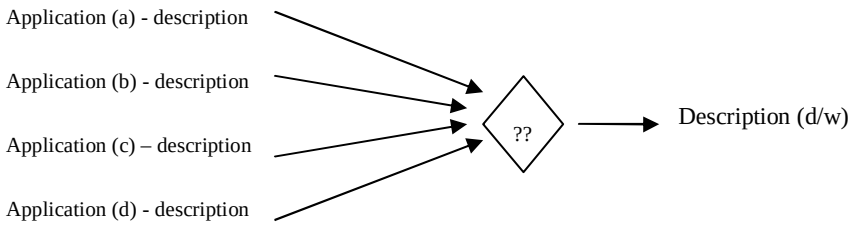

Figure 10. Example of a problem with multiple sources for the target element. (Adopted from Inmon et al. 2001, p.74).

Missing data problem

For different reasons which will not be further discussed in this work, there are according to Inmon et al. (2001), occasions when there are no data to be sourced to the target data. Under some circumstances, Gleason (1997) states, some DWs can manage to have nulls or blank fields stored when the data for a field is missing but for other DWs there must be a value stored in the field. The reason for the need of having a value stored is that if there are queries related to the table, these queries will not be valid if the field does not contain a valid value. As a result of this there is occasionally a need to create data to replace the missing data. This is usually done by obtaining and summarising the values in the adjacent fields above and below the target field and the average is put in as target data. In this way a curve based on the query would be smoothened instead of showing a huge gap. See Figure 11 for an example. Some corporations though use very advanced and complex techniques applying several variables resulting in a very realistic value.

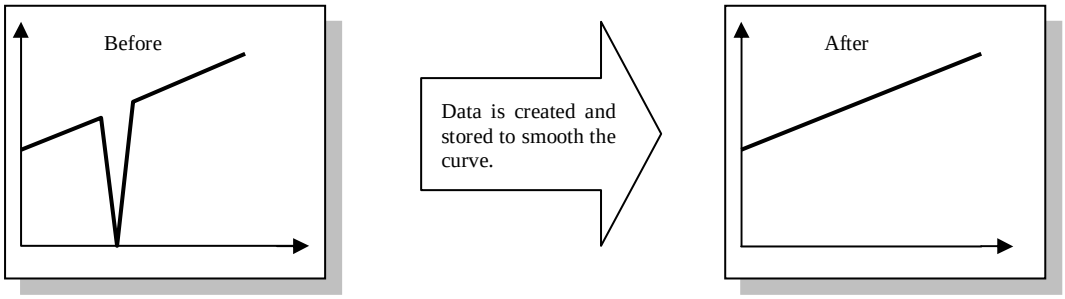

Figure 11. Example of a result from the solution of the missing data problem. (Based on an example of Gleason 1997, p.168).

2.4.3. Loading/ applying

When data has been extracted, cleansed and transformed it must be loaded into to the DW. This is according to Chaudhuri & Dayal (1997), usually performed by batch load utilities. The system administrator must have control over the loading of the DW and therefore the load utility used should enable the system administrator to be able to monitor the system, to cancel, suspend and resume the load process and to restart the load process if it failures with no loss of data integrity. Before the loading of the data takes place the data must be pre-processed. The pre-processing, often conducted by batch utilities, concerns according to Chaudhuri & Dayal (1997): checking integrity constraints, sorting, summarisation, aggregation and other computations to build the derived tables stored in the DW. It further concerns building indices and other access paths and partitioning to multiple target storage areas.

The loading of a DW must be well planned because of the amount of data involved. The load utilities for a DW deals for example with a much larger amount of data than loading utilities for an operational database. The DW is often loaded or refreshed at night because it needs according to Chaudhuri & Dayal (1997), to be taken offline to perform this.

There are according to Devlin (1997), different techniques of loading the data extracted and transformed to the target data store. These different techniques are: load, append, destructive merge and constructive merge.

Load

This is the most uncomplex technique of applying the source data set to the target data set. This technique loads or reloads the data that has been extracted and transformed into the DW completely replacing the existing target data.

Append

In this technique the loading technique appends the source data that has been extracted and transformed to the existing target data. The target data that already exists is saved either by duplicating the existing records or by rejecting the new data records. Which technique used is depending on the content of the captured data.

Destructive merge

This technique merges the data being uploaded to the existing target data. If the keys of the uploaded data match the existing records the content is updated and where the keys of the uploaded data do not match the keys of the target data, a new record is created.

Constructive merge

This technique always adds the uploaded records to the target data. Where the keys between the source data and the target data match, the old target data is replaced but not overwritten.

Which one of the different techniques that should be used depends according to Devlin (1997), on the time dependency of the target data. The constructive merge is according to the author, the technique most effective for maintaining periodic data since the records, using this technique, never are deleted.

3. Problem

In this chapter the aim and objectives of this work is given along with the delimitations of the work.

3.1. Problem area

In order for a corporation to be able to compete against the competitors in their business area, Marco (2000) claims, there is a need to be able to respond quickly to the market demands and customer needs. A DW is a decision support system many corporations use for this purpose. The DW is according to Devlin (1997), used to efficiently analyse and understand the performance of an corporation by enabling access to all the operational data in a structured way. However, according to Inmon (1999a) and Singh (1998), the internal data on its own is not sufficient for being able to understand the current situation of the corporation. There is also a need to acquire and integrate data originating from outside the corporation, i.e. external data, since such data allows the corporation to reveal information that would not been possible to compile from purely internal data (Inmon, 1999a; Singh, 1998).

To be able to acquire such external data and to integrate it into a predefined structure of a DW, the data must be handled according to given data structures. However, different types of data states different prerequisites on how to perform the acquisition and integration since the proposed usage of the data guides the activities performed. Therefore, there is no single approach that suites all needs. Instead, different data must be integrated according to its purpose and this raises the need for different types of integration approaches. Unfortunately, the literature covering external data acquisition and integration is rather vague on the characteristics of these approaches and the underlying problems. Therefore, this work is dedicated towards the creation of an increased understanding concerning these issues.

3.2. Aim and objectives

As a consequence of the above described problems, the following aim was formulated:

The aim of this work is to exploratory outline current approaches for acquiring and integrating external data into Data Warehouses and to give a brief overview of the future trends for external data integration.

The result of the work is supposed to increase the knowledge concerning the current standings and the future of external data acquisition and integration into DWs. The overall aim of this work will be satisfied by the fulfilment of the following objectives.

- Give an outline on how external data is currently acquired
- Give an outline of current external data integration approaches
- Identify the most common problems concerning the integration of external data into DWs.
- Identify future trends in integration of external data into DWs

The first objective, concerning the outline of how external data is acquired, will present information about how corporations acquire their external data, which distribution technologies are used and how the acquired data is transformed. This

objective was included since it gives the frame for how corporations manages the external data in the initial phases and it also pinpoints the importance of being aware of the purpose underlying the integration. The second objective, an outline of current external data integration approaches, will provide information about which different approaches are available and which one is most applied. The third objective about the most common problems concerning the integration of external data will describe the difficulties that are associated with the process of integrating external data. This objective was included in order to balance the descriptions and show on the fact that there are also issues that needs to be dealt with, before an unlimited acquisition and integration of external data is possible. This objective was also supposed to help in the identification of possible, future work. The fourth objective regarding future trends will present information about any new approaches that may be applied in the future. It was considered as important not only to investigate current state of the art, but also to scan the horizon for what the future holds, since it would also bring valuable insights for future work, but most importantly, it would give interested readers some hints on were the area is heading.

3.3. Delimitations

The DW technology is used in many different business areas and therefore this work will be restricted to the financial business area. Since approaches used for integration are the same throughout the different industry areas this work could have been accomplished in a different area but since there are several reports and case-studies from the financial business area and Data Warehousing, this area could be understood to be a more mature area. This area was further chosen in front of others because the researcher finds the financial business area more interesting and the research study conducted together with the co-researcher had an outlined aim concerning the financial business. By conducting the research in co-operation the same respondents could be used, this issue is further described in Section 4.3.

4. Research method

This chapter contains a presentation and a discussion about the method of how to conduct a research, suitable for this work. The method adopted is strongly influenced by the process (Figure 12.) by Berndtsson, Hansson, Olsson and Lundell (2002).

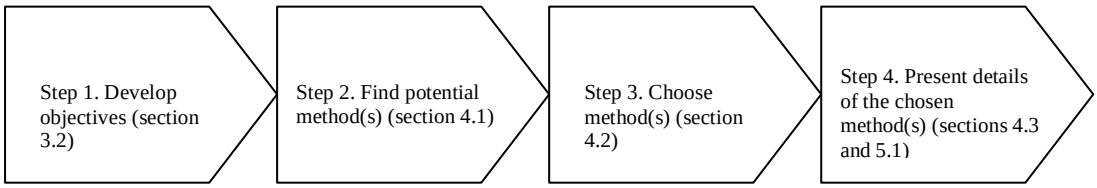

Figure 12. The four-step process presented by Berndtsson et al. (2002) The Figure gives additionally an indication about how the process is connected to the different sections.

4.1. Methodical considerations

In this section the reader will be given the motives for the method chosen to reach the aim stated in Section 3.2. The discussion is based on the process given by Berndtsson et al. (2002) introduced in Figure 12.

In order to achieve the overall aim of the research conducted, Berndtsson et al. (2002) claims that a number of different objectives (the objectives of this work are described in Section 3.2.) should be formulated; this is the first step of the process. An object is a minor, realistic and measurable unit and all objects should at the end lead to the answer of the aim of the work. To accomplish each one of these objectives developed, different methods of how to conduct a research could be used. The different objectives developed in this work are, as mentioned in Section 3.2., the outlining on how external data is currently acquired, the outlining of current external data integration approaches, to identify the most common problems concerning the integration of external data into DWs and finally, to identify future trends in integration of external data into DWs. The second step is to find a potential method(s) to conduct the research and the third step is to choose the most suitable method for the research. In this work only one method will be used for investigating the different objectives. The reason for this will be given in a later section. These three steps are further developed in Section 4.1. Step four outlines the details of the method chosen and will be introduced in Section 4.2. Finally, the research process will be given in detail in Section 4.3.

There are according to Andersen (1998), Berndtsson et al. (2002), Patel and Davidsson (1994) and Svenning (2000) two categories of methods of how to conduct a research: quantitative and qualitative. Quantitative research methods are according to Berndtsson et al. (2002), used to explain a specific substantive area. The research is usually driven by formulating a hypothesis and then testing it rigorously. The testing is usually performed conducting repeatable experiments. Qualitative research methods are, in distinction to the quantitatives, used when an increasing understanding of a substantive area is preferred. In this work a deeper knowledge is intended to be acquired about the aim given in Section 3.2. and therefore the method of how to conduct a research suitable for this is according to the authors, of a qualitative nature. The knowledge acquired in a qualitative research is according to Patel and Davidsson (1994) claimed to be of a deeper nature than the knowledge acquired in quantitative

research which often results in incoherent knowledge. This work aims at getting an increased knowledge of an area in which not much has been written. To get a wider knowledge, several different sources have been used, i.e. empirical data and literature. In this way the information gathered from the different sources is intended to give more descriptive information about the current state of external data integration approaches. There are according to Berndtsson et al. (2002), numerous methods/material collecting techniques that could be used, i.e. literature analysis, interview studies, case studies and surveys/questionnaires. There is no consensus among different authors whether for example, an interview study or a case study is a method or a material collecting technique. The terms used in this work for gathering material are therefore called method/material collecting technique.

As mentioned earlier, the aim of this work is to increase the knowledge of an area that is not well described in the literature. Therefore a literature study can not be performed due to lack of material. This leaves the gathering of information to be performed in two different ways. One way is to perform a case study as an in-depth exploration where a phenomenon is examined in detail. Since there is not much material describing the relevant problem of this work this method could have been used according to Berndtsson et al. (2002). Case studies make it possible to understand and explain a phenomenon which is not yet well understood. But, since the use of the case study method gives a deep understanding of a specific case and the aim of this work is to outline the state of several different cases, this method was not the most suitable method/material collecting technique for this work. Instead of a deep understanding about a specific case this work is interested in receiving more general answers while it enables the researcher to generalise the result. The other way of performing this work is to gather information from a larger amount of respondents to get an overview of their course of action in the matter, described earlier in Section 3.2.

For the purpose of gathering information from different respondents there are two specific methods/material collecting techniques that could be used in an exemplary way. Those two methods/material collecting techniques are: interview studies and surveys/questionnaires. Surveys/questionnaires are however, according to Berndtsson et al. (2002), usually used for exploring a well known phenomenon for which it exists a large amount of respondents having a relevant knowledge of the issue. The subject area of this work is not well known and there are therefore not a large amount of respondents having the relevant information. Another feature of the questionnaires is according to Andersen (1998), that they are not suitable for this work because the questions in a questionnaire should be structured and must be asked in a strict order. A researcher using a questionnaire is limited to only ask the preformulated questions and can not ask follow up questions. This leaves an interview study as the method/material collecting technique most suitable for this work.

4.2. Interview study

There are several different types of interviews. The interviews can be open, structured or a hybrid of those, i.e. semi structured. The interview technique chosen have both advantages and disadvantages. Open interviews are according to Berndtsson et al. (2002), usually chosen when performing a qualitative research. Since the method of how to conduct research of this work, as mentioned earlier, is of a qualitative nature this interview technique could be suitable. Open interviews are further used when the researcher has little knowledge of what specific issues will be discussed during the

interviews even though the general subject area of the interview is known. The interview questions asked by the researcher should be formulated in a way that makes it possible for the respondents to answer the questions using his/her own words. The advantage of using open interviews is according to Berndtsson et al. (2002), that if the interviews are properly conducted and performed by an experienced interviewer, the issues of real importance will be addressed. Disadvantages of open interviews are the difficulties for an inexperienced interviewer to find the right balance between open questions and the more important exploring questions and furthermore that it can be more difficult for the researcher to take notes during open interviews. The researcher of this work is not very experienced in conducting interviews and therefore a somewhat more structured interview form would be suitable. There is one more issue that must be taken into consideration when performing an open interview. This is according to Berndtsson et al. (2002), the question regarding the researcher's bias, i.e. the preconceptions, views and values that will influence the researcher while undertaking the research. This issue does not only concern the open interviews but all different types of interviews. In this work the researchers' bias is affected by the literature read in order to gather background information about the research area and the previously knowledge acquired in the subject.

Closed interviews are according to Berndtsson et al. (2002), interviews based on questions to which there are simple answers, i.e. questions that only return short answers or perhaps even an yes or a no. These are not the most suitable answers for this work since it is of a qualitative nature and therefore needs more descriptive information. The closed interviews are further on characterised by its point of view regarding questions asked during the interviews, i.e. it does not allow the researcher to add or remove questions depending on the answers received from the respondent. The question form can not be altered. For this work the possibility to add and remove questions and to post follow up questions is quite important while it can provide the researcher with more knowledge about the issues concerned. The advantage of the closed interviews is the possibility to repeat them several times, knowing that the same sets of questions always are put forward. This could be done as well when conducting more open interviews by writing down the questions that are important as reminders, i.e. semi structured interviews.

Semi structured interviews are according to Andersen (1998), used in a similar way as the open interviews. Semi structured interviews are common when the researcher has more knowledge of the phenomenon relevant for the interview. This technique is chosen in front of the other two because it gives the researcher the possibility to post follow up questions if necessary. The questions asked by the researcher could be asked in an unstructured way, i.e. the researcher does not have to follow a special order. Another important feature why the semi structured interview technique is chosen is according to Berndtsson et al. (2002), that the preformulated questions support an inexperienced researcher.

Despite the level of structure, interviews can be conducted in different ways. In this work the interviews are intended to be performed by telephone. Telephone interviews are chosen because the respondents are geographically scattered and the cost for conducting the interviews in person would be high. By conducting the interviews by telephone the respondents are according to Andersen (1998), more likely to take part in the research due to the saving of time compared to personal interviews. Disadvantages with telephone interviews are according to Andersen (1998), that the respondents can be unwilling to perform the interviews due to the fact that the respondents can not see the researcher and therefore be suspicious against him/her.

This could be avoided by sending the respondents an accompanying letter describing what the interviews will contain and presenting the general aim of the interviews. Another drawback regarding telephone interviews is that the researcher is not able to register the facial expressions or the body language of the respondents. Despite these drawbacks, telephone interviews were considered to be the most suitable method/material collecting technique.

4.3. Research process outlined

In this section the research process will be described in detail regarding the way it is intended to be used. The different steps of the process will be described and motivated to the reader. By presenting the research process in detail (Figure 13), the reader is enabled to get a comprehensive understanding on how the process is intended to be used without having to read the whole section.

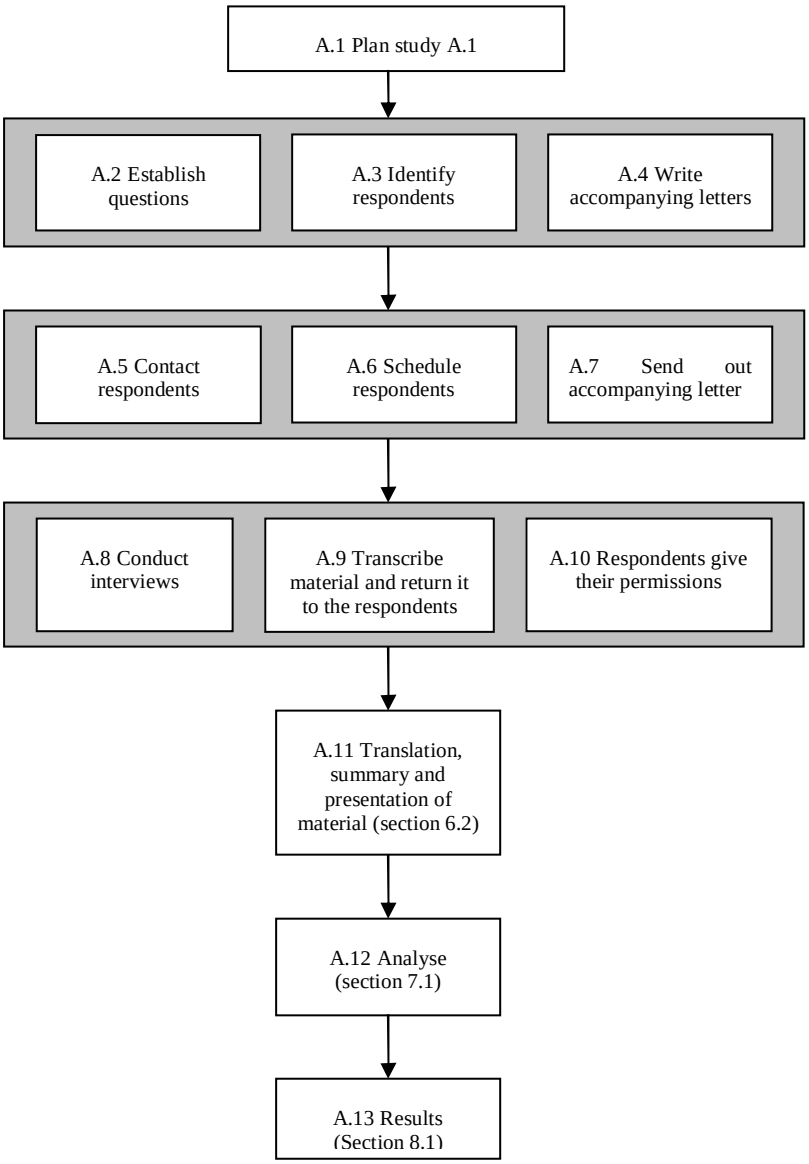

Figure 13. The research process in detail. The implementation of the research process will be further described in Section 5.1

When conducting a research with a specific method/material collecting technique, in this case semi structured interviews conducted by telephone, a plan makes it easier for the researcher to perform the different steps of the process and to accomplish the research in a more structured way. By performing a structured research study the possibility for a better research results is increased.

This work was closely related to another work at the Department of Computer Science at the Univerity of Skövde (Niklasson, 2003). This fact and the initial attempts to find respondents made the researcher aware of the shortage of information in this specific research area. The therefore insufficient numbers of respondents led to an agreement between the researchers of the two projects. They united the respondents found, co-operated during the interview phase and united the research material. The respondents were contacted once and combined interviews were performed. In this way the respondents time were more efficiently used and perhaps therefore, their positive attitude towards acting as a respondent were increased.

A.1 Plan study

The planning of the a research process should cover all steps that are supposed to be included in the process, i.e. issues concerning the interviews, the respondents, the questions, the collected material, the presentation of the material and finally the analysis of the material and the introduction of the results. The detailed plan is presented in Figure 13.

A.2 Establish questions /A.3 Identifying respondents /A.4 Write an accompanying letter

The different steps of the research process were planned to be performed simultaneously. The establishment of questions was thought of as a process in which a brainstorming of questions would lead to a foundation of questions to refine. These questions would be discussed with the supervisor and then, if needed, the questions would be changed. Support of how to establish questions would also be found in literature. The question form would be divided into different parts to facilitate for the researcher to ask the questions relevant for their work and especially to make it easier to divide the material later for the analysis.

The initial question concerning the respondents was how many were needed to get a broad enough base of material to analyse. By performing approximately ten interviews the material that would be gathered was considered to be sufficient to analyse and thereby have enough material to reach the aim stated in Section 3.2. Identifying suitable respondents would be accomplished by searching the Internet and the telephone directory for information about corporations within the financial industry. The aim was, as mentioned, to find ten respondents who had knowledge of both DW and external data.

The accompanying letter with information about the aim of the research work and an introduction of the researchers, together with a presentation of the important definitions used in the two different works would be established and discussed with the supervisor.

A.5 Contact respondents /A.6 Schedule respondents /A.7 Send out accompanying letters

The person responsible for the IT departments of the corporations found would be contacted by the researchers by telephone for further information about who would be suitable for answering questions about the topic. When the person who was supposed

to have the information asked for was to be contacted, a brief presentation of the work should be given and the person would thereafter be asked to take part in the interview.

If there was a positive answer from the person, the interview would be scheduled and further information about the work and the interview should be given. The address of the respondent should then be asked for, enabling the researchers to send the accompanying letter together with the questions. The questions should be sent in advance, enabling the respondent to prepare himself/herself for the interview.

A.8 Conduct interviews/ A.9 Transcribe material and return it to the respondent. / A.10 Respondents give their permission.

While conducting the interviews, the interviews would be recorded, making it possible for the researchers to later transcribe the material. The interviews would also be performed on a loudspeaker telephone enabling both researchers to listen. The interviews should be conducted in Swedish, facilitating for both the researchers and the respondents. This would perhaps give better and more detailed answers to the questions.

Transcribing the interviews would be done immediately after an interview has taken place in order to be able to send the transcribed material back to the respondent as soon as possible for approval. If a respondent would not give his/her permission to the researcher to use the material, it would be deleted and not further used in this or any other work.

When the material had been approved by the respondent the next step could be commenced.

A.11 Translation, summary and presentation of material.

When a respondent had given their permission to the researchers to use the material, it would be summarised and thereafter translated into English. One question at a time would be summarised and after that the material would be presented to the reader. The questions and answers would be divided between the researchers depending on who was in charge for the respective questions and the adjacent answers. If an answer contained information that both researchers needed, this will be described in the material presentation. The material will be presented to the reader in Section 6.2

A.12 Analyse

The material gathered during the interviews will be analysed and introduced in Section 7.

A.13 Results

The results from the analyse will be presented to the reader in Section 8.

5. Conducting the research

In this section the implementation of the outlined research process in Section 4.3 will be presented. The actual realisation sometimes differs from the plan and the reason for this will described in later sections. The interviews will later be presented together with the researcher's reflections and experience from conducting the interviews. The reader is referred to Figure 13 in Section 4.3 for an overview of the process. The conducting of the research process will only be described until phase A11. The analyse and results from remaining phases will be introduced in Section 7. and 8.

5.1. Implementation of the research process

A.1 The planning

A.1 The research process started, as shown in Figure 13, by planning the study. The planning was performed together with the co-researcher mentioned in Section 4.3 (A.1).

A.2 Establishing questions /A.3 Identifying respondents /A.4 Write accompanying letters

Once the plan was completed the realisation of the plan began. The first thing needed to be done was to find persons willing to participate in the interviews, so called respondents, and to establish questions for the interviews. The aim was to find ten respondents but the researchers only found eight with the relevant knowledge. These steps (A.2 and A.3) together with writing the accompanying letters (A.4) were according to the process plan intended to be done simultaneously and this was achieved. The two researchers established the questions by a brainstorming of questions relevant for the two works. Information about how to establish questions for interviews were found in literature. In the beginning the researchers concentrated mostly on the area that was relevant for their own work. The questions were then put together and presented to the supervisor. The researchers discussed the questions with the supervisor and made some changes before the final version of the question form was put together. The question form was divided into 6 parts: A, B, C, D, E and F. The questions were divided in different parts as they are based on the external data incorporation process given by Strand (2003). Strand's process is divided into four steps: identification, acquisition, integration and usage. The first part of the question form (A) contains introducing questions and the last part (F) contains closing questions. The remaining parts are divided due to the co-operation between the two projects. The parts C and D are included in this work and deals with the acquiring of external data and the integration of external data. The parts B and E are parts of the other work and are dealing with the identification of sources for external data and the use of the external data integrated in the DW.

The identifying of the respondents was performed by searching the Internet and the telephone directory for corporations within the financial industry. All insurance companies and banks that were represented in the yellow pages of the telephone directory were contacted. The respondents will be further described in Section 6.1.

The accompanying letter (Appendix 1) was written by the two researchers and included a brief presentation of the two works: the aim of their research, important

definitions (DW and external data), expected results, how the interviews would be conducted and managed and the interview questions.

The definitions were included since the researchers found it important that the respondents and the researchers understanding about the definitions coincided. If there was a considerable difference between the respondent's and the researchers outlook of the definitions, the respondent's answers will not be included in the analysis. Inmons definition of a DW, included in the accompanying letters is not the definition stated in this work. After the accompanying letter was sent out, the researcher found a newer definition and the researcher chosed to use this instead. The newer definition is based on the previous definition and is established by the same author and the meaning of the two definitions is still the same. The researcher therefore believes that the choise of using the newer definition of Inmon did not affect the work. The interview questions were included so that the respondents would have the possibility to reflect upon the questions in advance and thereby become more comfortable during the interviews.

A.5 Contact respondents /A.6 Schedule respondents /A.7 Send out accompanying letters

The next step in the process was to contact (A.5) and schedule the respondents (A.6). When this was done, the accompanying letters (A.7) were sent out to the respondents. These different steps were carried out simultaneously i.e. when a respondent agreed to participate in the interview he/she was also scheduled and later on that day the accompanying letter was sent out.

After finding the telephone number for a suitable corporation, the work of finding a person in charge of the IT department began. When the person in charge was not available someone else with the relevant knowledge was asked for. When a person with knowledge of the relevant issues had been found, the researchers asked if he/she was interested in participating in the interview. If the answer was positive the interview was scheduled and the researchers then asked for the respondent's address enabling the researchers to send out the accompanying letter.

A.8 Conduct interviews / A.9 Transcribe material and return it to the respondent. / A.10 Respondents give their permission.

The next step of the research process was to conduct the interviews (A.8). The interviews were, as mentioned earlier, conducted in Swedish on a loudspeaker telephone. The researchers used a loudspeaker telephone that enabled both researchers to listen to the answers of the respondent and to get a better understanding of the whole interview. The researchers always started the interview with asking for the permission to record the interviews. The interviews were then recorded on an electronic device, facilitating the work when transcribing the interviews. The transcription (A.9) was carried out directly after an interview had been completed because the transcribing of an interview took long time and the researchers wanted to get the transcribed material ready as soon as possible to send it back to the respondents for approval (A.10).

A.11 Translation, summary and presentation of material.

When the respondents had given their approval to the researchers to use the material, the researchers started to summarise the material (A.11). The material was summarised according to the questions and then translated into English. The summarisation was carried out by revision of the transcripts and the answers were

then cut for the relevant questions and after that the answers were compiled. This had to be done since the respondents occasionally answered a question belonging to another part of the interview. By gathering the answers relevant for a specific question the material would also be easier to divide between the researchers. The researchers then assorted the parts and translated the material.

5.2. Motivating the questions

The questions were, as mentioned earlier, based on the external data incorporation process of Strand (2003). Strand's process is divided into four steps: Identification, Acquisition, Integration and Usage. The identification step deals with the identification of sources for external data. The acquisition step deals with how corporations obtain external data. The integration step deals with how the external data is combined with the internal data and finally, the usage step deals with the usage of external data in a DW. In this work only covers the acquisition and integration.

The questions that were established were aimed to gather a qualitative material, i.e. they were not supposed to generate simple answers as a yes or a no. The main purpose of the questions was to get information sufficient to reach the objectives and further on, also to fulfil the aim presented in Section 3.2. The questions were divided into different parts. The questions of the introduction were established to gather information about the respondent in order to obtain a comprehensive picture of the task performed by the respondent and his/her experience in working with DWs. The introduction questions were also a possibility for the researchers to ensure that the respondents and the researchers shared an understanding concerning the definitions used in this work. The following section of the questions was aimed at the acquisition of data from external sources. These questions were aimed at gathering relevant information about how and in which format the data was received from the data suppliers. Information about the data and in which format it is received is important for what approach will be chosen for the integration. The next section of the questions was aimed at gathering information about the integrating approaches. As integration can be accomplished in several different ways, questions about what approach was beeing used was of relevance to find out if there was one approach used that could be singled out in front of the others. Questions about how data is stored in a DW were of importance to obtain an understanding about how data was integrated in a DW. Further on the researcher aimed the questions to find out how the respondents looked upon the future. This information concerns whether there is an obvious trend in what integration approaches are being used and if the external data will be even more integrated with the internal data in the future. The following part, i.e. the closing questions, was aimed at the respondents opinions of external data in general and what the benefits and pitfalls of external data could be.

5.3. Reflection on the interviews

In this section the conducting of the interviews will be discussed in order to evaluate the execution of the interviews as well as the information gathered.

The interviews were preceded by sending accompanying letters to the respondents. This enabled the respondents to prepare for the interview and to get an increased knowledge about the questions. In some cases however, the respondents did not seem to have read the letter or the questions and the consequence of that was that some of the respondents could not answer all questions, resulting in incomplete answers.

When the researchers conducted the interviews, the answers received from the respondents were first thought to be exhaustive. However, as the summary of the material was performed and the information gathered was revised, the researchers realised that in some cases, the respondents had not really answered the question stated, rather another subject, resulting in lack of information about the relevant issue. A reason for this might be that the question was not clearly formulated or that the respondent did not have enough knowledge regarding the issue. This was not always discovered during the interviews and the reason for this is thought to be the researchers lack of experience in performing interviews. The researchers sometimes felt stressed and were therefore not always focused on the respondents answers, resulting in unfortunate voids and lack of information. However, with more interviews conducted, the researchers felt increasingly secure as interviewers and this resulted in improved results. When interviewing Respondent 3, the researchers had a problem with the recording device. As a result of this, the answers were not recorded properly. The researchers discovered the problem with the recording device during the interview. Respondent 3 was asked to go through the answers again and accepted to do this but the answers given the second time were not as exhaustive as the first time. The researcher believes despite the problem stated above that the most relevant information was acquired.

6. Information presentation

In this section the information gathered during the interviews will be presented. The respondents who participated in the interviews will first be introduced and after the introduction a compiled form of the interviews will be given. The complete interviews with detailed answers are presented in Appendix 2.

6.1. The respondents

The respondents who participated in the interviews are described from the answers given in the introducing questions of the interviews. The respondents work in either large, medium or small corporations. The researcher has chosen to classify the well established and generally known corporations that mainly work with banking as large corporations. The newer corporations that work with banking as a supplementary task are classified as medium sized corporations. Finally, the researcher has chosen to classify corporations that work only in a specific region or only worked with banking for a few years as small corporations. As there was an agreement with the respondents, i.e. they wanted to be anonymous, their name and other information that could lead to the identification of a respondent are excluded below as well as in appendix 2. Respondent 4 did not want his/her material to be included in the appendix.

Respondent 1.

Respondent 1 is working for a large bank in Sweden and is responsible for the DW system in general. The main responsibility for respondent 1 is the connections between the systems and the DW. Respondent 1 has approximately two years of experience in working with DWs.

Respondent 2.

Respondent 2 is working for a relatively small bank in Sweden. His/her main responsibility is customer analysis. Respondent 2 has been working in different environments using DWs for five years.

Respondent 3.

Respondent 3 is working for a relatively small bank in Sweden. Respondent 3 is employed as system administrator manager and the main responsibility is the customer systems. The respondent has prior to this work no experience in working with DWs.

Respondent 4.

Respondent 4 is working as a team leader for a large bank in Sweden. Respondent 4 is responsible for the core activity of the DW where one part is the ETL layer. Respondent 4 has several years of experience in working with DWs and has taken part in several projects developing DWs where he/she has had different roles.

Respondent 5.

Respondent 5 is working for a medium sized, combined bank and insurance company in Sweden. The respondent holds the position of customer analyser and the main area of responsibility involves customer analysis, the developing of sales strategies,

customer strategies and the follow ups on different campaigns. Respondent 5 has approximately two years of experience in working with DWs

Respondent 6.

Respondent 6 is employed as systems administrator in a medium sized, combined bank and insurance company in Sweden. The main areas of responsibility for respondent 6 are the DW, the customer systems and the marketing systems. The respondent has been working with DWs for approximately twenty years

Respondent 7.

Respondent 7 is working as a maintenance manager of the DW in a back office company for banks and financial institutions. His/her main responsibilities are the maintenance and development of the DW, the financial systems and the reporting systems. The respondent has approximately two years of experience in working with DWs.

Respondent 8.

Respondent 8 is an IT-system development manager in a large insurance company in Sweden. The respondent is in charge of a group of system developers working with analysis and different decision support systems in the corporation. Respondent 8 has five years of experience in working with DWs.

6.2. The interviews

In this section the answers concerning the main interview questions (parts C and D) will be presented. Sections 6.2.1 to 6.2.6 covers the acquiring phase and the Sections 6.2.7 to 6.2.12. covers the integration phase. In cases where information is missing, the researchers have not received answers from the respondents and the reason for this could be that a respondent could not answer the question or did not want to.

The acquiring phase

6.2.1. In what way is data acquired from your external sources?

There are according to the respondents two different ways to receive external data. Either by Subscription Service Approach or by On-demand Approach. Respondent 2, 5 and 6 state that they are using a Subscription Service and thereby receive new data on a regular basis from their data suppliers. They use the On-demand Approach as a complement to the Subscription Service Approach when they order data for occasional need. Respondent 3, 4 and 8 only use the Subscription Service Approach. None of the respondents answered that they only use the On-demand Approach. The compiled answers from the respondents are presented below in Table 1.

Table 1. Presentation of answers regarding the acquisition approach of external data.

Acquiring	R1	R2	R3	R4	R5	R6	R7	R8
On-demand		X			X	X		
Subscription		X	X	X	X	X		X

6.2.2. How much of the data you acquire is custom-made for your data warehouse?

Respondent 1 answers that the service company responsible for their data supply, arranges for all data to be structured the way they want it. Respondent 2 describes that

the data they receive is custom-made concerning the data structure but not concerning the substance. Respondent 3 does not give a direct answer to the question regarding how much of the data is custom-made but describes that the data received from two of their suppliers is custom-made. Respondent 3 further answers that they have more than four suppliers of data. Respondent 4 divides the suppliers into two categories: corporations and authorities. Data received from corporations is usually custom-made while data received from authorities only is delivered unadjusted. Respondent 5 answers that almost all data received from their external sources is custom-made. There are however, some sources being used that supply unstructured data for the DW. Respondent 6 answers that none of the data supplied from external sources is custom-made but all data is adjusted before integrated into the DW. Respondent 7 states that whether the data is structured according to their needs or not is depending on what kind of data they need and therefore depending on which supplier they use. Respondent 7 further answers that if the data the corporation needs has not been assembled before, the corporation can have it adjusted the way they want it. Respondent 8 states that most of the data must be restructured and it is not custom-made for the DW.

6.2.3. To what extent are your suppliers able to custom-make the data for your needs?

Respondent 2 states that the suppliers of the external data can supply the corporation with the data they need, structured the way they want. It is just a question of money. Respondent 4 claims that data supplied from corporations as suppliers can be structured the way his/her corporation wants it but when the data is supplied from authorities it is not possible to receive it in another way than in a predefined format. Respondent 5 states that data suppliers do presumably not have the possibility to supply the corporation with data structured to their needs. Respondent 6 answers in the same way, i.e. the data suppliers supply data in a standard format only but there is occasionally, an opportunity to affect the way the data is structured. Respondent 7 describes the situation as following: the corporation has the possibility to have the data they want, structured in the way they want. It is just a question of how much money the corporation is willing to spend. Respondent 8 describes a similar situation: the corporation can get the data they want, structured in the way they want, but since it would mean a huge cost for the corporation this approach is almost never used. But nevertheless, the possibility still remains.

6.2.4. Can you ask your suppliers to adjust occasionally needed data in the way you need it?

Respondent 2 answers that it is possible to receive other data than the normal data received, if there is a need for it. Respondent 3 expresses that he/she is not aware of that their data suppliers do this. Respondent 5 answers that the suppliers can supply the corporation with the data they need. Respondent 6 answers that he/she thinks it is possible but expresses that there are laws limiting what data can be used. Respondent 7 claims that it is possible to get any data requested from the supplier, it will perhaps take a long time before receiving it but it is really just a question of money. Respondent 8 answers that he/she can ask for any type of data but the supplier lacks time to prepare it.

6.2.5. What ways of distribution are you using when acquiring data?

Respondent 1 mentions that the data is received online as a file and downloaded from the suppliers. Respondent 2 receives the data online using File Transfer Protocol

(FTP) or E-mail or by sending the data that is to be integrated to the data supplier where the supplier prepares the file and then returns it to the corporation ready for applying. Respondent 3 answers that their external data is stored at a Web-hotel and when the data is updated it is automatically downloaded online, using file transfer. Respondent 4 answers that they receive the data online. Respondent 5 describes that the data is received in two different ways. The data can either be received as a file online or by receiving a CD-ROM that contains the data. Respondent 6 answers that the data is received either by receiving a CD-ROM or by a file transfer online depending on what kind of data it is. Respondent 7 answers that they download a file from their supplier. Respondent 8 claims that they receive their external data online, downloading files using FTP. The answers given by the respondents are presented below in Table 2.

Table 2. Answers regarding the distribution ways when receiving the external data.

Distributing ways	R1	R2	R3	R4	R5	R6	R7	R8
FTP	X	X		X	X	X	X	X
CD-ROM					X	X		
E-mail		X						
Web-hotel			X					

6.2.6. Can you think of any new approaches for the acquirement of data in the future?

Respondent 1 states they have recently acquired an ETL-tool to use for this operation in the future. They will then receive a file and out of this, extract the records they need. Respondent 2 does not see any new approaches for acquiring the data. Respondent 3 thinks that they will continue to download the data and not use online transactions in the near future. Online transactions can be described as when the corporation always is online with the supplier. Respondent 4 can not think of any new approaches for how to acquire external data. Respondent 5 thinks that there will be an increased use of Internet for the acquiring of external data in the future. Respondent 6 answers that Internet will be used more if there will be a secure way to send data. Respondent 7 answers that a possible approach is the further development of the File Transfer Approach. The respondent further discussed the security issues regarding the possible use of keys and the feature of being online with the supplier in the future. Respondent 8 thinks that the use of transaction servers and the ability to be online and connected to the supplier will become more widespread.

The integration phase

6.2.7. What approaches are used by your corporation when integrating external data into the corporation's data warehouse?

6.2.8. Which one of the approaches mentioned is most frequently used? (Please rank)

Respondent 1 states that the only way they integrate external data into their DW is by applying it into their operational systems and thereafter extract it together with the internal data. Respondent 2 presents that they use three different approaches for the integration of external data. The different approaches are ranked according to which is most frequently used.

- Integrating the external data directly into the DW
- Integrating the external data into operational systems
- Integrating the external data manually

Respondent 3 claims that they match the keys of the data stored in the DW with the keys of the external data and thereby, updates the relevant records. Respondent 4 also uses the approach where they match the external data keys with the keys of the data, already stored in the DW. The key that is being used is organisation number. Respondent 5 claims that they use the same approach as respondent 3 and 4. They match the keys of the external data with the keys of the data previously stored in the DW. Example of keys that could be used is civic registration numbers or postal numbers. Respondent 6 integrates all external data manually by handwritten procedures in SQL. Respondent 7 integrates the data as a dimension in a data cube. Respondent 8 states that they use a Visual Basic program together with SQL to integrate the external data. The SQL together with the Visual Basic program performs the extraction, transformation and loading of the data. Another common approach is the approach where external data is stored into operational systems and then extracted together with the internal data.

6.2.9. Do you know of any other approaches than the one/ones you use?

Respondent 2 has examined an approach where an interface would enable the integration of data that is acquired occasionally, directly but the functionality is much the same as the one before. Respondent 4 mentions that it could be done manually. Respondent 5 can not think of any other approaches integrating external data. Respondent 6 can think of an approach where a software program performing this functionality is developed. The respondent has heard that there are suppliers that supply their customers with this kind of software. Respondent 8 describes a way of using DB2 tables.

6.2.10. How is the external data stored in the Data Warehouse?

Respondent 1 answers that the external data is stored in star schemes and dimension tables in the DW. Respondent 2 states that the external data is stored completely integrated in the tables. This makes it impossible to separate the external data from the internal data. Respondent 3 stores the external data in the database tables. Respondent 4 stores the external data in tables in the DW as well. Respondent 6 stores the external data with the internal data which makes it impossible to see where the data originates from by only looking at it. Respondent 7 answers that both the external data and the internal data are stored in the database tables. Respondent 8 claims that the external data is stored together with the internal data in tables. The answers given by the respondents are presented below in Table 3.

Table 3. How the external data is stored.

Storing	R1	R2	R3	R4	R5	R6	R7	R8
Reference table								
Dimension level	X						X	
Attribute level		X			X	X		
Tupel level			X	X				X

6.2.11. Do you think the integration of external data will increase?

Respondent 1 answers that they definitely think that there will be an increase in the integration of external data in the future. As the DW is used more and the users get an increased understanding of what can be accomplished, they also need more external data. Respondent 2 does not think the integration of the external data will increase. Respondent 3 think they will increase the integration of external data because there is a demand of more information in the corporation. Respondent 5 answered that if the security is improved it could be possible but not in the near future. Respondent 6 states that there will not be an increase in integration of external data but adjustments of the integration is a continuous process. Respondent 7 thinks that the integration of external data will increase in the future as more and more users are using the DW. This increases the need for different data and to integrate more systems. Respondent 8 states that the increase in external data is depending on what laws and restrictions there are and which are to come in the future.

6.2.12. Do you see any obvious trends concerning the integration?

Respondent 2 answers that one trend could be the Probability Theory integration. This integration approach is supposed to match external data with the internal data when there are no matching keys. This could be done when the keys can not be released because of different laws. This is performed by calculating the probability of a match depending on the other fields. Respondent 3 can not think of any trends. Respondent 4 can not think of any noticeable trends either. Respondent 5 answers that they could not see any trends at the moment. Respondent 6 answers that system solutions using Drag and Drop could be the next thing. The tables could then be integrated by dragging one table onto another. But respondent 6 is sceptical to this approach. Respondent 7 answers that one possible trend is that the end-users will have easier access to the DW by the use of web interface instead of expensive licenses to analyse-tools. This is however, not a trend concerning the integration but a trend in the access of the information. Respondent 8 states that instead of having several different small DWs, the trend is to have one large, detailed centralised DW from which other minor DWs are supported. The compiled answers from the respondents are presented below in Table 4.

Table 4. Possible trends in integration approaches.

Trends	R1	R2	R3	R4	R5	R6	R7	R8
Probability Theory		X						
Drag and Drop						X		
End-user interface							X	
Central Data Warehouse								X
None			X	X	X			

7. Analysis

In this section the material gathered during the interviews will be analysed. The analyse will be divided into four parts, one for each objective presented in Section 3.2. The information gathered from the respondents will be discussed regarding the differences and the similarities exposed in the answers. The interviews will be the main foundation for the analysis but the material will also, when possible, be compared to the information in the literature analysis, i.e. Section 2. Data Warehouse. First, the different acquisition approaches will be introduced and discussed. Second, the external integration approaches will be presented and discussed and thereafter, the most common problems concerning the integration of external data into DWs and finally, future trends will be discussed.

7.1. An outline on how external data is currently acquired

There are different approaches on how to acquire external data. The purpose of the external data for a corporation affects in what way the external data is acquired. Corporations could according to Strand, Wrangler and Olsson (2003), either use external data on a strategic level, an operational level or in both ways. The way in which a corporation acquire external data affects how the data is integrated into the corporation's DW. This will be further discussed in Section 7.2. If the data is to be used on a strategic level the external data could be integrated directly into a DW. If the external data however, is to be used on an operational level the data could be applied into the operational systems. As some corporations use external data on both a strategic level and an operational level, corporations could apply external data either by first applying it into the operational systems and then extract, transform and load it together with the internal data into the DW or to integrate it directly into the DW. The approach chosen by a corporation is in this way depending on what kind of external data it is and the purpose of the external data. According to the interviews, five out of eight corporations stated that they integrated the data directly into the DW. Based on this information the researcher believes that these corporations use the external data on a strategic level. The remaining three corporations can be assumed, according to the answers given to use the external data on an operational level.

The Sections (7.1.1., 7.1.2. and 7.1.3) presents how the external data was acquired by the corporations participating in the interviews. In this work, the acquiring phase concerns how corporations receive external data, i.e. by using the Subscription Service or the On-demand service and how the data is distributed and further on, how the data is prepared for integration.

7.1.1 Subscriber Service / On-demand

The most common approach for acquiring external data is according to the interviews, to use the Subscriber Service Approach. Subscriber Service means that the corporations receive external data on a regular basis according to a contract between the data supplier and the corporation. This approach was used by six out of eight corporations. The second most common approach for acquiring external data is the On-demand Approach. Corporations using the On-demand Approach contact their suppliers when they need new external data. This approach was used by three out of eight corporations. The reason for using the On-demand Approach rather than the Subscriber Service Approach could be a money issue. The researcher believes that

corporations that have no need of large amount of external data or only need some external data for occasionally use do not need to buy all expensive external data on regular basis, only when needed.

7.1.2. Different approaches for distribution of external data

External data could be distributed from the data suppliers in several different ways. As the information gathered during the interviews shows in Table 4, the corporations receive external data by using File Transfer Protocol (FTP), CD-ROM, E-mail or Web-hotel. The FTP was the most common approach used by the corporations. The reason for this could according to the researcher, be the increased use of Internet and that FTP is a fast and quite secure way to distribute the external data. Seven out of eight corporations used FTP to receive external data. The second most used approach to receive external data was to use a CD-ROM. This approach was never used alone but as a complement to FTP. The reason why FTP is not used in every case when it comes to distribute external data could be that not all of the external data needed is used on a regular basis. Corporations might sometimes need additional external data from other data suppliers or the supplier may additionally not have the ability to distribute the data using FTP and therefore distributes the external data on a CD-ROM. The service of distributing external data using FTP can also be more expensive than the usage of CD-ROM as more technical equipment is needed. Another reason why FTP is more common is for example that when a corporation establish a contract with a data supplier and the external data is acquired using the Subscriber Service the data suppliers provide the FTP as a service included. Other ways to distribute external data mentioned in the interviews, were to receive external data by E-mail or to download it from a Web-hotel. The E-mail Approach is according to the researcher, supposed to be used when the size of a file containing the external data is small. This approach was also used as a complement to FTP and the reason for this could be the same as the CD-ROM Approach, i.e. the approach is used when external data, from others than the main supplier, is acquired for occasional needs. The last approach mentioned in the interviews was the approach when a Web-hotel was used to distribute external data. In this case, no other approaches were used. When new external data is available, the corporation receives a message from their supplier and can then access the Web-hotel and download the external data.

7.1.3. The Automatic or Semi Automatic Approach

External data received from a data supplier could be of different structure and content. In order to prepare the external data, corporations could use either the Automatic Approach using commercial ETL-tools, the Semi Automatic Approach where corporations have developed ETL-tools on their own or the Manual Approach. Since the Manual Approach was not used by any of the corporations participating in the interviews, this approach will not be further discussed. The researcher believes that there could be different approaches used within a corporation. A corporation could for example use self developed ETL-tools for one purpose and commercial ETL-tools for another purpose.

According to the interviews, five out of eight corporations used commercial ETL-tools for the purpose of preparing the data. The reason for this could according to the researcher, be the advantages mentioned by Gleason (1997) from using commercial ETL-tools, i.e. the automatic generation of meta data and the elimination of the costly process for a corporation that develops ETL-tools on their own. The importance of

meta data is discussed in Section 2.3.3. and also described in literature by Devlin (1997), Inmon et al. (2001) and Marco (2000).

One of the respondents expressed that they have just purchased a commercial ETL-tool but not yet applied it. The researcher believes that more corporations will purchase commercial ETL-tools in the future. The main reason for not using commercial ETL-tools is according to the material gathered during the interviews, that they are too expensive to purchase. This issue will be further discussed in Section 7.3.4. The researcher believes that the usage of commercial ETL-tools are more common within larger corporations. Larger corporations usually have huge amounts of data stored and commercial ETL-tools facilitate the handling of these huge amounts of data. Another reason why large corporations are thought to be using commercial ETL-tools is that they probably have the sufficient funds needed to purchase the expensive commercial ETL-tools. Smaller corporations might very well not be able to come up with the funds needed to purchase commercial ETL-tools. This hypothesis is supported by the information gathered during the interviews as it shows that usage of self developed ETL-tools are more common in smaller corporations.

Three out of eight corporations had developed ETL-tools on their own. This approach is in this work called the Semi Automatic Approach (SAA). One reason mentioned in the interviews, for a corporation to use the SAA is that the corporation can preserve control over what data enters the DW. The issue of controlling data originates from the fact that one of the main problems concerning integration of external data is poor data quality. The problem of poor data quality will be discussed in Section 7.3.3. and it is also mentioned by Strand and Olsson (2003). The researcher believes that having control over the data could be possible if the amount of external data to be integrated is of a manageable size.

The corporations using the SAA have usually developed their own software to perform the same tasks as commercial ETL-tools. Software developed in the corporations however, usually performs just some of the activities. This enables the corporation to be fully flexible in the handling of data but on the other hand, additional programs are needed. Respondent 8 for example, states that they have developed software for cleansing and loading the external data using Visual Basic together with SQL. One disadvantage using the SAA is according to Gleason (1997), that self developed software usually lacks the ability to generate the important meta data describing what transformation rules have been implemented. To manually generate documentation containing the meta data is a time consuming and costly process, as always when producing documentation in systems development.

The procedure of the SAA was outlined by respondent 6. Respondent 6 describes the course of action as follows: first, the external data is prepared using SQL. Second, the external data is transformed to fit the DW's structure and the data to be integrated is handled again with SQL and stored procedures. If the source of the data and the external data are to be used again, the stored procedures are stored in the Workmanager for future usage. Respondent 6 declares that every new case of integration of external data is dealt with as a new case for which the approach is adjusted to fit the structure and content of the received external data. Respondent 6 further states that the only way to really acquire knowledge about the data received and what transformation might be needed is to manually control the code. Respondent 6 has built the DW from scratch and has full knowledge about all the systems in the corporation. This enables the respondent to know exactly how the data should be structured in order to fit the DW. By preparing the external data using the SAA,

respondent 6 has got complete control over what data is stored in the DW. The reason why the SAA is used rather than the more common Automatic Approach (AA) is according to respondent 6, a strategic choice. The DW in respondent 6's corporation has developed slowly and the main idea was to have a flexible system. When a complete system (a DW) is purchased it is a very big step for a corporation. As mentioned before, the cost for a complete system is very high. Respondent 6 further states that the expectations in the performance of a system are sometimes out of proportion. When a complete system is purchased there are usually constraints limiting what is possible to accomplish. By developing the system in minor steps the ability to keep the system flexible is maintained. Additional parts are custom-made to fit the existing system and the administrator thereby preserves full control over the system. Having control of the system and the data stored in the system is according to respondent 6, the key to success. The other respondents did not discuss this issue in details.

The researcher believes there are three main reasons for using the SAA to prepare external data. First, to maintain control over the external data and second, the cost of commercial ETL-tools that handle the integration, is too high. The issue of expensive commercial ETL-tools is further discussed in Section 7.3.4. By using the SAA, DW technicians have more control of the data, i.e. regarding what transformations must be done and quality issues concerning the received external data. Poor quality is as mentioned by Strand, Wrangler and Olsson (2003), one of the main problems when integrating external data. By managing the preparing process manually the researcher believes that corporations feel more secure about the quality of the external data integrated.

7.2 An outline on how external data is currently integrated

As discussed previously, depending on what a corporations purpose for the external data is, the approach on how to acquire it is affected. If a corporation is to use the external data on a strategic level, the data is usually received in a preformulated way from their data supplier, according to the interviews. But as mentioned before, external data is also sometimes received from other sources. The content of the external data is then not prepared. Considering that the quality of the external data could be poor, corporations sometimes chooses not to unite the external data with the internal data. This could be achieved by integrating the external data into a reference table or into a separate dimension. When external data is acquired from a data supplier that delivers external data of good quality regarding its content, the external data could be united with the internal data. Approaches on how to integrate the external data together with the internal data are by integrating it into an attribute level or into a tupel level. These four approaches of integrating external data are described below.

- to integrate data into dimensions
- to integrate data into attributes
- to integrate data into tupels
- to integrate data into reference tables

The first approach is to integrate the external data into a separate dimension in the DW. By integrating external data into separate dimensions the external data will not unite with the internal data. This could as mentioned before, be an advantage as

external data sometimes is of poor quality. Example of external data that could be integrated this way could be: a business partner's customer record as shown in Figure 14. This approach was used by two out of eight corporations.

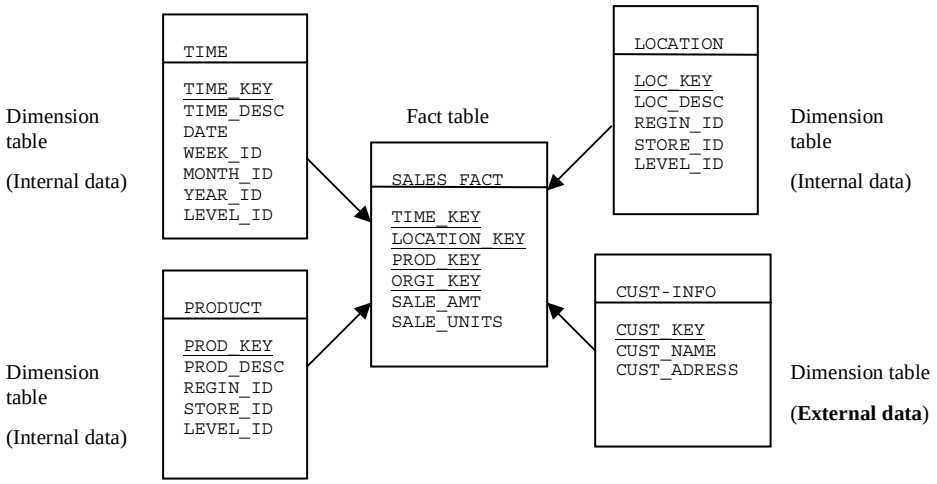

Figure 14. Example of external data integrated in a dimension.

The second approach is to integrate the external data into attributes. By integrating external data into attributes the external data is united with the internal data and the external data can then not be separated from the internal data. Example of an external data that could be integrated in this way was mentioned by one of the respondents to be customer ratings. The respondent explained that customer ratings from the data supplier Upplysnings Centralen could be used for segmentation of customers. An example of attribute integration is presented in Figure 15. This approach was, according to the information gathered during the interviews, used by three out of eight corporations.

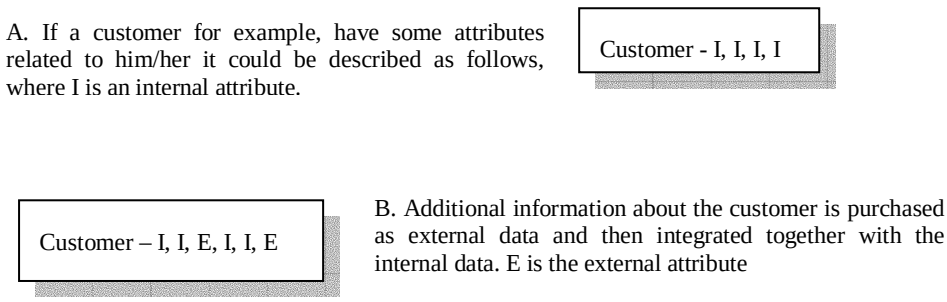

Figure 15. A) Example of internal data as attributes. B) Example of when attributes based on external data are integrated into an attribute level.

The third approach is to integrate the external data into tupels. In this way the external data is united with the internal data and it is impossible to separate the external data from the internal data. An example of integrating data into a tupel could be that a corporation have the addresses of their customers integrated into a table as tupels. An example of this is shown in Figure 16 A. When a customer moves, the new address is

purchased from a data supplier and the old value is updated and replaced with the new one as the DW is refreshed. An example of this is shown in Figure 16 B. This approach was used by three out of eight corporations.

A) Tupels only containing internal data

C_ID	C_Name	C_Adress
I	I	I
I	I	I
I	I	I

B) Tupels containing both internal and external data

C_ID	C_Name	C_Adress
I	I	I
I	I	E
I	I	E

Figure 16. A) Example of internal data on tupel level. (I=internal data) B) Example of internal and external data integrated into a tupel level. (I=internal data, E=external data)

The fourth and last approach on how to integrate external data is to integrate it into a reference table but here the data is not actually integrated into the DW. Corporations sometimes choose to store external data separately from internal data as one of the main problems with external data is the sometimes poor quality. The issue of external data's poor quality is further discussed in Section 7.3.3. When a data analyst needs to use external data stored in a reference table with internal data stored in a DW, an interface is used. An example could be: a corporation receives a branch index as external data and this is stored in a reference table. In an interface a data analyst could analyse the corporation's internal data and by using the external data stored in the reference table he/she can compare the corporations result with the branch index. The external data is not always as described above, integrated into the DW but it can also be integrated into analyse-tools. An analyse-tool plots the branch index on the screen enabling the analyst to compare it with the corporations return. An example of this is shown in Figure 17. This approach was not used by any of the corporations but as Devlin (1997) mentions reference tables in his literature, the researcher would like to present it as a possible approach.

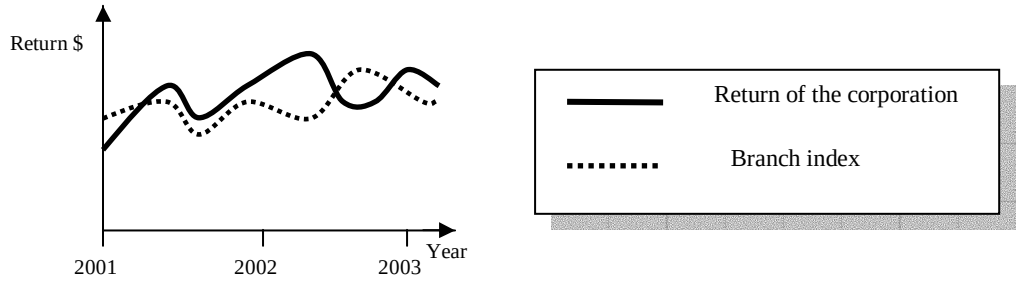

Figure 17. Example of external data integrated with internal data using an interface. The internal data is stored in the DW and the external data is stored in a reference table. The user can not see the difference in how the data is stored.

When the integration of the external data is discussed on a higher level, the information gathered during the interviews shows that it is common to integrate the external data with the internal data into the DW. The researcher believes the reason for this to be the improvement in quality of external data as most respondents use data suppliers. The researcher further believes that as data quality improves, corporations feel more secure about integrating external data into the DW.

7.3. Common problems concerning integration of external data.

There are according to the information gathered from the interviews, several different problems concerning the integration of external data into DWs. In this section the different problems will be described and discussed.

7.3.1. The problem of data structure

The most common problem regarding integration of external data into DWs is the difference in structure of external data compared to the structure of internal data in DWs. This problem is also described in the literature by Devlin (1998) and Inmon et al. (2001). The data must be transformed to fit the data structure of the DW and this is a very time consuming and costly process. This problem could be solved if the data suppliers adjust the external data in the way that the corporations request. According to the information gathered during the interviews, external data could be divided into three categories: external data from authorities, structured data from data suppliers and unstructured data from data suppliers. (Strand, Wrangler and Olsson, 2003).

External data received from authorities is not structured in a specific way, i.e. the data is not structured to fit the systems of the corporations. Usually the authorities have no interest in what structure the data has, they just supply it and then it is up to the corporations if they want to use it or not. The external data distributed from authorities is not intended for DWs. If corporations want to use the external data supplied from authorities they need to transform it, otherwise it will not fit their systems. The researcher believes that since authorities have nothing to gain from structuring the data in the way corporations need, the problem with the data structure will remain in the future.

External data received from data suppliers is as mentioned, either structured or unstructured. Information retrieved during the interviews shows that most of the external data suppliers could structure external data in the way the corporations request. Seven out of the eight corporations that participated in the interviews receive

at least some external data that has been custom-made to fit their DWs. Generally, customisation of external data concerns only changing the structure of the data in order to make it fit the DW's structure and not the content of the data. The corporations usually purchase both structured and unstructured data. The reason why corporations do not receive all external data custom-made, both in structure and content is according to the interviews a money issue. In a case where a supplier custom-makes the entire external data for a corporation, the data would be an enormously expensive purchase. External data is according to the information gathered during the interviews, expensive to purchase even if the data is not customised.

7.3.2. Restricting laws

Another problem concerning the integration of external data is that it exist laws that prevent corporations from integrating some categories of data. Information gathered during the interviews describes that corporations find it difficult to know exactly what they are allowed to do with external data. Person Uppgifts Lagen (PUL) for example prevents some data to be stored. Examples given by the respondents of categories of data prevented to store in a DW by PUL are: the income of a customer, number of children and the marital status of a customer. Corporations are in this way prevented from segmenting their customers using this information. Another example mentioned is that purchased information about civilians can contain civil registration numbers and these are sometimes not allowed to store into a DW. This issue could possibly be avoided by the use of a new integration approach. This integration approach is further in this work called the Probability Theory Approach and will be described in Section 7.4.1. Information about corporations is according to the interviews, not as restricted as information about civilians. Furthermore, information about how external data is allowed to be used, what data is allowed to be united as well as what and for how long external data is allowed to be stored is a general problem to corporations. The researcher believes this issue to be a problem that will be of less importance in the future as the knowledge and maturity of corporations in the area, working with external data, will increase. However, as new laws and restrictions probably will appear in the future, this problem will continue to be an issue for corporations to take into consideration before integrating external data. The researcher believes that the issue of restricting laws could be different if another business area had been researched. Financial corporations are believed to have more restrictions affecting what kind of data they are allowed to use or integrate than for example industrial areas have.

7.3.3. Poor data quality

The problem with poor data quality is a well known problem and is mentioned in literature by Adelman (1997), Strand and Olsson (2003) and Inmon et al. (2001). The information gathered during the interviews presents three problems regarding data quality. First, the age of external data is mentioned as a source to problems. Data acquired from external sources could be old and when old external data is integrated and then used in a DW, the result is not accurate. Internal data is according to Inmon et al. (2001), timestamped before integrated into a DW. External data must also be timestamped but as the origin of the external data sometimes is unknown, this could be difficult. From the researcher's point of view, this problem is usually related to data received from other sources than data suppliers as data suppliers generally deliver data of good quality. This issue is also discussed by Strand and Olsson (2003).

The second problem was the above already mentioned issue of the sometimes unknown origin of the external data. As corporations do not know from where some parts of the external data originates, they find it difficult to rely on. The researcher believes that if corporations are not able to rely on the external data they are about to integrate, it could be difficult to know what data can be extracted as the content could be of poor quality. The researcher believes that this problem will slowly vanish as more corporations acquire their external data using data suppliers. By receiving external data from a data supplier, corporations know the origin of the external data.

The third and last problem concerning poor data quality mentioned during the interviews, was the issue of dirty data. Dirty data is data containing information that is wrong in some way; either incorrect or insufficient information could then be stored. The issue of dirty data is from the researcher's point of view, possible for corporations to avoid if they acquire their external data from data suppliers. When a corporation acquire external data from a data supplier there is usually a contract between the two parts describing what structure and content the data must have to fit the needs of the corporations.

One disadvantage using data suppliers is according to the information gathered during the interviews and also the researcher's point of view, that purchasing all external data will result in major costs for the corporations. The researcher believes that even if today's prices of external data are high, the overall cost of acquiring external data will decrease in the near future due to more data suppliers arriving on the scene. Expenses due to poor data quality include errors in analysis, errors when integrating external data as well as time and personal efforts transforming external data into suitable format.

7.3.4. Expensive tools

Three out of eight corporations did not use commercial ETL-tools when integrating external data into their DWs. Information gathered during the interviews shows that a reason not to purchase commercial ETL-tools is the high cost. It is the belief of the researcher that corporations integrating large amounts of data from several different sources would in the end benefit from using commercial ETL-tools rather than self developed ETL-tools. The researcher further believes that to develop or adjust software for every system in a corporation with continuously changing systems environments will in the end be even more expensive than to purchase an commercial ETL-tool. Commercial ETL-tools could however, be difficult to purchase for smaller corporations that do not have sufficient funds. Kimball (1998) states that commercial ETL-tools are not always cost effective for smaller corporations with smaller DWs. The researcher does however believe, in line with Kimball (1998) that the cost for commercial ETL-tools will decrease as more commercial ETL-tool software will enter the market. The competition between software developers will affect the price on the commercial ETL-tools in a positive way for customers.

7.4. Future trends concerning integration of external data

The information gathered during the interviews shows as presented in Section 6.2.12, no obvious trends regarding future approaches of the integration of external data into DWs. There are however, three different approaches that are mentioned as possible trends from the respondents. These will be presented below and finally, a general discussion on further acquisition and integration of external data will be given.

- Probability Theory

- Drag and Drop
- Centralised Data Warehouse

7.4.1. Probability Theory

The most interesting approach according to the researcher, is the Probability Theory Approach (PTA). This approach was described by respondent 2. the PTA is used when external data is integrated into a DW where the external data does not have keys at all or not the same keys as the internal data already stored in the DW. The reasons for why external data occasionally does not have keys could be various. Five out of eight respondents mention that one of the main problems concerning external data is that the corporations are not allowed to store all existing data about their customers because of PUL and other related laws. These laws occasionally prevent the corporations to store for example, social security numbers. The course of action using the PTA is described by respondent 2. The external data that is received by the corporation is usually structured in a way that a person, whose data is part of the dataset, can not be identified within the external data delivered, i.e. the social security number could for example be left out. The data already stored in the DW usually contains a lot of information about for instance, customers. When a new set of data is delivered, containing additional information about these customers and there is no key to match the records, the PTA is applied. This approach uses the records already stored in the DW with the records in the external data. Occasionally, a lot of the external data is already stored in the DW and the new data set is only providing various new records of information about the customers. The existing records, already stored in the DW are compared to the records in the external dataset. The probability that two customers compared are the same customer is then calculated. If the probability is high enough, the new records of information stored in the external dataset are applied to the customers records stored in the DW. In this way the corporation can integrate the additional information in the external data without having to use the keys as a match and thereby avoid restrictions stated by the laws.

The researcher of this work has searched in literature and on the Internet for further descriptions of this approach but no material was found. The reason for this could be that the PTA has different names unknown to the researcher.

7.4.2. Drag and Drop

The other approach the researcher thought were of special interest in this work was the Drag and Drop Approach (DDA). This approach is described by respondent 6 as a system where the data is integrated using an interface. DW technicians can designate the table of which they want to integrate and then drag it onto the table of which they want to integrate the data, then drop it and it will integrate automatically. This approach is however, very technically dependant. This means that for this to be possible, the system must be able to solve all different transformations, both simple and complex as presented in Section 2. The DDA was mentioned by a corporation that used the SAA. The researcher believes that this approach compares to the commercial ETL-tools used by several other corporations. It is the researcher's point of view that the usage of commercial ETL-tools will be more common in the future

because of the different benefits gained from using these. Benefits using commercial ETL-tool are presented in Section 7.2.1.

7.4.3. Centralised Data Warehouse

One of the respondents answered that they have several smaller DWs in the corporation at the moment. The respondent believed a possible trend could be to replace the different smaller DWs with a large centralised DW and instead use data marts for the different departments. This is discussed in literature by Chaudhuri and Dayal (1997), Inmon et al. (2001) and Sperley (1999). The issue of having a large centralised DW and from this feed the different data marts is presented in Section 2.3.5. The researcher believes that if a large centralised DW is to replace the several smaller DWs with data marts, these should be dependent data marts as this would result in that the data marts will have the same source even if the content of the data marts would differ. As the source of the data marts is the same, all data marts will have access to the same data enabling the different departments to make decisions based on the same information. The researcher believes further that a disadvantage using a centralised DW as a source could be that if data of poor quality is integrated into the DW, all departments will be affected.

7.4.4. The increase in integration of external data

Three out of eight respondents answered that they believe that the integration of external data will increase. The reason for this were according to the respondents that as the DW is more used and the users mature and get an increased understanding of what can be accomplished, the need for external data will increase. The researcher agrees with this as the DW area is relatively immature in Sweden and will continue to develop. As more experience is acquired, more possibilities are believed to be discovered. Support for this issue can also be found in literature where Devlin (1997), Inmon (1999a) and Singh (1998) discuss the possibilities with external data.

Two out of eight respondents answered that the increased integration of external data is depending on some issues that has to be dealt with. One of the respondents stated that he/she did not think that there would be an increase in the integration of external data due to security issues. What those issues were was not further developed. The other respondent believed that the increase in the integration of external data would be depending on if there will be any new laws or restrictions concerning the issue. The researcher believes that the most probable reason for corporations not to increase the integration of external data will be future laws. This hypothesis is supported by the answers of the respondents where they express the problems with current laws since they prevent in what way the external data is allowed to be used.

Another issue that was discussed by a respondent during the interview whether to increase the integration of external data or not was to consider the possibility to use the already integrated external data more. If the external data already integrated into the DW could be used more efficient, there is perhaps no need for more external data. The researcher believes that as corporations mature in their knowledge and usage of external data and realise what potential possibilities there are, new kinds of external data can be integrated.

8. Conclusions

In this chapter conclusions of the analysis will be presented. The result will be presented, based on the four objectives presented in Section 3.2 and these objectives together will reach the aim of this work.

For repetition these objectives are:

- Give an outline on how external data is currently acquired
- Give an outline of current external data integration approaches
- Identify the most common problems concerning the integration of external data into DWs.
- Identify future trends in integration of external data into DWs

Each objective have been assigned one section for further discussion.

8.1. An outline on how external data is currently acquired

The purpose of the external data for a corporation affects the way the external data is acquired. The external data can according to Strand, Wrangler and Olsson (2003), be used on either a strategic level, an operational level or in both ways. If a corporation is to use external data on a strategic level they could integrate it directly into the DW but if they should use it on an operational level they could apply it into the operational systems. Strand (2003) presents a process model on the whole process needed for corporations to incorporate external data. As this work has outlined how external data is currently acquired the researcher believes that the acquisition phase can be divided into sub-phases, Figure 18. These sub-phases are concerning the frequency of acquiring the data and involves if a corporation uses the Subscription Service Approach, receiving external data on regular basis, or the On-demand Approach, receiving external data for occasionally use. The second sub-phase is concerning how the external data is distributed from the data supplier. The third and last sub-phase concerns whether the external data should be prepared using commercial ETL-tools or self developed ETL-tools.

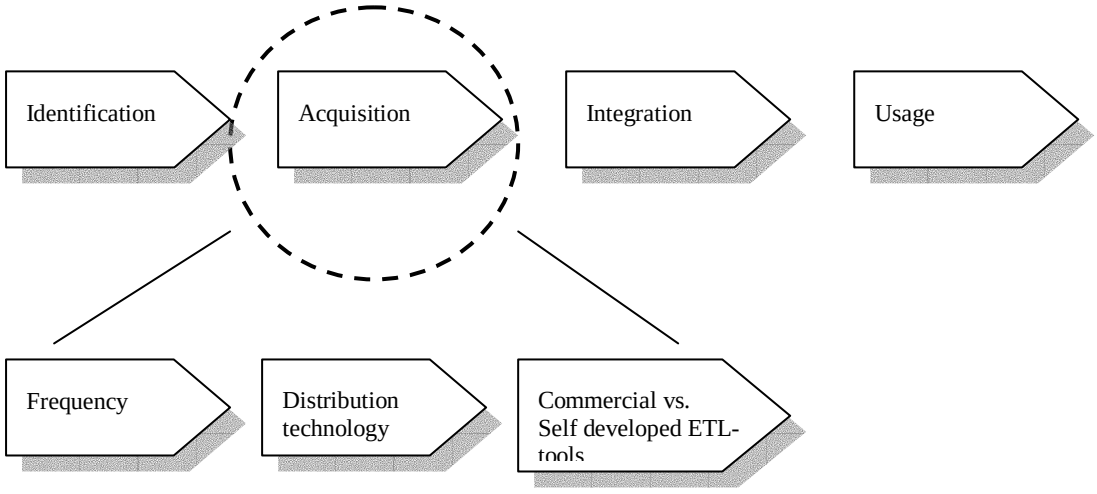

Figure 18. Example of the external data incorporation process with outlined sub-phases for acquisition. (Adopted from Strand 2003, p.4).

Information gathered during the interviews presents two main approaches on how external data is acquired. These two approaches are: the Subscription-Service Approach and the On-demand Approach. The Subscription Service Approach is used by corporations that use external data frequently. Corporations with these demands usually establish a contract with their data supplier to assure the quality and the structure of the data to be purchased. When some external data is needed occasionally, the On-demand Approach could be used as a complement to the Subscription Service Approach. Corporations that do not use external data on regular basis also prefer to use the On-demand Approach. One reason for using the On-demand Approach in front of the Subscription Service Approach could, from the researcher's point of view, be a money issue as external data is expensive to purchase.

The corporations that participated in the interviews usually received the external data from their data suppliers by using the File Transfer Protocol (FTP). Other ways of receiving external data were: by CD-ROM, by E-mail and by Web-hotels. None of the corporations used CD-ROM or E-mail as the only source to receive external data but as a complement to the FTP. The researcher believes that corporations receive additional external data that is not used regularly, on CD-ROM. According to the researcher these two approaches are used when external data, from other sources then the main data supplier, is needed for occasional use.

The last approach mentioned in the interviews was the Web-hotel Approach where a Web-hotel was used to distribute external data. In this case, no other approaches were used.

To prepare the external data received from external sources, the information gathered during the interviews, presents two approaches used for this purpose. The Automatic Approach (AA) using commercial ETL-tools and the Semi Automatic Approach (SAA) using self developed ETL-tools.

Five out of eight corporations used the AA. The reason for this is not clear but the researcher believes one major issue to be the advantages of commercial ETL-tools have compared to the self developed ETL-tools. It is the belief of the researcher that there will be an increased use of commercial ETL-tools in the future. The reason for this is believed to be that the prices of the commercial ETL-tools will probably decrease as more commercial ETL-tools will enter the market. This statement is also supported in the literature by Kimball (1998). One advantage using commercial ETL-tools is that they generate meta data automatically.

Three out of eight corporations used the SAA. The reason for this approach to be used rather than the AA was mentioned to be that commercial ETL-tools are very expensive to purchase. One additional reason to use the SAA rather than the AA is that purchasing expensive commercial ETL-tools is not always cost effective for smaller corporations.

8.2. An outline on how external data is currently integrated

The approach a corporation chose to apply when integrating external data could be depending on the source used and thereby the quality of the external data received. If the external data is of poor quality a corporation could chose not to integrate the external data with the internal data. Two out of eight corporations integrated their external data into a separate dimension table. In this way the external data is stored separated from the internal data.

If the quality of the external data is good enough, the external data can be integrated with the internal data. Three out of eight corporations integrated their external data into attributes. This way the external data is integrated with the internal data. Three of the eight corporations participating in the interviews stated that they integrated the external data into a tupel level. In this way the external data is also integrated with the internal data. According to the above discussed results there is no integration approach more common than another but when reflecting over the results on a higer level, the results show that it is more common to integrate the external data with the internal data into the DW. It is the researcher's point of view that the reason for this is the improvement in quality of external data as most respondents use data suppliers.

8.3. The most common problems concerning integration of external data into Data Warehouses

The most common problems concerning the integration of external data into DWs were according to the interviews, the problems with data structure of external data, restricting laws, poor data quality and finally, expensive tools.

The most common problem regarding the integration of external data into DWs is the difference in structure of external data compared to the structure of internal data in the DW. In order to integrate the external data, the external data must be transformed to fit the data structure of the DW. This is a very costly and time consuming process. According to the information gathered during the interviews, external data could be divided into to three categories: external data from authorities, structured external data from data suppliers and unstructured external data from data suppliers. The problems concern the received external data acquired from either authorities, as they do not prepare the structure of the external data in any way, or unstructured external data from data suppliers. The reason why not all data is received in a prepared way is that it would be enormously expensive to purchase for the corporation.

The second problem concerning integration of external data into DWs is existing laws that are preventing corporations to integrate some categories of data. Information gathered during the interviews describes that corporations find it difficult to know exactly what is allowed to do with external data, what data is allowed to be integrated as well as what and for how long external data is allowed to be stored. The problems corporations have with this issue is, the researcher believes, a problem that will be of less importance in the future as corporations knowledge and maturity in the area, working with external data, will increase. However, as new laws and restrictions probably will appear in the future this problem will continue to be an issue for corporations to take into consideration before integrating external data.

The third problem was the problem of poor data quality. The information gathered during the interviews presents three problems regarding data quality. These are: the age of external data, the unknown origin of external data and the issue of dirty data.

The age of external data

The age of external data is mentioned as a source to problems. It is sometimes difficult to know what age data acquired from external sources has and when old external data is integrated and then used in a DW, the result might not be accurate. Another issue concerning the age of external data is that data stored in a DW should be timestamped. If the age of the external data is unknown, to timestamp it could be difficult.

The unknown origin of external data

Another problem concerning external data is the sometimes unknown origin. As corporations do not know from where some of the external data originates they find it difficult to rely on.

Dirty data

The last problem concerning poor data quality is the issue of dirty data. Dirty data is data containing information that is either wrong or insufficient.

It is the belief of the researcher that these three problems mentioned can be avoided if the corporations acquire their external data from data suppliers. Data suppliers should be able to guarantee the quality of the external data in terms of: the age of the data, the origin of the data and also the content of the data. One disadvantage using data suppliers is however, that external data is expensive to purchase.

The fourth and last problem concerned expensive ETL-tools Three out of eight corporations did not use commercial ETL-tools. The reason for the three corporations not to purchase commercial ETL-tools was mentioned to be the high cost. However, the researcher's point of view is that competition between software developers will affect the price on the commercial ETL-tools to decrease. This is in line with Kimball (1998) as the author states that the cost of commercial ETL-tools will decrease as more commercial ETL-tools enter the market. The three corporations that did not use commercial ETL-tools used self developed ETL-tools.

8.4. Future trends in integration of external data into Data Warehouses

The information about possible future trends concerning the integration of external data into DWs could be divided into two categories. First, approaches for integrating external data into DWs and second, the increase in integration of external data.

The Probability Theory Approach

The information gathered during the interviews did not point out any obvious trends concerning the integration of external data. One respondent stated that one approach that might be used more in the future is the Probability Theory Approach. This approach could be used to integrate external data and internal data where the keys do not match, or if the external data have another key than the internal data or no key at all. It is the researcher's belief that this approach will be increasingly used as this approach enables corporations to use and to integrate the additional information in the external data without having to use keys as a match and thereby avoid restrictions stated by the laws.

The Drag and Drop Approach

Another approach mentioned as a possible future trend was the Drag and Drop Approach. The course of action for this approach is: the table the DW technician wants to integrate is designated and then dragged onto the table of which he/she wants to integrate the data. The data will then be integrated automatically. The researcher believes this approach to compare with commercial ETL-tools used by several other corporations.

Centralised Data Warehouse

Example of another possible future trend is to have a large centralised DW and instead use data marts for the different departments. There are two different kinds of data marts, dependent and independent. The researcher believes that dependent data

marts would be more suitable than independent data marts. By using dependent data marts the different departments data marts is fed from one source resulting in that the same data is used throughout the corporation.

The increase in integration of external data

Three out of eight respondents answered that they believed in an increase in the integration of external data since the increased usage of DWs will result in an increased understanding in what is able to accomplish. It is the researchers point of view that external data will be more and more used by corporations and the external data's importance will grew in the future as the knowledge and understanding of what is allowed and how it is accomplished will increase.

Two out of the eight respondents described issues that might affect the increase in integration of external data in a negative way. These issues were: future laws preventing what is allowed to do with external data and security issues that must be dealt with before increasing the integration of external data.

One respondent stated that corporations would perhaps not increase the integration of external data in the future but instead use the already integrated external more. If the external data already integrated into the DW could be more efficiently used there is perhaps no need for more external data. The researcher believes that as corporations mature in their usage of external data and realise what potential possibilities there are, new kinds of external data can be integrated.

9. Discussion

In this chapter the researcher will discuss the experience gained from this work. Furthermore, this work will be evaluated in a wider context and finally, ideas of future work will be given.

9.1. Experience gained from this work

While conducting this work the researcher has gained a lot of experience of different natures. The experience is divided into three categories: experience in conducting a relatively large work, experience in conducting interviews and experience in writing in a foreign language.

9.1.1. Experience in conducting a relatively large work

While conducting this work the researcher has gained experience in what efforts are involved in performing a research. A simple plan of how to conduct this work was established in order to allocate enough amount of time for the performance of each one of the different parts of the work. The researcher soon realised that the time first allocated for each one of the different parts would not be sufficient. This led to several changes of the original plan. The researcher believes that there were many reasons for the misjudgement of the time needed to perform each part. One reason according to the researcher, was the lack of knowledge in what efforts are involved in establishing a suitable aim for a work. Before the final aim was established the researcher had to change it several times.

Another misjudgement made by the researcher was the time effort involved in finding literature on the subject. The researcher knew from the beginning that there was a shortage of relevant literature on the chosen subject but as the work proceeded the researcher realised that the lack of literature would lead to a problem with keeping up with the time first allocated. The researcher believes that despite this problem, the amount of literature found on the subject is sufficient for presenting enough background material.

9.1.2. Experience in conducting interviews

The method/material collecting technique used for gathering information to this work was the interview study. The researcher had, before conducting the interviews in this work, little experience in conducting interviews. From conducting the interviews, the researcher has gained experience in how to establish questions, how to find suitable respondents and how to transcribe material. The transcriptions took for example, longer than the researcher first expected. Conducting the interviews was also a much more demanding task then the researcher first thought. As the researcher conducted more interviews, the results improved. The importance of establishing easily understood and well-formulated questions is the issue the researcher feels is the most important one. The researcher realised, when reflecting over the material from the interviews, that some questions had been indistinct. This realisation is a great experience for the researcher. Despite the sometimes indistinct questions, the researcher believes sufficient material was gathered enabling the researcher to present conclusions about the aim of this work.

Finding suitable respondents to participate in the interviews of this work was not as easy as the researcher first believed. The researcher aimed for ten respondents but the initial attempts to find respondents made the researcher aware of the shortage of information and experience in this specific research area. The therefore insufficient numbers of respondents led to an agreement between the researcher of this work and a researcher conducting another work within the same area to co-operate during the interview phase. Even though the person in the corporations with most relevant knowledge on the subject was asked for, the persons that later participated in the interviews did not always have the answers to all questions. The reason for this could be the difference in aim between the two works. One respondent able to answer the questions asked by one researcher were not always able to answer the questions asked by the other researcher. A test interview could have been preformed by the researcher before the main interwievs had taken place. If this had been done the researcher could maybe have realised if the questions were distinct or not. The reason why the researcher did not perform a test interview was that the researcher thought that since there was already a shortage of respondents, to use one only as a test would mean too much of a sacrifice for this work.

9.1.3. Experience in writing in a foreign language

The researcher believes that by writing the report in English, the researcher's knowledge in the English language has increased. The researcher further believes that the choice of writing in English has affected the efforts involved in writing the report and it has meant a more time consuming process than if it had been written in Swedish. One problem the researcher had with using the English language was the difficulties to sometimes express exactly what the researcher meant due to a limited vocabulary. Another issue concerning writing in English was that it was difficult to translate the interviews, word for word, from Swedish into English.

9.2. Evaluating the work in a wider context.

This work is, as mentioned in Section 3.3., delimited to corporations within the financial business area in Sweden. Because of this delimitation some of the results of this work are difficult to generalise as corporations in this business area have special characteristics that do not apply to other business areas. Corporations within the financial business area are prevented to use some data by several different laws concerning civilians while other business areas as the manufacturing business area, might not have an interest in using such restricted data. As different business areas use different kinds of data, specific issues concerning the data used can be difficult to generalise. The issue on how to acquire external data is believed by the researcher to be easier to generalise and apply for other business areas since the results of this work show that depending on if the data is used on a strategic level or on an operational level, the acquirement and later the integration of external data is affected. The researcher believes that the issue of corporations purposes for the external data affects the way it is integrated, i.e. applied into operational systems, integrated directly into a DW or integrated into a DW through operational systems, could be applied to other business areas as well.

One conclusion is that the acquisition approach a corporation uses is depending on what a corporation's purpose for the external data is. If a corporation uses external data on an operational level the external data acquired could be applied into the operational systems and used with the internal data. If a corporation uses the external data on a strategic level the external data could be integrated into the DW and stored

separately from or integrated with the internal data. In cases where external data is used both on an operational level and on a strategic level the external data is first applied into the operational systems and is then extracted together with the internal data. Having the above information in mind when reading the following definition of external data by Devlin (1997 p.135.)

"business data (and its associated meta data), originating from one business, that may be used as part of either the operational or the informational processes of another business"

the researcher believes this definition could be altered as the external data could be used in both the operational and the informational processes of another business. The altered definition could be

Business data (and its associated meta data), originating from one business, that may be used as part of the operational *and*/or informational processes of another business.

The purpose of this work was to find out and acquire knowledge in current approaches for acquiring and integrating external data into DWs and to give a brief overview of the future trends for external integration.

This was intended to be achieved by reaching the aim of each and one of the different objectives presented in Section 3.2. When evaluating the result compared with the objectives the researcher believes that the overall aim of this work has been reached. However, when reflecting on the results for each and one of the different objectives the researcher believes that some of the results could have been more exhaustive.

Concerning the first objective, give an outline on how external data is currently acquired, the results acquired are believed to be relatively exhaustive. This as the respondents presented the current situation and presented a relatively common view on what different approaches there are for acquiring external data. This is also believed to be the case when reflecting on the second objective, give an outline of current external data integration approaches. Concerning the third objective, identify the most common problems concerning the integration of external data into DWs, the researcher believes that there are more problems related to the integration phase than presented in the results. If a precise question would have been established instead of the sometimes posted follow up questions, concerning existing problems with the integration of external data, the researcher believes that this could have resulted in a more exhaustive material to analyse, i.e. other problems could have been exposed. As there was no precise question concerning this issue, the problems with the integration of external data are not presented in the material presentation. The reason for the results to be somewhat limited concerning this objective could also be that the respondents participating in the interviews did not have the relevant knowledge of this issue and were therefore not able to think of any other problems. The material gathered during the interviews contained information that was not relevant for this work as the questions were sometimes not as distinct as the researcher first believed. In despite of this it is the meaning of the researcher that the choice of using the interview study as the method for how to perform this research, was suitable. When the researcher reflects over the other methods, there is no one believed to have resulted in better material to analyse than the one used. The fourth objective, identify future trends in integration of external data into DWs, was believed to give more exhaustive information about future trends. The reason why no more information was received concerning this issue could again be the respondents lack of knowledge. As

mentioned in Section 5.1., the researcher first aimed for ten respondents that would participate in the interviews but only eight respondents accepted. It is the researcher's point of view that this might have affected the results. The researcher believes that if more respondents had participated in the interviews there could have been more exhaustive information gathered. The researcher believes that the objective concerning problems with integration of external data and the objective concerning possible future trends of integration of external data are the two objectives where more information could have been exposed. The other two objectives are believed to be relatively exhaustive. When reflecting on the respondents participating in the interviews of this work, some of them did not have very much experience in working with DWs. The issue of lack of knowledge concerning DW could be related to that DW is a relatively new concept and have not been used by the corporations for very long. Using external data together with DW is an even newer occurrence but there will most likely be an increase in the usage of external data as the knowledge about the opportunities external data can result in increase. (Singh, 1998; Inmon, 2003).

9.3. Ideas of future work

This work has in a comprehensive way outlined the acquiring and integration of external data into DWs. As the work also dealt with possible future trends, an interesting continuation of this work could be to outline the course of action for the Probability Theory Approach. This could be achieved by performing a case study. The researcher finds this approach particular interesting as the PTA could facilitate the integration of some external data not allowed to integrate today, due to different laws.

As this work was delimitated to only contain corporations within the financial area, an interesting issue to research could be if there is any differences in another business area concerning the acquisition and integration approaches used for external data and if another business area has the same problems concerning the integration of external data or if there are any different trends within other business areas.

The issue of expensive commercial ETL-tools could be another issue interesting to research. Kimball (1998) mentions that purchasing commercial ETL-tools is not always cost effective for smaller corporations. A reason not to purchase commercial ETL-tools was in this work mentioned to be that they are too expensive. A possible future work could be to research what factors must be taken into consideration to assure the cost effectiveness of purchasing commercial ETL-tools.

A summary of the ideas of future work is:

- To outline the course of action using the Probability Theory Approach
- To outline current approaches for acquisition and integration of external data into Data Warehouses in another business area and to compare the results with this work.
- To outline what factors must be taken into consideration to assure cost effectiveness of purchasing commercial ETL-tools.

References

Adleman, S. (1997) Data quality. In: J. Bischoff and T Alexander (eds.), *Data Warehouse: practical advice from the experts* (p. 122-134). New Jersey: Prentice Hall.

Agosta, L. (2000) *The essential guide to data warehousing.* New Jersey: Prentice Hall.

Andersen, I. (1998) *Den uppenbara verkligehten: Val av samhällsvetenskaplig metod.* Lund: Studentlitteratur

Bischoff, J. & Alexander, T. (1997) *Data Warehouse: Practical Advice from the Experts*. New Jersey: Prentice Hall.

Berndtsson, M., Hansson, J., Olsson, B. och Lundell, B. (2002) *Planning and implementing your final year project - with success!*, London: Springer

Chaudhuri, C. & Dayal, U. (1997) An overview of data warehousing and OLAP Technology, SIGMOD record, Vol.26, No. 1, pp.65-74. Available at Internet: http://www.acm.org/sigmod/record/issues/9703/index.html. [Accessed 2003.02.23]

Collet, Stacy. (2003) *Incoming.* EBESCO host, Computerworld, Vol.36, Issue 16, p34. Available at Internet: http://web8.epnet.com. [Accessed 2003.02.19].

Connolly, T. & Begg, C. (2002) *Database Systems: a practical approach to design, implementation and management* (3rd edition). Harlow: Addison Wesley Longman.

Devlin, B. (1997) *Data warehouse: from architecture to implementation.* Harlow: Addison Wesley Longman.

Gleason, D. (1997) Data transformation. In: J. Bischoff and T Alexander (eds.), *Data warehouse: practical advice from the experts* (p. 160-173). New Jersey: Prentice Hall.

Inmon, W.H. (1996) Buliding the data warehouse (2nd Edition). New York: John Wiley and sons.

Inmon, W.H. (1999a) *Integrating internal and external data,* The Bill Inmon.com library LLC. Available at Internet: http://www.billinmon.com/library/articles/intext.asp. [Accessed 2003.02.23].

Inmon, W.H. (1999b) *Data mart does not equal data warehouse, DM review.* Available at Internet: http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=1675. [Accessed 2003.02.28].

Inmon, W.H., Imhoff, C. & Sousa, R. (2001) *Information corporation factory (2nd edition).* New York: John Wiley & sons.

Jennings, M.F (2001) *Strategies for custom data warehouse ETL processing,* DM review. Available at Internet: http://dmreview.com/editorial/dmreview/print_action.cfm?EdID=3573. [Accessed 2003.02.14].

Kelly, S. (1997) *Data warehousing: the route to mass customization updated & expanded.* Chichester: John Wiley & sons.

Kimball, R. (1998) *The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses.* New York: John Wiley and sons.

Mancuso, G. & Moreno, A. (2002) *The role of OLAP in the corporate information factory.* DM review. Available at Internet: http://www.dmreview.com/master.cfm?NavID=198&EdID=5990. [Accessed 2003.03.01].

Marco, D (2000) *Building and managing the meat data repository: a full lifecycle guide.* New York: Johan Wiley & sons.

Niklasson, M. (2003) *Current and future application areas for external data in data warehouses.* Final year project number: HS-IDA-EA-03-410. Department of Computer Science, University of Skövde, Sweden.

Oglesby, W. (1999) *Using external data warehouses and warehouses to enhance your direct marketing effort. DM review.* Available at Internet: http://www.dmreview.com/master.cfm?NavID=198&EdID=1743. [Accessed 2003.02.23].

Salmeron, J. L. (2001) *EIS data: findings from an evolutionary study.* Journal of systems and software. Vol.64, Issue 2, p. 87-172. Available at Internet: http://www.sciencedirect.com. [Accessed 2003.02.23]

Sperley, E. (1999) *The enterprise data warehouse; planning, building and implementation*. New Jersey: Prentice Hall.

Strand, M. (2003) Incorporating external data into data data warehouses. In: B. cronqvist (eds.), Proceedings of the knowledge in organization (KIO) doctorial consortium, part 2, 5-6 February, 2003, Västerås, Sweden.

Strand, M. & Olsson, M. (2003) The hamlet dilemma on external data in data warehouses. Presented at the 5th *International Conference on Enterprise Information Systems* (ICEIS'03), 23-26 April, 2003, Angers, France.

Strand, M., Wrangler, B. & Olsson, M. (2003) Incorporating External Data into Data Warehouses: Characterizing and Categorizing Suppliers and Types of External Data". *Americas Conference on Information Systems (AMCIS) 2003*, August 4-6, Tampa, Florida, USA. (To appear)

Svenning, C. (2000) *Metodboken: Samhällsvetenskaplig metod och metodutveckling. Klassiska och nya metoder I IT-samhället.* Eslöv: Lorentz förlag

White, C. (1997) A technical architecture for data warehousing. In: J. Bischoff and T Alexander (eds.), *Data warehouse: practical advice from the experts* (p. 84-90). New Jersey: Prentice Hall.

Appendix 1 - Accompanying letter

Dear Sir/Madam,

First of all, we would like to thank you for the interest you are showing for our research projects and also the fact that you are willing to participate in the interview study targeted towards organizations incorporating external data. Your knowledge and experiences are vital for us, especially since there is very little written with respect to the incorporation of external data into Data Warehouses. In addition, it is our strongest belief that everything in theory must be related and compared with practice, for being truly interesting and contributing, and since our work is theory driven, we need you to be able to acquire the empirical perspectives.

In the following, we will shortly describe the aim of our research, important definitions (Data Warehouse and external data), expected results, how the interviews will be conducted and managed, and the interview questions. The definitions are included since we find it utmost important that we share the same definitions, for being able to include your answers in the analysis. The interview questions are included so that you have the possibility to reflect upon the questions in advance and thereby become more comfortable during the interviews.

The aim of our research projects

Our research projects are final year projects for the fulfillment of our Bachelor's degrees and they are coordinated under the research project MIDAS-EXIT (Multiple Integrated DAta Sources – EXternal data: Insights and Trends) administered by our supervisor Mattias Strand at the Department of Computer Science, Högskolan Skövde. The MIDAS-EXIT is a joint description of all the activities conducted under the fulfillment of our supervisors Ph.D. studies, aimed at characterizing and categorizing the opportunities offered and challenges experienced, when incorporating external data into Data Warehouses. The MIDAS-EXIT project focuses on two empirical perspective; suppliers of external data and consumers/users of external data. Our research projects have the common aim to outline the experiences and knowledge among the consumers/user organizations, with respect to external data incorporation into Data Warehouses. In detail, our projects are aimed at characterize and categorize current approaches when integrating external data into Data Warehouse and to classify and characterize different application areas for external data incorporated into a Data Warehouse.

Important definitions

We have chosen to use the following definitions in our work.

Data Warehouse

We have chosen the definition of W.H. Inmon as his definition is the one that is most widely used and the one most referred to.

Inmon defines a Data Warehouse as "*a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process.*"

Appendix 1 – Accompanying letter

External data

We have chosen to adopt the definition stated by Devlin as this definition is in line with our view of external data. Devlin defines external data as *"business data (and its associated metadata), originating from one business, that may be used as part of either the operational or the informational processes of another business"*

Expected results

Expected results of our work could be divided into two parts. First, to increase the knowledge regarding if there is a more common way of integrate external data into the Data Warehouse, if there is a reason why perhaps a peculiar method is used in front of the other and if that is the case, what are the factors taken into consideration. Secondly, to give an outlook regarding how external data is used in organizations and to be able to make a categorization of the different application areas that could be found. Also to give a hint toward application areas for external data not exploited today, which may be exploited in the future.

The interviews

The interviews will be conducted by phone and if permitted the interviews will be recorded. The motivation for recording the interview is that it gives us a higher freedom during the interview and allows us to concentrate on acquiring as much of your knowledge and experiences as possible. After the completion of the interview, the recorded material will be transcribed into text and the call recorded on the tape will be deleted. In addition, the transcribed text will thereafter be sent back to you, allowing you to reflect upon the transcribed material, to correct improper interpretations, and add new information that was not included during the interview. By using this approach, good research ethics are applied, since you, by this, have the ability to confirm the recorded material. Also, if you want to withdraw parts of the material, this is also possible. The approach chosen also contributes in the validation of the results, since as you are able to give remarks and perform changes on the raw material; the final material is already reviewed once by an area specialist.

The interviews are approximated to span between 60 - 90 minutes. For you to be able to prepare yourself before the interviews, the interview questions are included in this document (see next section). The interviews will be semi-structured, meaning that the interview questions will be used as a discussion framework, rather than a strict, sequence of questions. The idea of semi-structured questions is to have a common point of reference (the questions), but still have the freedom to elaborate on specific matters that comes up. Thereby, we will hopefully become more comfortable in the situation, as the interview is transformed into something more a like a coffee-break discussion.

Appendix 1 – Accompanying letter

The interview questions

Introduction

1. Name?
2. Position in the corporation? (As. IT-manager)
3. Name of the corporation?
4. Line of business? (Bank or insurance company)
5. Areas of responsibility within the corporation?
6. Experience of Data Warehouses?
7. What is your definition of a Data Warehouse?
8. What is your definition of external data?

The phase of identifying external data

9. What sources are used for the acquirement of external data?
10. Why are these sources chosen?
11. How are these sources identified?
12. What other sources do you know of?
13. Is finding a data suppliers for external data a problem?
14. Regarding your corporation, what does the future look like? Will other sources be used?

The phase of acquiring external data

15. In what way is data acquired from your external sources?
16. How much of the data you acquire is custom-made for your data warehouse?
17. To what extent are your suppliers able to tailor-make the data for your needs?
18. Can you ask your suppliers to adjust occasionally needed data in the way you need it?
19. What ways of distribution are you using when acquiring data?
20. Can you think of any new approaches for the acquirement of data in the future?

Appendix 1 – Accompanying letter

The phase of Integration external data

21. What approaches are used by your corporation when integrating external data into the corporation's data warehouse?

22. Which one of the approaches mentioned is most frequently used? (Please rank)

23. Do you know of any other approaches than the one/ones you use?

24. How is the external data stored in the Data Warehouse?

25. Do you think the integration of external data will increase?

26. Do you see any obvious trends concerning the integration?

The phase of using external data

27. In what different application areas are you using external data in combination with a Data Warehouse?

For each one of the application areas:

27a. What possibilities does external data open up for/ what is its contribution to this application area?

27b. Are there any problems related to this area of application?

27c. What different kinds of external data is included?

28. Are there application areas you reject due to insufficient funds or problems to big to overcome?

29. Will the usage of external data in your corporation increase in the future?

Closing

30. What general advantages do you see with the usage of external data in the Data Warehouse?

31. What general disadvantages do you see with the usage of external data in the Data Warehouse?

32. Is external data used for other purposes?

33. How much of the total external data used is integrated in the Data Warehouse?

34. Is there anything you would like to add?

35. Would you consider to participate in future interviews?

36. Would you like to take part of the material produced in this work?

THANK YOU FOR PARTICIPATING!

Appendix 1 – Accompanying letter

Contact information

Markus Niklasson
Telephone number: 0500-483759
E-mail address: e00marni@student.his.se

Carl-Fredrik Laurén
Telephone number: 0500-427745
E-mail address: b00carla@student.his.se

Best Regards,

Markus Niklasson and Carl-Fredrik Laurén
Department of Computer Science, Högskolan Skövde

Appendix 2 - Transcribed material

RESPONDENT 1

Namn?

NAMN

Tjänstetitel?

(tvekan) jag, min titel som i det här jobbet är att jag är systemägare för BANKENS s Data Warehouse och jag är även anslutningsansvarig.

Mellan?

För dom anslutningar som ska göras till Data Warehouse.

Företagsnamn?

FÖRETAGSNAMN

Typ av företag?

Bank

Och dina ansvarsområden, det har du ju nästan svarat på.

Ja

Har du några tidigare erfarenheter av datalager?

Nej

Hur länge har du varit inblandad i det här projektet?

Det är två och ett halvt år och det är inte längre ett projekt. Det var ju det från början, det var det innan jag kom. Nu mera så jobbar vi alltså i förvaltning. Men vi ansluter ju fortfarande.

Din definition på ett datalager?

Just det, det står ju här. Hur definierar ni datalager, oj då. Vi har ju en så tjusig definition, den skulle ju jag tagit fram. Kan vi vänta med den frågan så kan jag få återkomma med den. Jag vill gärna använda den definition vi har jobbat fram.

Ja det går jättebra, och den externa datan har du ju redan definierat då.

Ja det är information som kommer utifrån, inte inom banken utan utifrån banken.

Då kommer vi in på den fasen vi kallar identifiering. Vilka källor inhämtar ni extern data ifrån?

Inga.

Inga?

Nej.

Var får ni extern datan ifrån då?

Vi tar inte in någon extern data. Det vill säga vi i warehouset gör inte det, det är så här, vi tar in information från många olika system i banken, som i sin tur tar in extern data. Vi har till exempel ett kundregister som är helt uppbyggt på extern data.

Vet du vilka källor dom inhämtar extern data från?

Ja, jag känner ju till vad dom hämtar det ifrån, delvis, jag är inte helt hundra, dom hämtar ju från flera, men det är något som heter DAFA tror jag som dom får sina kunduppgifter från. Men som sagt vi som, BANKENS Data Warehouse vi hämtar inga uppgifter från externa system, eller externa leverantörer. (tvekan) Nu blev det kanske lite, första gången här när ni ringde till mig, så fick jag denna frågan och då svarade jag precis som jag svarade nu. Så då blev jag lite förvånad när jag fick frågorna för dom bygger ju helt på att vi skulle hämta in extern data.

Men för grejen va ju där att ni fick in extern data via dom systemen ni i så fall integrerar. Så får ni ju på ett sätt ändå in extern data.

Exakt. Men frågorna relaterar mycket till, om vi tar fråga 10 till exempel, varför hämtar ni just från de här källorna, det har jag ingen aning om, för jag vet inte varför källsystemet, det som är källsystemet för mig har valt att hämta sin information från just. Den externa källan.

Men vi får helt enkelt göra så att vi, försöker få fram svar på dom frågorna som det går. Det är ju begränsat med användningen överhuvudtaget med extern data i Data Warehouse i alla fall i Sverige.

Ja

Så vi är glada för den informationen vi kan få.

Ja, det är ju så pass nytt i Sverige.

Ja precis, vi gör ju en sådan här explorativ studie, för att se hur långt vi har kommit.

Ja nästa fråga då, som du kanske kan svara på då, vilka andra källor känner ni till?

(kort paus) Nja, egentligen inga som jag känner till, jag är ju alltså helt övertygad om att vi hämtar extern data men som sagt det är ju då andra källsystem i banken som inte jag då jobbar med. Så det kan jag inte säga att jag känner till, vad vi hämtar ifrån riktigt.

Hur tror du framtiden ser ut då? Kommer ni att lägga till mer källor?

Det är jag helt övertygad om att vi kommer att göra.

Ok kan du ge några exempel på vad du tror kommer att hända?

Jag kan säga så här att i dagsläget så (kort paus) det kommer nya regler kring hur krediter ska hanteras till exempel (kort paus) nu det senaste är det något som heter Basel2 vet inte om ni har hört talas om det men det handlar om kapitaltäckningsregler, där man, alla banker blir tvungna att sätta en rating på varje enskild kund. Det är ett sånt exempel, där finns det sådana externa, vad ska jag kalla dom för, leverantörer som har rating på kunder till exempel upplysnings centralen, UC. Dom har rating system på kunder, det är en typisk sån uppgift som jag är helt övertygad om att vi kommer att köpa (kort paus) och använda. Och stoppa in i vårat warehouse, som ett bidrag. Det är ett exempel (kort paus)

sen jobbar vi i BANKEN kanske inte som så många andra banker gör för vi jobbar ju då väldigt mycket utifrån vad vi vet själva om kunden, och vi har ju inte varit så intresserade av att segmentera kunder, klassificera kunder, utan varje kund är unik och ska beaktas därefter. Så därför har vi ännu inte diskuterat så väldigt mycket vad vi behöver för, vad vi eventuellt skulle behöva för information som finns utanför banken och då för att hjälpa oss att få en bild av kunden, men det finns och jag tror att det här kommer vi att vara intresserade av i framtiden.

Anskaffningsfasen då, om vi går in lite på den här då. Då ska vi se, då blir det hur ni hämtar data ifrån era andra system då? Då är allting redan egentligen skräddarsytt för erat datalager eller? Behöver ni transformera datan ni hämtar in från era interna system då?

Vi jobbar efter en modell som ni säkert har hört talas om eftersom ni redan har, ni har ju Devlin, IBM, som heter SSDM det har ni hört talas om va?

Hört talas om, kan den inte särskilt väl.

Det är ju då en sån där universal modell då för att definiera begrepp som finns inom finansvärlden, som man använder för att fastställa, för att tvätta information. Och vi har då en egen modell som vi har jobbat fram som heter BANKENS datamodell och utifrån den så jobbar vi tillsammans med anslutande system, analys och när vi har kommit fram till vilken information det är så skräddarsyr vi den enligt den här modellen. Så ingen, vi stoppar aldrig in någonting i vårat datalager som inte har gått igenom den tvätten.

Nej precis utan allting körs där igenom?

Yes.

Ok, och det sker online det här allting då, det är uppkopplat genom nätverk, ni kör det ingenting med att ni använder DVD eller så?

Vi får filer, ni ska veta att jag kommer från verksamheten, jag är inte från våran leverantörssida, inte från datasidan. Utan jag är, före detta banktjänsteman ute på kontor alltså som har börjat jobba internt med den här typen av saker så att ibland om jag uttrycker mig lekmannamässigt vad gäller sådana där datatermer, det kan hända att jag gör det.

Jaja, ingen fara, det är ingen fara alls.

Om man också tar på framtiden då, finns det några sätt för inhämtning då? Som det var just nu så hade ni en fil eller?

Ja, vi, nya sätt och nya sätt, vi har precis eller vi ska precis börja jobba med så kallade ETL verktyg, vi har precis köpt ett, och ska börja jobba med det och det blir lite annorlunda, (kort paus) egentligen, ja det är mera hanteringen av, med hjälp av det så kan vi ju jobba på det sättet att vi kan ta hem en hel fil från ett system och själv plocka ut den information vi vill ha, det är det det är till för, extract, transform, load. Så har vi inte gjort nu utan i dagens läge jobbar vi så att dom system som ansluter sig får fri genomgång och sen får de själv skapa filen. Och stå för innehållet och leverera den till oss och garantera innehållet

Ok, ja, men det är bra och så sätt för då kan man kalla dom undersystemen till er som era leverantörer i så fall då.

Ja.

Precis, men det är ju bra.

Var det svaret på frågan.

Ja det är bra där. Det förklarar ju ert sätt för framtiden om man säger så.

Ja, vi kommer väl, tror vi då helt gå över till ETL, men riktigt vad det innebär, det har vi inte riktigt, det kan vi inte säga ännu eftersom vi då inte har hunnit ens köra en pilot.

Ok, om man säger ska vi se, det här till exempel, hur lagrar ni er data i datalagret? Använder ni starscheman eller stjärnscheman och sånt där?

Jo det gör vi, vi använder, o nu var vi inne på min den här, alltså vi har ju tabeller, DB2, och dimensionstabeller, och sen jobbar vi med starschemas för att när vi då, för dom slutliga användarna så att säga, dom ställer frågor mot starschemas eller vi har ju fördefinierat frågor men vi har ju även användare som jobbar och bygger egna starschemas.

Ok, så dom bygger efter sina behov då?

Ja. Mot vissa givna tabeller, inte det stora, vi har den strukturen att vi har ett stort detaljlager och sen har vi då dimensionstabeller och dimensionstabellerna föder martarna.

Ok, absolut, jättebra. (kort paus) då ska vi se här, om vi kan plocka fram någon annan här. Det blir väl samma sak här då i så fall, det kanske nästan är en upprepning av den frågan Markus ställde tidigare. Det är om ni kommer att öka integreringen av extern data, det blir väl att, då ökar ni, tror ni att det kommer användas mer så kommer ni antagligen integreras mer data också kan jag tänka mig.

Ja det tror jag, alltså vi går från noll så det går ju fort. (skratt) Så att, jo, det är mycket troligt ja. Det är ju så att säga det här är ju fortfarande, vi (kort paus) jag var ju inte med från början, det här var ju ett projekt från början, jag tror att det påbörjades 97, så vi har hållit på i sex år, på en konferens med IBM igår så berättade BANK1 att dom fick genomslag först efter tio år, för sitt warehouse, vad man kan använda det till. Vi känner att vi har börjat få väldigt kraftigt genomslag nu, och i och med det så kommer ju också efterfrågan att öka också, då kommer också dom här kraven som gör att vi kommer att bli intresserade av att hämta in extern data.

Ja precis. Det låter bra det här.

Ja det är roligt.

Ja men det låter som ni är på rätt väg i alla fall.

Ja det tycker jag att vi är.

Det är ju väldigt olika hur lång tid det tar för respektive företag, det beror ju på, ja vad man har haft förut för erfarenheter av det då så att säga.

Ja jag tror alltså att det är väldigt olika vilken bransch man är i då. Bank, eller bankerna har väl inte riktigt (kort paus) sett nyttan kanske, så tydligt. Men nu tror jag att man är på väg att inse det.

Själva användningen då, det kanske du kan svara på lite mer då. Användningsområden för extern data, du nämnde tvättning, är det det enda som ni använder extern datan till?

Ja vad sa du, jag nämnde?

Tvättningen av kunduppgifter som sker i era system innan dom kommer in i datalagret. Är det det enda sättet som ni använder extern data?

Ja alltså extern data då (tvekan) det är det enda system som jag kan säga på rak arm levererar data som kommer ursprungligen extern ifrån. Annars har ju vi affärssystem som producerar information som har tillkommit inom banken. Jag menar inlåningsreskontran, utlåningsreskontran, alla dom tunga kärnsystemen i banken finns i warehouset. Alla dom tvättas.

Ja, känner du till några problem med det här att tvätta datan?

Ja visst, det är jättetungt att få organisationen med på noterna. Att förstå nyttan med tvätten. (kort paus) Det är liksom en inlärningsprocess för alla inblandade, så att det är ganska tung, tidskrävande process.

Finns det även användningsområden som ni avstår ifrån, nu har du ju talat om att ni kommer integrera lite men finns det andra användningsområden ni avstår ifrån om ni har otillräckliga resurser eller stora problem eller någonting?

Nu hör jag lite dåligt vad du säger, men om det finns användningsområden som vi inte.

Som ni avstår ifrån eftersom ni har för otillräckliga resurser eller för stora problem?

(tvekan) Ja alltså användningsområdet för ett warehouse, för oss som jobbar med det är det ju hur stort som helst. Det är ju allt från analys till statistik till ja, alltså det är, går ju använda precis hur som helst, det som är, det är oerhört kompetenskrävande, det kräver specialkompetenser bland dom som jobbar med det, för man måste förstå hur man ska bygga (kort paus) ihop det, för att användarna ska kunna dra nytta av det (kort paus) på bästa möjliga sätt. Man kan inte bara, jag vet inte om jag ska, det är väldigt lätt att bygga ett warehouse och stoppa in en väldig massa information men det som är svårt är att sen bygga vidare, att bygga nivåerna, plattformarna, varifrån användarna sedan lätt kan söka den information de är intresserade av och inse att dom kan vrida och vända på den på ett sätt så att dom verkligen kan analysera den, ur olika vinklar. Det har jag, anser jag att vi har inte tillräckligt med resurser för i dagsläget.

Ok, är det sådant som kommer då eller?

Hörru du man hoppas när man märker hur intresset växt, vi får, vi har alltså precis i dagarna nu ett enormt intresse ifrån bankens högsta ledning och då hoppas man ju liksom att det föds en förståelse för att, att det ska, man måste ju ha resurser för att göra. Men än så länge är ju vi väldigt, väldigt slimmad organisation, vi är väldigt få som jobbar med det här.

Användandet av extern data, kommer att öka i organisationen, det har du redan svarat på.

Ja

Ok då ska vi ta lite avslutande frågor då bara.

Vilka generella fördelar ser du med användning av just extern data i datalagret?

(suck) Ja (kort paus) fördelar det beror ju alldeles på vad det är, om man tar en sådan sak som jag nämnde för er, rating av kunder där man, någon redan har gjort det jobbet utifrån vissa givna parametrar, en klar fördel. (kort paus) Nä det vet i sjufåglarna, det har jag inte funderat så mycket på. Det är ju alltid så att när (kort paus) när det kommer, när man ska göra den typen av bedömningar utav saker och ting då är det ju alltid både tidskrävande och ofta en väldig massa mängd data som ska samlas in då, på rätt sätt, analyseras, då kan man tjäna mycket tid på om det finns färdigt. Men sedan som sagt så beror det ju alldeles på vilket område vi pratar om. Jag har ju som sagt väldigt lite erfarenheter av det, så det, jag är väl inte så bra på att svara på den frågan tror jag.

Ok, ser du några generella nackdelar då? Med extern data?

Ja det man kan säga direkt är att det man har själv och har stoppat in och vet hur det är sammansatt ner till minsta beståndsdelen har man ju kontroll över. Det har man ju inte av extern data, utan där får man ju lita på, på kvalité. Det är möjligtvis nackdelen.

Är det så att ni använder extern data till något annat i företaget, förutom det här som du berättat med, som kommer in i Data Warehouset. Är det andra system som också extern data som inte kommer in i Data Warehouset?

Ja det är jag helt övertygad om. Ja det är det, det finns det.

Har du någon uppfattning om hur mycket av den här extern datan som verkligen kommer in i datalagret?

Ja alltså om man nu säger att hela vårt kundregister är uppbyggt på extern data så är ju det en mycket, det är en betydande del, det är ju alla våra kunduppgifter i princip. Så det är ju en betydande del.

Något i övrigt du har att tillägga?

(lång paus) Nej, jag tycker jag har försökt att svara uttömmande. Nej jag har inget speciellt att tillägga. Nej.

RESPONDENT 2

Ditt namn?

NAMN

Arbetsuppgifter?

Kundanalytiker.

Vad har du för ansvarsområden?

Mina ansvarsområden är ju (kort paus) det är ju kundanalys, hitta målgrupper för alla kanaler. Hur vi ska bearbeta kunderna, erbjudande till kunder, se var du hittar lönsamhet någonstans.

Vad har du för erfarenheter av datalager? Har du jobbat med något tidigare?

På datalager? Min bakgrund är att jag jobbade på FÖRETAG i 5 år med marknadsundersökningar. Där är ju inga jättedatalager, utan där är det ju snarare platta datafiler. Och på slutet när jag jobbade där så började det börja komma upp så där med att fler och fler blev intresserade av CRM och koppla undersökningsdata till sina datalager. Det tyckte jag lät intressant så då bytte jag till BANK 1. Och där kom jag då i kontakt med stora datalager. Där utvecklade vi också det befintliga datalagret, så vi bytte version av det, från det första kunddatabasen till en databas som var betydligt större då.

Vad är eran definition på ett datalager?

Jo, på BANKEN här, det vi kallar vårt datalager är det som föds från våra källdatasystem, värdedatasystemen, och där föds det med, dels bank, transaktionsdata och kunddata och sen föder vi också med externdata så att det blir en kunddatabas, vill vi kalla det för.

Hur definierar ni externdata då?

Ja, det är ju då all data som vi inte har själva egentligen, och (kort paus) data som vi köper in från någon extern leverantör.

Det låter ju bra. Då går vi in på det här med källorna då. Vilka källor inhämtar ni externdata ifrån?

Ja, marknadsanalys har vi ju, dom har ju något system som kallas Mosaik som är klassificering av privatpersoner efter postnummer. Det använder vi och sen på företagssidan är det Dun and Bradstreet då. Och sen är det då adressuppgifter då från SPAR.

Vad är det för data ni hämtar från Dun and Bradstreet? Vilken typ av data är det?

Det är i stort sett allt om företaget som, dom gånger vi hämtar det så handlar det om att lägga upp en kredit på företaget, så det är kreditdata och bokslutsdata och så vidare.

Finns det någon anledning till att ni hämtar just från de här källorna, finns det andra källor som skulle kunna ersätta de här?

Ja, det finns det. Ja just det, vi har ytterligare en, vi har ju samma sak på personsidan, på privatsidan hämtar vi från UC, upplysningscentralen då. Det är ju också (kort paus) data som är nödvändig för att bestämma om personen är kreditvärdig eller inte. Och ja, det finns andra. Som Dun and Bradstreet och UC dom skulle vi kunna switcha, eller kasta ta ut båda två och ta ett tredje företag, det finns just på kreditsidan finns det ju, ja fyra-fem stycken att välja mellan i Sverige. Men vi har valt dom största, på privatsidan är UC störst och på företagssidan är D och B störst. Så att det (lång paus).

Det är anledningen till att ni har valt dem då?

Nej, det är väl, (kort paus) D och B tycker vi håller en, dom ser lite annorlunda på det så det är kvalitén. UC har väl varit så att (kort paus) ja dom var störst och var villiga att hjälpa oss när vi startade upp.

Hur identifierade ni dessa källor, hur hittade ni dem?

Hur vi hittade dem? Alltså företagen eller källor, eller vad menar du där?

Att ni förstod att de hade data som ni ville åt?

Ja just när det gäller kreditsidan så var det ju inte så svårt, det var ju frågan om, jaha, hur löser vi det här. Och om upplysningscentralen ägs ju av bankerna så att det var (kort paus) i stort sett internt lite grand då. Alla känner ju till dem, däremot har vi ju Mosaik, som är mer i kommunikation mot marknaden, den är ju svårare att hitta. Fast dom söker ju ofta upp en, där är det ju dom som har hittat oss, och berättat vad dom har och kan tillföra.

Finns det andra källor som ni känner till?

Ja, just på marknadssidan är det ju mycket TEMA har ju mycket, där vi inte köper något speciellt idag och ja, vad finns det mer? Ja på företagssidan finns det ju också, där finns det ju en uppsjö av stora och små som säljer företagsinformation.

Tycker du att det är svårt att hitta leverantörer av externdata?

Nä, det tycker jag väl inte, det är i sådana fall (kort paus) att hitta dom som har datan är inte svårt, det svåra är att få klart för sig vad man får göra med externdata. Där kan det vara riktigt lurigt. Dom säger ju att, ja det här kan vi göra, men då får ni inte koppla externdatan till ert datalager, utan då måste vi göra vissa grejer, ni skickar data till oss och sen kan vi göra det. Så att det, där kan det vara lite smålurigt.

Hur ser framtiden ut då, kommer ni använda andra källor tror du?

(Bandspelaren slutade fungera)

Vilka generella fördelar ser du med användningen av externdata?

Ja det är framför allt att kunna ta beslut om, ja både kredit och var kunden bor och hur kunden, att lära känna någon snabbt utan att egentligen ha egen information om det.

(Bandspelaren slutade fungera igen)

(Respondenten kontaktades igen för kompletterande frågor i och med bandspelarproblem)

Anskaffningsfasen. Hur hämtas datan in från era externa källor?

Som sagt, vi hämtar in på tre sätt kan man säga. Dels direktkopplade mot extern datan, dels via flata filer som vi läser in med batchjobb eller ja via körningar och sen data som, ja vad sa jag förut då, hade ju en tredje också. Och sedan då vi går genom värdesystemen då, och läser in där först.

Och ni hade även blandad prenumeration och on-demand va?

Ja just det.

Hur mycket av den data ni hämtar in var skräddarsytt efter ert datalager?

Ja just det, det var ju utformningsmässigt i stort sett allt skräddarsytt för att passa antingen direkt i datalagret eller via värdesystemet. Men innehållsmässigt ingenting. Och sedan från värdesystem till datalager är det skräddarsytt däremellan.

I vilken utsträckning kunde era leverantörer skräddarsy datan efter era behov?

Det ska jag ju säga, att det gör dom oftast tillsammans med våran egen IT-avdelning, så att vi får ju skräddarsytt och oftast så tycker jag att det är inga problem för dom utan att dom hjälper gärna till men tar ju då betalt för det.

Kan ni begära från era leverantörer att dom iordningställer annan data efter era behov?

Ja det skulle vi kunna göra, om det är något vi behöver ja, så skulle vi kunna få till att antingen be dom eller metodvis via våran egen IT då.

Och de distibutionsvägarna som ni använder för inhämtning av datan då?

Ja det är ju dels direktkoppling då, vi använder oss av FTP i vissa delar och sedan är det ju e-post och vi skickar filer fram och tillbaks.

Och du nämnde någonting tidigare med att ni skickade iväg en fil som iordningställdes hos leverantören.

Ja just det.

Skulle du kunna dra lite snabbt om det igen?

Om dom gånger vi skickar iväg. (kort paus) Nu kommer inte jag ihåg exakt, om vi nu pratade om marknadsdata eller vad det var. Där finns det ju tillfällen vi skickar iväg en fil, dom iordningställer den, vi får tillbaks den och läser in då i datalagret med ny information.

Såg du några nya sätt att inhämta data i framtiden?

Nej det gjorde jag inte. Som jag sade, här använde vi oss av vår egen IT-avdelning om vi skulle vilja komma på någonting.

Ok, då går vi över på fasen integrering här. Vilka tillvägagångssätt använder ni er av vid integrering av extern data i företagets datalager?

Det var ju där det var tre. Där använde vi direktkoppling och sedan använder vi då jobb som går via värdesystemen och sedan använder vi ju direkta (kort paus) ja där vi kör lagringsjobb in i systemet då, och det var ju den ordningen på dom också vad gäller rangordningen så att.

Känner du till några andra tillvägagångssätt?

Vi tittar lite på det, det är när vi ska läsa in (kort paus) data som vi får då och då så kör vi det via IT med batchjobb där vill vi ha ett interface som vi kan läsa in själva då direkt i datalagret då som är samma sak men det sker på lite annat sätt då.

Hur lagrades datan i, den externa datan, i datalagret?

Den var ju lagrad helt integrerad så att man ser inte skillnad på vad som är externt och vad som då är internt.

Kommer ni att öka integreringen av extern data?

Nej, det blir ju svårt där eftersom jag tycker att den är integrerad.

Du nämnde någonting tidigare med integrering mot systemen då.

Ja just det, i datalagret sker, behövs det nu ingen utökad integrering utan det är integrering med övriga system där man vill se data från datalagret som behöver integreras för att även kunna se extern data där.

Och såg du några tydliga trender på tillvägagångssätt beträffande integreringen?

Nej det var ju lite så där att det integrera med hjälp av sannolikheter istället för integrera direkt på personer.

Och kan du dra lite snabbt vad det gick ut på igen?

Ja om vi på, får in extern data som ligger på personnivå (kort paus) så har man oftast data avidentifierade personer. Vi har ganska mycket information om personen och sedan har vi information om de här personerna i vårat datalager, så kan man då uppbygga en modell för att få över dom här, den information man vill ha så använder man gemen, den information som är gemensam på både extern- och intern-datan för att skaffa sig en sannolikhet för vem det hör hemma på, den extern data man inte har då.

Ok, då var det bara ett förtydligande på fråga 21 då igen. Dom tillvägagångssätten, de var som du nämnde förut att ni, (kort paus) integrerade, att ni läste in extern data direkt in i datalagret.

Ja

Och så tog ni in det via de operativa systemen.

Ja

Vilket var det tredje?

Det var ju, att vi (kort paus) ja externa filer som vi lägger över hit och läser in dom manuellt i stort sett då.

Ok, jättebra.

Användningsområdena då igen. Det var alltså kredit, adressuppdatering och marknadsföring då som var dom stora?

Det stämmer.

Skulle du kunna dra lite om dom här möjligheterna och bidraget är för just marknadsföring?

Extern datan har ju mycket att tillföra, när kunden är ny när vi inte känner till vem vi har att prata med. Sedan varefter man ja, lär känna personerna i datalagret, får mer och mer information om dem så minskar extern datan i, ja hur viktig den är och då är intern datan då och transaktionsdata det vi använder framförallt då.

Ok. Fanns det några problem, du pratade om någonting om att det krävdes tillstånd eller någonting sådant?

Just det. Det är ju också, det som är lite lurigt här, så fort det är på personnivå, företag är inte riktigt lika svårt men när man ligger på privatpersoner, vad man får göra och inte göra, hur man får samköra och inte samköra och vilken data man får spara och hur länge man får spara och så vidare.

Ok. Sedan nämnde du något problem med adressuppdateringen också. Att det var något internt problem som ni hade.

Ja just det. Det med?

Flera adresser.

Ja själva adressen är ju korrekt när man begär en uppdatering men att få alla system och ta samma adress och om dom nu ska ha samma. Att en person kan ha flera adresser och så vidare. Där har man ett problem men det är ju internt att styra över adresserna och vilken adress som ska finnas i vilket system.

Fanns det några användningsområden som ni avstår ifrån?

Ja det var ju lite om man kommer in på ekonomiska data där vi ser att, där man kanske skulle kunna utveckla det i framtiden där vi inte gör idag.

Kommer användning att öka tror du, av extern data?

Ja det tror jag och det tror jag också att det kommer att dyka upp fler leverantörer som förstår vad man kan göra med externdata och dom kommer försöka skräddarsy den mer.

Ja har egentligen bara en kompletterande fråga som inte står med och som vi inte diskuterade förut. Vilka är dom vanligaste problemen vid integreringen?

(kort paus) mellan, ja.

Har ni något sådant som är just vanligast eller så där? Generella problem vid integreringen.

Nja det är väl lite sådär att (kort paus) Nej. Som sagt, att veta hur bra extern datan matchar egentligen mot den interna. Att veta hur stark kopplingen är och, men sedan är det ju själva integreringen inga större problem. Sedan är det ju sådan här småsaker som att, ja hur gammal är extern datan, men det bygger man ju upp själv. Det är lite sådant man kan ha problem med i början innan man kommer på att det ska nog finnas ett datum också på all extern data. Men inga problem annars så att.

Ok, jättebra. Tack så mycket att du ville hjälpa oss direkt.

RESPONDENT 3

Namn

NAMN

Tjänst titel?

Systemförvaltningsansvarig

Företag?

FÖRETAGSNAMN

Dina ansvarsområden?

Ja det är kundsystemet och ett länsbonussystem som det heter som ingår i kundsystemet.

Har du några tidigare erfarenheter av datalager?

Nej det kan jag inte påstå.

Hur definierar du datalager?

Ja, jag definierar det som en mängd information som är relativt lättåtkomlig (kort paus) och externa data det är sånt som vi får utifrån alltså, definierar vi som inte kommer från våra egna datasystem då.

Identifieringsfasen då, vilka källor inhämtar ni data ifrån.

Ja det är då från SPAR, ja det är mitt system, vi hämtar ju (kort paus) till FÖRETAGET hämtar vi från SPAR, Basun och UC.

Ok, varför just dessa källor? Finns det några andra källor som skulle kunna ersätta dom? Som du vet om?

Nej det enda som kan ersätta dom är val av leverantör då..

Ok, kan du ge exempel på vad det finns för andra leverantörer

Ja på basunen och företagsinformationen finns det ju Semainfodata och så finns det Statistiska centralbyrån.

Hur identifierade ni de här källorna som ni har? Hur hittade ni dem så att säga? Att dom har informationen som ni vill ha.

Ja den där frågan förstår jag inte riktigt.

Ja dom här SPAR, basun och UC. Hur kom ni fram till att det var dom här källorna som ni ville ha in data ifrån?

Ja det visste vi ju var dom har för slags information. Jag vet inte, har man jobbat ett tag så vet man väl det.

Ok, är det svårt tycker du att hitta leverantörer av extern data i allmänhet?

(Lång paus) Nej, jag kan inte svara helt säkert på det, inte den här tycker jag inte har varit något svårt, inte i de här fallen. I andra fall kan jag nog inte ge svar på det.

Hur ser framtiden ut? Tror du att andra källor kommer att användas i eran organisation?

Nja, vad menar du med andra källor? Det är också en sådan här fråga som jag inte…

Att ni tar in från, nu tar ni in från SPAR, UC och basun som du nämnde om ni kommer ta in från även andra företag.

Nej det tror jag nog inte faktiskt. Vi tar ju i och för sig in från bilregistret till andra försäkringssystem.

Det är inget som kommer in i erat Data Warehouse? Eller kommer det också in i ert datalager?

Det kommer in via våra försäkringssystem sådär men inte direkt för datalagret. Jag kom på en sak till, däremot kanske vi skulle kunna hämta in telefonnummer (kort paus) det kanske vi skulle behöva hämta externt så småningom. För det brukar vara svårt att få tag i aktuella telefonnummer på våra kunder.

Ok, för anskaffningsfasen då, hur hämtas datan in från era källor. Prenumererar ni på datan eller?

Ja vi prenumererar, ja gör prenumeration på det och sen görs det via filöverföring då.

Ok, är det online då eller laddar ni ner filen?

Ja, på ett ställe så hämtar vi från webbhotell som dom kallar det för, då ligger den där och så får vi ett meddelande när. så är det liksom någon snurra som går som talar om att nu finns det något att hämta och så startas ett jobb igång då o hämtar.

Hur mycket av den data ni inhämtar är skräddarsydd efter erat datalager?

Ja det är två stycken, det är både basun och UC som är skräddarsytt för oss. Vilket jag tycker är lite dumt för att det är bättre att få in den informationen och sen skräddarsyr vi den själva för då har vi chans att bygga vidare sen.

Ok, vet ni om era leverantörer kan skräddarsy annan data efter era behov?

Nä, inte annan data mer än vad dom har i sin grunddata.

Dom kan inte samla in eller iordningställa annan data?

Inte vad jag känner till, inget som jag känner att vi skulle behöva liksom.

Ser du några nya sätt för inhämtning av data i framtiden eller? Hur tror du det kan se ut där? Kommer ni att fortsätta ladda ner det?

Ja i den närmaste framtiden tror jag att det kommer vara så i alla fall, jag tror inte att vi kommer ha några sådana här online-transaktioner på det här nej. Det ser jag som ganska långt framöver. Och sen ser jag inte sånt där jättestort behov utav det.

Ok, då på integreringsfasen, vilka tillvägagångssätt använder ni er av vid integrering av den externa datan i ert datalager.

(Lång paus) Ja, jag vet inte vad du menar med integrering. Vi tar in det här och sen tittar vi om dom här, att vi har dom här kunderna och då gör vi uppdateringen. Det är data som vi får då.

Ja integrering då det är mer liksom menar jag hur ni beblandar den då, sammankopplar den med den interna datan ni redan har.

Ja visst, vi bygger upp dem i våra kundregister där vi behöver information då som vi får.

Det är ett sätt som ni gör, ni bygger upp det så.

Ja förutom dom här företagen som vi har, dom har vi tillåtelse att lagra dom(kort paus) har vi då och försöker få om kunder som vi (kort paus) så att vi får in samtliga företag, samtliga som inte är enskilda företag, dom får vi (kort paus) och dom har vi rätt att kundbearbeta, försöka göra säljkampanjer mot då.

Ok, hur lagras den externa data i ert datalager?

Nja vi hämtar alltså de termer och de fält vi behöver uppdatera. Sen när det gäller, sen är det vissa lagar och bestämmelser på personer, då får vi inte ha dem lagrade hur länge som helst, men ett par veckor sen ska det förstöras då.

Men ni lagrar dem alltså i själva databastabellerna direkt då?

Ja det kan man säga, samma dag eller dagen efter så kör produktionskörningarna igång.

Tror du ni kommer att öka integreringen av externdata, att ni integrerar den datan ni har mer?

Ja det tror jag för det finns önskemål om det. Få in mera information om företagen

Ser du några trender i tillvägagångssättet för att integrera extern data i datalagret.

Nej det gör inte jag i alla fall.

Ok, då går vi in på användningsfasen då. Vilka tillämpningsområden använder ni externdatan med erat datalager? Om man säger så, vad har externdatan för roll i ert företag.

Ja, det är mer som information och sen externdatat får rätt namn och adress på våra kunder. Och sen från UC är det rätt ekonomisk information.

Ser du några problem med någonting inom de här användningsområdena?

Nja det ligger nog mera inom huset och tekniken**.**

Finns det några användningsområden ni avstår från som ni har otillräckliga resurser för eller för stora problem eller någonting?

Nej

Tror du att användandet av externdata kommer att öka i framtiden. Att själva externdatans roll kommer att bli större om man säger så? Mer värdefull?

Ja jag tror det om vi ska utöka marknader och sånt där så tror jag vi kommer behöva det. Att man ska kunna bearbeta, ja rätt kunder för oss då va. Man gör en viss inriktning.

Då har vi bara några avslutande frågor också. Vad ser du för generella fördelar med användning av extern data i datalager?

Nä, men ofta när man tar från de här leverantörerna då vet man att man får mera korrekt information.

Ser du några generella nackdelar med externdatan?

Ja det kanske tar lite längre tid innan vi får uppdaterat det.

Använder ni externdata till något annat i erat företag förutom i data lagret så att säga?

(Kort paus) Ja, det gör vi till våra försäkringssystem, använder vi dem.

Har du någon uppfattning om hur stor del av den totala mängden extern data som används och integreras i datalagret?

Hur stor del? (Kort paus) delmängd, ja. (lång paus) jag är inte riktigt säker på din fråga där.

Alltså, utav all den extern datan ni tar in hur mycket kommer verkligen in i datalagret så att säga?

Jaha, ja det är olika. När det gäller SPAR så kanske vi får in (kort paus) ja det är knappt hälften som vi får in. Och på övriga UC, Basun så använder vi allting, för att SPAR får vi in hela rikets alla förändringar en gång i veckan. Och då uppdaterar vi bara våra kunder och det är bara dom vi får uppdatera resten får vi inte ha lagrade ju.

Är det något övrigt du har att tillägga.

Nej det tror jag inte.

RESPONDENT 5

Namn?

NAMN

Tjänstetitel?

kundanalytiker

Företagsnamnet?

FÖRETAGSNAMN

Vilket typ av företag?

Försäkringsbolag och bank

Dina ansvarsområden?

För mig? Ja vad ska man säga, analys av lönsamhet på kunder och kampanjer, uppföljning på kundutveckling, kampanjer och så vidare, strategiarbete vad gäller kund (kort paus) försäljningsstrategier, kundutvecklingsstrategier kan man då säga.

Har du någon tidigare erfarenhet av datalager?

Sedan FÖRETAG, nej. Inte mer än det jag jobbat med nu den sista tiden.

Hur länge har du jobbat då?

På FÖRETAGET? I sju år.

Och med datalager hela tiden?

Nej, i sammanlagt ett år ungefär, ett och ett halvt år kanske, jag har varit borta en period och varit mammaledig. Så där, det ganska nytt hos oss med datalager, det är väl ett par, tre, fyra år gammalt bara så att det är inte förrän nu det har landat egentligen, ordentligt.

Hur är då din definition av ett datalager?

O ja, det var ju dom där svåra. Vi har egentligen ingen uttalad definition av ett datalager tror jag. Jag pratade som sagt med dom andra här som har varit med från början och byggt och våra konsulter och allt, men vi kan inte säga att vi har någon direkt datalager mer än att vi det är ett ställe där vi samlar bestånds och försäljningsdata egentligen om våra olika produkter, för att kunna jobba med den ur analyssynpunkt och uppföljningssynpunkt på ett vettigt sätt. Så vi har ingen direkt definition så.

Men den funkar ju så?

Ja

Och definitionen av extern data i så fall?

Det kan man väl säga, vi definierar det som sånt som så att säga inte vi har i våra egna produktsystem utan som vi köper eller får någon annanstans ifrån. Det är extern data för oss som vi behöver för att kunna, ja, jobba på ett bra sätt men där vi inte sitter på informationen själva utan någon annan gör det bättre.

Då kommer vi in på nästa fas, identifieringsfasen. Vilka källor inhämtar ni extern data ifrån?

Vi hämtar från SPAR och från Posten, och sedan Marknadsanalys, ett företag som vi samarbetar med.

Ok, varför hämtar ni från dessa källor? Finns det några andra källor som skulle kunna ersätta dom här källorna?

(tvekan) Det gör det nog säkert, SPAR vet jag inte men det är ju dom som så att säga passat oss bra, vi har, SPAR har ju, det är ju adresser framförallt vi köper av dom och det är ett bra system som funkar med oss och vi har haft avtal sedan tidigare och det finns säkert andra leverantörer men det är dom här vi har valt just nu. Marknadsanalys finns, har konkurrenter men (kort paus) där tror jag inte att det är någon som kan ersätta just den modellen som dom har. Vad gäller Posten så är det ju postnummer som vi köper därifrån och det är ju, det finns nog inte heller någon annanstans. För att jag tror att dessa är dom största leverantörerna och dom som, bästa för vår del på dessa sakerna.

Hur identifierade ni dessa källor? Hur hittade ni att dom hade information ni vill åt?

Ja, vad gäller marknadsanalys så är det en tidigare marknadschef här som kom in med det samarbetet så att säga, han hade jobbat med dom tidigare så att han visste vad dom kunde och så vidare, så det var den vägen. Vad gäller SPAR och Posten så är det också på erfarenheter i huset, att man har jobbat med dom tidigare och att dom har väl så att säga varit lite aktiva själva och hört av sig. Försäkringsbolag behöver ju ofta denna typen av tjänster, så att man kan nog säga att det är folk i huset här som har talat om att informationen finns på dessa ställen, tidigare erfarenheter.

Vilka andra källor känner ni till? Som ni inte använder då.

Vi, det finns ju fastighetsregister, bilregistret, centrala bilregistret till exempel. Vi har ju sådana här kreditupplysningstjänster, UC och sådana saker och det är ju tjänster som vi använder men ingenting vi kör med i DW:et, utan det ligger i produktsystemen och sådana saker men ja, (kort paus) det är väl dom jag kan säga så här, som jag kommer på på rak arm. Men det finns ju väldigt mycket externa källor egentligen, tror jag, man kan väl få veta det mesta.

Är det svårt att hitta leverantörer av extern data?

Nej. Vi har inte upplevt det så. Inte för våra behov ska jag väl säga.

Hur ser framtiden ut då tror du, kommer ni att använda andra källor också?

Jag som jobbar med det jag gör hoppas på det för att det kommer göra att vi kan bli ännu bättre på att hitta våra lönsamma kunder och screena bort dom som vi inte är så intresserade av att jobba vidare med och så vidare. Så, men jag vet inte, men jag hoppas att vi kommer använda andra externa källor.

Vad hoppas du mest på att det ska komma in för andra källor?

I mitt fall hoppas jag att man kan, och det är ju lite begränsat i och med PUL som har kommit men mer information om saker som samboförhållanden med inkomster och sådana saker. Det kan man ju aldrig få i detalj men man kan få uppfattningar om sådana

saker och sedan riskbeteenden, det här med kreditbeteenden eller vad man ska säga, betalningsanmärkningar och sådana saker att det kan användas på ett mer effektivt sätt när vi klassar våra kunder så att säga. För min del skulle det vara väldigt användbart.

Ok, då kommer vi in på nästa fas som gäller anskaffning då. Hur hämtas datan in från era externa källor? Är det, prenumererar ni på den eller är det så att ni har sådan där on-demand tjänst på det. Att ni säger till att nu behöver vi mer data.

Vi har båda delar, vad gäller SPAR så uppdateras det en gång i månaden. Marknadsanalys är mer on-demand, vi har en fil med folk här och vi behöver veta detta om dom och så vidare, uppdatera, så att det är på förfrågan. Även Posten är sådant som vi gör på förfrågan. SPAR har vi ett avtal med så det löper, kommer löpande.

Ok, hur mycket av den data ni hämtar in är skräddarsytt efter erat datalager?

(tvekan) Dom flesta, vi har ju definierat vilka fält vi vill ha så att säga, vad som ska uppdateras hos alla dom här, så att det är väl det mesta kan man säga. SPAR har ju mycket mer information än vad vi tar (kort paus) så att säga. Så det tror jag nog att det mesta är. Om det är så man menar med att det är skräddarsytt efter datalagret.

Ja precis, eller hur mycket ni, andra sådana sätt att få det ni kanske bara får en stor fil och sedan får ni göra i ordning den själva och lägga in ert datalager. Men här låter det som att allt det här arbetet, transformeringen och det här plocka bort onödig data som ni inte ska ha är redan gjort när ni får den, det är det som är liksom att den är skräddarsydd då.

Ok, ja du menar så (kort paus) nä man kan säga så här, det kommer från SPAR, så kommer det en stor fil egentligen och där uppdateras förändringar bara och det är ju skrivet någonstans då att uppdatera detta och detta så att när (kort paus). Marknadsanalys och Posten är väl skräddarsytt och när det gäller SPAR så är det väl inte det då utan där har vi då satt en regel om detta eller detta har hänt på den här personen så ska det uppdateras med adress och så vidare.

Vet du i vilken utsträckning era leverantörer kan skräddarsy data efter ert behov. Kan ni, till och med SPAR skulle dom kanske kunna gå in och göra det åt er om ni hade det behovet?

(kort paus) Jag tror inte vi skulle släppa in dom. Men (kort paus) ja, nej egentligen tror jag inte att dom gör det, men det vet jag inte (kort paus) faktiskt.

Kan ni begära att era leverantörer i ordningsställer annan extern data efter era behov?

Ja

Ok, vilken distributionsväg använder ni er av vid inhämtning av data?

Med distributionsväg, är det att det går på fast lina då eller, eller kommer på skiva?

Ja, är det cd-rom eller är det on-line, eller får ni någon fil på något sätt?

Ja det (tvekan) vissa saker kommer på skiva, cd-rom, Marknadsanalys till exempel brukar komma på en cd som vi trycker in, en textfil på något sätt som vi läser in. On-line är mer,

SPAR är ju on-line, det kommer över lina bara (kort paus) så det är väl dom två vägarna egentligen (kort paus) som det går.

Ser du några nya sätt för inhämtning av data i framtiden?

Ja vad ska man säga (kort paus) vet ej, får jag nog säga där. Det blir väl säkert ännu mer över Internet och så där kanske, men ja, vet inte.

Ok, då kommer vi till integreringsfasen härefter då. Vilka tillvägagångssätt använder ni er av vid integrering av extern data i företagets datalager?

Vi sitter ju, det är SQL-server så att vi laddar filer via DTS.

DTS vad är det?

DTS-paket som man kör, det är textfiler egentligen tror jag som man kör in. Så att vi laddar så, eller vilken, hur, vad är frågan egentligen annars?

Frågan som jag menar den, är att ni har ju intern data i era system och när ni lägger på den externa datan, hur mappar ni ihop det där då liksom hur får ni in den.

Jaha ok, det är ju någon unik nyckel. Så att personnummer är det ju ofta, när det gäller adresser och så vidare så är det ju den att det matchas upp helt enkelt. Hitta personnumret och fyll på med den information som det ska vara, så att det är ju det, det är ju fördelen för oss jämfört med många som inte är bank eller försäkringsbolag skulle jag tro, att det finns ett personnummer som är unikt och när det gäller markandsanalys så är det postnummerkoder som man får för att lägga på mosaiktyper (kort paus) så att det är alltid någon unik nyckel som man matchar på, information.

Jättebra, och det är det sättet, du känner inte till flera sätt?

Vad skulle det kunna vara? Ge något exempel bara så.

Ja på ett sätt, man kan ju, först kan man läsa in extern data i ungefär som en, till något operativt system där ni sedan på det sättet ni utvinner intern data och lägger in i Data Warehouset att ni behandlar även externa datan på det sättet och suger in den där också så att allting transformeras och tas in i Data Warehouset. Förstår du vad jag menar?

(tvekan) Ja, det tror jag att jag gör (tvekan) när det gäller det som går in i Data Warehouset så ja, nej jag tror inte, för vi har ju ett antal produktsystem som så att säga levererar information där så att säga den här försäkringen och det och det som tankas in, nattetid. Och sedan har vi så att säga en del där kunderna ligger och det är ju där den externa datan går in egentligen, så att, nej jag kan inte se att det är på något annat, vi matchar så på personnummer, just så där på rak arm, nej det tror jag inte.

Känner du till några andra tillvägagångssätt än dom ni använder er av då?

(tvekan) Nej det kan jag väl inte säga att jag gör då. Jag är nog fel person just för dom här frågorna kan jag säga. När du nämner det du nämner nu så känner jag inte igen att jag har hört att det är någonting sådant som vi gör egentligen till DW:t utan det tror jag inte.

Tror du att ni kommer öka integreringen av extern data? Alltså just graden av integrering, att ni kanske integrerar det hårdare eller det kanske är så pass integrerat som det går att få det.

(lång paus)

Det låter ju nästan som det när ni matchar varje nummer med person eller nycklar och sådant.

(tvekan) Jag vet faktiskt inte. Om det är så att det finns ett effektivare eller bättre sätt som funkar säkerhetsmässigt och allting så kanske det skulle kunna gå men just nu så är det ingenting som det har talats om i alla fall.

Då har jag bara en avslutande fråga på integreringsfasen här, det är om du ser några tydliga trender men det har du ju i stort sett svarat på här, eller nej det har du inte men jag vet heller inte om du kan göra det.

Nej jag tror inte det kan man väl säga. Jag är ju som sagt en vanlig användare om man säger.

Då passar ju nästa område bra då kanske, användningsfasen.

Ja jag hoppas det.

Inom vilka tillämpningsområden använder ni extern data med erat datalager?

(tvekan) När vi (kort paus) plockar fram kampanjer till exempel, då är det ju väldigt viktigt, dels matcha med adresser och leta upp rätt ställen och så vidare. För analys, (kort paus) det är väl dom stora delarna.

Vad är det för typ av analyser ni gör?

Vi jobbar med analyser ofta på mosaikkoder och så vidare i, egentligen lönsamhetsanalyser är det ju frågan om på kunder för att försöka rangordna dom och då tittar vi lite på var dom bor och vad man kan tro att dom är för typer av personer, om dom är välutbildade eller inte och sådant. Mycket kan man ju spåra beroende på vilka områden människor bor i, hur grannarna är så att säga. Så det är den typen av analys och uppföljning också för att se vem var det som köpte och vem köpte inte och försöka lista ut varför vissa gör det och andra inte, så det är då vi oftast kombinerar den externa datan.

Vad ser du att externdatan ger för möjligheter när det gäller dom här kampanjerna som du pratade om?

Det är framförallt med adresser. Det är dom här som jag nämnde, mosaiktyperna till Marknadsanalys som har gjort ett klassningssystem för ja, alla hushåll i Sverige egentligen. Så att det är dom bidragen vi ser där vi har nytta av då i alla dom här stegen, både för kampanj, analys och så vidare.

Finns det några relaterade problem med just någon av dom här tillämpningsområdena som du ser det?

Nej det enda är väl att vi är lite beroende av den externa leverantören. Men annars så, vad menar du med relaterat problem?

Ja det kanske är något problem som uppstår i och med att ni har, tar in extern data som ni kanske har svårt att tolka den eller att ni har svårt att använda den på det sätt som ni vill?

Ok, nej inte som vi gör just nu. I, ja det kanske jag också skulle sagt, men i datalagret så tolkas ju vissa fält i datat så att säga och kodas om så att vi kan använda det för analys och så vidare. Så att nej just nu är det inte så utan att det funkar bra, det är väl snarare så att det finns vissa saker som vi skulle vilja ha som vi inte får idag, för att personuppgiftslagen säger stopp och så vidare och (kort paus) där vi har gammal data. Men annars så ser jag nog inte det.

Vad finns det då för användningsområden som ni avstår ifrån som du pratade om?

Ja det är just att det finns ju väldigt mycket data som man kan få tag på men av lager så är vi ju hindrade från att använda allt för mycket information om folk, om deras inkomster och vilka, hur många barn dom har och sådana saker. Vi kan egentligen få reda på att folk bor på ett visst ställe, att dom möjligtvis har avlidit och så men självklart så skulle det ju vara mycket intressant om man kunde få reda på andra saker när man börjar titta på sin kundbas och vilka man tror skulle vara dom bästa kunder och så vidare. Men otillräckliga resurser, för stora problem har vi inte stött på hittills utan det är nog snarare så, ja otillräckliga resurser, vi har valt kanske att inte gå in i, att koppla upp oss direkt mot till exempel UC och sådana här upplysningsföretag men där köper vi istället tjänsten när vi behöver det då. Så att hittills har vi inte direkt, ja det är klart jag skulle vilja ha mycket och det är alltid pengar som sätter stopp. Problem tror jag inte att vi stött på några egentligen.

Men det är mest lagarna då som sätter gränserna?

Ja det kan man säga i vårt fall. All data finns nog att köpa men vi får inte använda den på det sätt vi vill.

Tror du att användandet av extern data kommer att öka i framtiden i just erat företag?

(lång paus) Ja kanske, det tror jag. För jag tror också att vi kommer bli bättre på, via DW Data Warehouset att egentligen att använda den information vi redan har i egentligen som vi sitter på via våra produktsystem men som idag är svår att komma åt. Men det som jag ser som en stor fördel med Data Warehouset är att man kan jobba med det på ett annat sätt, datat. Så att vi kommer säkert att använda mer extern data men även den interna datan i ännu större utsträckning.

Då har vi egentligen bara några avslutande frågor kvar. Vilka generella fördelar ser du med användning av extern data i datalagret?

Generellt sett så fyller dom på med kvalitativ data som vi inte kan komma åt själva på ett så bra sätt.

Generella nackdelar då?

Att vi är beroende av den leverantören. Funkar det inte en dag så sitter vi där liksom, om vi ska skicka ut ett brev och inte har färska adresser eller någonting.

Använder ni extern data till någonting annat? Du nämnde UC och så att ni använde det i era.

Ja det gör vi, banken gör det i kreditupplysningar och så vidare, i kreditbedömning, dom använder extern data och vi använder ju bilregistret (kort paus) bland annat och lite andra sådana här Pararisk, ett risksystem och så vidare men det är ingenting som går till DW:t som sagt utan det används direkt i produktsystemen. Men vi använder extern data ganska mycket.

Har du någon uppfattning om hur stor del av den totala mängden extern data som verkligen kommer in i datalagret?

Ja jag vet inte. Nej det var jättesvårt tycker jag att säga en mängd men, av extern data som används (kort paus) nej jag vågar inte säga någon andel där.

Något övrigt du har att tillägga?

Nej egentligen inte.

Respondent 6

Namn?

Det var NAMN

Din tjänstetitel?

Ja jag har väl ingen speciell men systemansvarig kan man väl kalla mig för.

Företagsnamnet?

FÖRETAGSNMAN

Typ av företag, bank där?

Ja det är väl bank och försäkringsbolag skulle man väl kunna säga. Därför att vi är distributör utav BANK1, försäkringsbolagets produkter också, förutom att vi är en bank.

Dina ansvarsområden då?

Ja det är Data Warehouse och kundsystem, marknadssystem kan man också säga.

Har du någon tidigare erfarenhet av datalager?

Ja, jag har jobbat med det här från och till de senaste 20 åren.

Ok, då har du varit med ett tag där.

Ja

Din definition av ett datalager i så fall?

Ja vi är inte så nog med det, jag satt och funderade på det och egentligen är ju den exakta definitionen att det som är vårt datalager, det är det som finns i den här servern som vi använder som datalager. Det är ju en knasig definition egentligen men det beror på att vi inte har (kort paus) funderat särskilt mycket på det där utan helt enkelt, det har varit viktigt för oss att separera produktionsdata från analysdata. Så att det som vi har åstadkommit är att vi skyddar våra produktionssystem från den här typen utav frågor som bara besvärar produktionssystemet egentligen så att i vårt datalager så har vi, kan vi ställa frågor som vi normalt inte kan ställa i produktionsmiljön beroende på dels att vi har samlat informationen på ett ställe och dels beroende på att vissa typer av frågor ställer man lämpligen inte i produktionsmiljö utan där är det mer att vi frågar efter en kund eller ett konto i taget eller någonting sådant. Det är väldigt väldefinierade frågor som vi har i produktionsmiljö som är, som ska vara snabba och så vidare.

Och din definition på extern data?

Ja den är lika trivial den egentligen. Det är data som vi skaffar in utifrån kan man väl säga. Som alltså inte produceras inom företaget.

Då kommer vi in på identifieringsfasen. Vilka källor inhämtar ni externdata ifrån?

Ja det är ett flertal olika källor kan man väl säga, det är våra olika produktionssystem, till exempel ett för bankreskontra, vi har ett för fondhantering, aktiesystem och sedan så eftersom vi också är distributörer utav liv och pensionsförsäkringar sådant där, och

linkförsäkringar så får vi då information från de systemen, men det kallar vi inte för externa data utan det betraktar vi mera som internt data. Så att det är från alla produktionssystem och lite, ja rätt mycket olika typer utav data som vi tar in.

Kan du ge exempel på olika typer utav data som ni tar in?

Jag brukar säga att vi har tre olika typer utav data, dels har vi en, det är kundinformation, kundadressinformation, det egentligen bara en tabell men den är väl rätt stor, där vi hela tiden underhåller och ser till så att vi har korrekt adress till våra kunder.

Ok, och var hämtar ni denna information ifrån?

Ja den har vi alltså i våran produktionsmiljö men används också i Data Warehousemiljö och den underhålls då utifrån, varje vecka från SPAR och svensk adressändring. Det är den första och den andra då, vårt engagemangsregister som vi sammanställer då i vårt Data Warehouse så där tar vi då från olika delar utav företaget som jag nämnde tidigare, bank, fond och aktier, liv och link och vad det kan vara för någonting. Och den sista biten är, skulle jag vilja kalla för händelseloggar där vi sparar information vad som har hänt egentligen i vår relation till kunderna, om vi har gjort någon kampanj, om vi har gjort utskick, om kunden själv har varit inne på vår Internetbank, om vi har ringt kunden eller allting sådant där. Med dom tre bitarna så har vi, tycker vi väldigt komplett Data Warehouse och kan hela tiden se vad som har hänt och hur vi ska bearbeta våra kunder, för att om vi går tillbaka till definition igen så såg jag den som ni hade där och den var ju enbart för decision support. Men vi har den även för praktiskt arbete, det vill säga vi kan göra, ta fram DR-utskick och sådant där via vårt Data Warehouse också. Så att det är både analys och produktion kan man väl säga. Ja och vi, varför vi hämtar just de här källorna, det har väl vuxit fram under årens lopp här, vi började med ganska liten skala och det byggs på hela tiden med olika typer av ny information som vi känner att vi behöver.

Och hur identifierade ni dessa källor?

Ja det blir efter hand som vi känner att det blir ett nytt informationsbehov som ska tillgodoses eller tillfredställas.

Känner du till några andra externa källor som kan ge data till er?

Ja om vi pratar extern information så är vi då, att vi tar in till exempelvis ett förmögenhetsregister på svenskar som har hög inkomst eller hög förmögenhet, vi har ett företagsregister till exempel som vi tar in en gång i månaden med alla svenska företag. Vi har tagit in lite grand sådana här statistiska uppgifter men det är ganska blygsamt det vill säga befolkningsstatistik och sådant där att jämföra då med våra uppgifter.

Varifrån tar ni in dessa, vilka företag är det som ger er dessa?

Ja det mesta hämtar vi in från SEMA faktiskt, eller SchlumbergerSema som det heter nu numera och vi köper ibland adresser utifrån men det är ganska blygsamt nu för tiden eftersom vi har så mycket egna adresser så vi behöver inte köpa så mycket nu för tiden, utan vi jobbar med mycket med korsförsäljning.

Ok, men tycker du att det är svårt att hitta leverantörer av externdata?

Nej, det tycker jag inte att det är. Det är snarare tvärtom.

Hur ser framtiden ut, kommer ni använda andra källor, tror du , ta in externdata från andra leverantörer?

(kort paus) ja, (kort paus) jag har inga konkreta planer på det, men det skulle nog inte förvåna mig.

Då kommer vi in på anskaffningsfasen här. Hur hämtas datan in från era externa källor? Är det alltså on-demand, alltså ni efterfrågar hos leverantörerna eller prenumererar ni på den?

Ja det är väl både och kan man säga. Det här företagsregistret, det får vi som jag sa till exempel en gång i månaden så det är ett abonnemang, medan det här företagsregistret köper vi från fall till fall, det är ganska dyrt så att, vi har. Planerna var att vi skulle köpa ett nytt varje år men nu hoppade vi över det här förra året för vi tyckte att vi klarade oss bra med det vi hade. Så att det är väl en kombination utav abonnemang och styckeköp.

Hur mycket av den datan ni hämtar är skräddarsydd efter just ert datalager?

Jag vi ser till att den blir skräddarsydd, (kort paus) Jag tar in filerna och stoppar in dom där dom, och ser till så att dom passar in bland den andra informationen så att man får ju aldrig skräddarsytt material utifrån, utan man får det man får så att säga så brukar jag se till att de blir användbart.

Ok, så att ni får det bara i en given form så får ni transformera och fixa i ordning den själva då?

Ja så har det vart hittills i alla fall. Ja det kan man väl säga, man får ibland säga om var man ska ha det och lite grand men oftast är det så att man får det man får så att säga.

Ok så dom, leverantörerna skräddarsyr liksom då egentligen inte efter ert behov utan dom har en standard egentligen som dom skickar ut?

Ja, mer eller mindre. Det kan väl vara så ibland att man resonerar om vad exakt man ska köpa men det är inget större problem för oss utan vi ser vad vi får, oftast får man med en filspecifikation och så tittar man på den och ser, ja men det här är precis vad vi behöver då. Man kanske får med lite data då som man inte behöver, då är det liksom bara att strunta i det.

Kan ni begära av era leverantörer att de iordningställer annan data efter era behov, om ni skulle behöva något annat?

Det tror jag säkert att vi skulle kunna göra. (kort paus) Inom vissa givna ramar, det är alltid känsligt det här med personnummer så det är sällan vi får med när vi köper utifrån, till exempel om vi köper personuppgifter får man inte lämna med det, i alla lägen.

Och vilka distributionsvägar använder ni för inhämtning av data? Får ni det alltså, ligger ni on-line mot leverantörerna eller skickar de den kanske via cd-rom eller?

Ja det är både och kan man säga. Företagsregistret får jag på cd-rom till exempel och (kort paus) om vi kallar det här som jag får från SPAR då varje vecka så, det vill säga adressändringar och sådant där så får vi det via filöverföring till oss. Exempelvis.

Ser du några nya sätt för inhämtning av data i framtiden?

Ja man kan väl tänka sig att man hämtar via Internet men (kort paus) ja dels så har vi ganska besvärligt med att hämta in data, eller besvärligt och besvärligt men det är rätt mycket säkerhetsproblem. Vi vill inte ha för mycket hål i våra brandväggar och så vidare så att. Det här med cd-rom fungerar faktiskt ganska utmärkt.

Ok, jättebra där. Integreringsfasen då. Vilka tillvägagångssätt använder ni er av vid integreringen av externa data i företagets datalager?

Ja det är hårt arbete höll jag på att säga men (kort paus) jag gör ju det där själv hela tiden och det gör jag med hjälp utav SQL-bearbetningar som är då när jag har fått i ordning på det så sparar jag det, alla de här, i form utav lagrade procedurer och är det så att det blir ett återkommande jobb så läggar jag det så att säga i vår jobbhanterare. Så att, jag börjar med att experimentera med handskriven SQL då, när jag ser att det funkar så blir det mer och mer rutin utav det.

Du gör alltså detta manuellt då?

Ja det kan man väl säga att man måste börja där någonstans så att, vi har inget färdigt system som tar hand om alla ny fall.

Detta är enda sättet ni använder er av eller?

Ja, jag kan inte se något annat att lösa det på egentligen. Man måste ju in och hacka i koden och titta exakt vilka uppgifter man får och hur dom ska transformeras och var man ska lägga dom någonstans och så vidare.

Känner du till några andra tillvägagångssätt än de ni använder er av?

Ja (suck) känner till och känner till, man kan väl tänka sig att man (kort paus) gör något program eller att leverantören möjligtvis tillhandahåller något program, det har jag väl hört talas om vid något fall i alla fall men då hade jag redan gjort det själv så att.

Ok, hur lagras den externa datan i datalagret?

Ja den lagras tillsammans med den interna datan så att, vet man inte vad som är vad så kan man inte se om det är externt eller internt.

Tror du att ni kommer öka integreringen av extern data, i den meningen att graden av integrering? Eller den är kanske total? Det kanske inte går att integrera den mer?

Nä det kan jag väl inte påstå att den är, men det sker ju förbättringar och finslipningar av det här hela tiden.

Ja, men det är något du tror kommer öka i alla fall?

Ja det gör det.

Ok, ser du några tydliga trender annars i tillvägagångssätt för integrering?

(lång paus) Ja jag har ju sett sådana här systemlösningar där man med mera drag-and-drop teknik kan integrera tabeller men jag är lite skeptisk mot sådant så att.

Vad är detta, mer drag-and-drop, kan du utveckla det någonting?

Ja det finns alltså färdiga system där man plockar ihop olika (kort paus) ja, tabeller eller filer eller någonting sådant där för att lägga in i sitt Data Warehouse. Men nä jag ser inga problem med det sätt vi gör idag, då har man tycker jag lite bättre kontroll över det också.

Ok, då kommer vi in på användningsfasen då. Inom vilka tillämpningsområden använder ni extern data kombinerat med erat datalager?

Ja, det finns lite olika kan man säga, dels finns det olika rapporter som vi lämnar ut till olika myndigheter och företag med återkommande som, (kort paus) tas ut från vårt Data Warehouse, det är väl en bit och sedan har vi det som jag nämnde att vi gör olika marknadsbearbetningar, PR och telemarketing kanske och så vidare. Och sedan är det då analys och statistik och sådant här, det är väl dom tre sakerna, möjligtvis att vi tar fram lite sådana här, Data Warehouset är lite grand också ett komplement till vår produktionsmiljö, det vill säga att vi snabbt kan ta fram rapporter som man på sätt och vis lika väl skulle kunna göra i produktionen men det tar mycket längre tid att ta fram dom där och så vidare. Så att man kanske har det som en provisorium och ibland så blir provisoriumet permanent att man tar fram de här listorna, kan vara olika underlag som personal har inom banken för olika aktiviteter.

Vad ser du att externdata ger för möjligheter för dom här olika användningsområdena?

Ja det är väl mest för analys och kundbearbetningar som vi har nytta utav dom sakerna.

Ok, kan du utveckla det lite, vad bidraget är och så?

Ja exempelvis så kan vi använda det här förmögenhetsregistret till att selektera ut dom kunderna vi vill bearbeta, både egna kunder och presumtiva kunder, och även företagsregistret använder vi till sånt.

Finns det några problem med detta, med externdatan i detta?

Ja det finns ju ett problem med det här förmögenhetsregistret det är att vi inte har personnummer utsatt, jag får alltså matcha, men vi har faktiskt fått födelsedatum, men det är inte hundraprocentigt i alla fall så att det är väl ett problem som vi har, med den biten att det kan bli fel där ibland att vi tappar en del som inte får match på för att jag får match både på namn och adress och namn och sånt där, och det kan vara lite knepigt ibland.

Finns det andra användningsområden som ni avstår ifrån eftersom ni har otillräckliga resurser eller för stora problem eller något? För just externdata då?

Nej, inte externdata men däremot så har vi väl (kort paus) försökt i något sammanhang att ta in transaktioner och bankreskontra, det har liksom, det orkar vi inte med, rättare sagt, maskinerna orkar inte med. Men det är inte externdata utan det är interndata, det kan vara externdata som ligger hos en underleverantör hos oss eller är outsourcat men det, det är väl någonting som jag kom att tänka på som vi har liksom, något jag velat göra men inte kunnat.

Tror du att användandet av externdata kommer att öka i framtiden?

Ja, det tror jag nog att det kommer att göra, men det är inte något dramatiskt.

I vilka användningsområden tror du att det kommer att öka mest?

Jag tror det kommer öka mest när det gäller analys och jämförelser bland vårt eget kundregister då och demografiska data då som man kan köpa utifrån.

Ok, då har vi väl egentligen bara några avslutande frågor kvar. Vilka generella fördelar ser du med användningen av externdata?

Ja det är väl helt enkelt så att det är information som vi inte har själva som vi behöver.

Vilka generella nackdelar ser du?

Ja, möjligtvis kan nackdelarna vara att man inte känner riktigt hur datat uppkommer och så vidare, det kan väl kanske vara problem ibland, att tolka datat, eftersom det inte är sitt eget data och, men det får man väl undersöka framför allt, i fall då sådana här saker, det är väl någonting som har att göra med, jag menar när man tar in nya externa data att man får undersöka och titta på vad det verkligen är för någonting. Det kan vara lite kryptiskt ibland.

Använder ni även externdata till någonting annat som inte kommer in i erat datalager?

(tvekan) ja, vi överför viss del utav extraherad, bearbetad information till vår produktionssystem faktiskt. Så att det är inte externdata direkt in utan det kan vara så att vi har matchningskört för att kolla till exempel vissa uppgifter och så, dom finns också i vårat produktionssystem.

Hur stor del av den totala mängden externdatan kommer verkligen in i erat datalager, har du någon uppfattning om det?

Ja det är väldigt svårt att säga för att vi har väldigt mycket information och väldigt mycket intern information, skulle tippa att det rör sig om runt fem procent någonting sådant där, inte mer i alla fall om man räknar antalet poster eller någonting sådant där.

Jag har bara en liten kompletterande fråga beträffande integrering som jag satt och tänkte lite på här. Har ni några problem när det gäller integreringen egentligen som du har tänkt på, reflekterat över, alltså har ni några generella problem?

Ja vad ska man säga om det, (kort paus) idag har jag haft problem till exempel, beroende på att mycket av dom här körningarna som görs då det är stora bearbetningar som ibland kan ta flera timmar, som då är inlagda som schedulerade jobb och ibland så kan det hända att man får något fel då, och ligger felet i början av kedjan så har man kommit en bit på väg och sedan så spricker allting och det är väl mycket sådant som kan lägga sig in under rubriken problem med det här, och det är väldigt svårt också att förbestämma alla situationer som kan uppkomma, jag menar man kan få någon, nästan fysiska fel ibland och, det här jag nämnde med extern data kan också göra att man inte har koll riktigt på det, hela processen och då kan få in oväntade typer av data som man inte har haft tidigare som ställer till det för en och sådana där saker. Så det är väl ett litet problem som jag har brottats med hela förmiddagen idag till exempel. Det händer då och då.

Har du något ytterligare att tillägga?

Ja, jag brukar ge lite goda råd ibland och det är många som misslyckas med sina datalager, det är ett allmänt känt faktum. Och som vi har gjort då, börjat i liten skala och växer därifrån, är väl ett gott råd och att hela tiden förstå vad man håller på med är också ett annat råd, det låter väl kanske lite konstigt men det är inte alltid man gör det alla gånger. Man har en övertro på att system ger en, ja löser problem och ger svar på frågor som man inte kan lösa på något annat sätt, men så är det väl nästan aldrig. Vi har väl satsat på det, eller jag, kan man väl säga, för det är jag som har byggt upp alltihopa i princip, med (kort paus) på flexibilitet men vilket har gjort att vi har liksom inte köpt något färdigt system men vi har ett rapportverktyg som heter business objects som är ett inköpt verktyg för att ta ut rapporter som användarna själva kan göra, men annars så bygger det väldigt mycket på den funktionalitet som finns i databasen, vi kör på Microsoft SQL server och där finns det ju att man kan bygga dom här, lagrade procedurerna och man kan schedulera jobb, det räcker långt för att skapa ett sådant här Data Warehouse. Och det gör att man, vi kan göra det väldigt flexibelt också, så att vi har liksom inga större begränsningar utan vi skräddarsyr varenda liten grej egentligen kan man väl säga. Köper man ett stort system, då får man för det första ta ett stort steg med en gång och för det andra så gör det det det ska. Ibland kan det vara en fördel, ibland en nackdel.

RESPONDENT 7

Namn?

NAMN

Tjänstetitel?

Projektledare och förvaltningsansvarig.

Företagsnamnet?

FÖRETAGSNAMN

Ja, vad är det för typ av företag?

Det är ett nystartat företag som är ungefär ett år gammalt, man kan säga att vi är nischat då alltså, ett backoffice företag för banker och finansiella institut, inkluderar även IT då, fast vi har jobbat med sådana här frågor för de här två bankerna som vi har nu, BANK 1 och BANK 2, och så har det blivit ett eget bolag egentligen för den avdelningen.

Ok, och dina ansvarsområden är?

Ja som det är nu då så har jag bland annat då (kort paus) jobbat som projektledare generellt då i ett IT projekt men också då speciellt ansvar för utvecklingen och förvaltningen av bland annat då datalagret, och jag har även då några rapporteringssystem och finansiella system som jag också är typ då ansvarig för så att säga, utveckling och rättningar och att det drivs framåt.

Har du några tidigare erfarenheter av datalager, har du deltagit i några tidigare projekt eller något sådant?

Nej, det (kort paus) för mig var det väl (kort paus) jag har jobbat med det i två år nu ungefär och innan dess hade jag inte jobbat med datalager mer än, en del av datalagret är ju också själva databaserna och jag har jobbat rätt mycket med databaser innan. Just med kuber och så där.

Ok, hur definierar ni datalager? Vad är din definition av datalager?

(Kort paus) Som jag ser det så är det ju egentligen dom här, att man kan vrida och vända på informationen från olika håll på olika djup. Att gå ner på en kund och titta på. (Kort paus) eller att uppifrån ett kontor och gå ner på en handläggare ner till en kund om man vill det eller så titta på en produkt, man säljer en viss produkt och går ner och tittar på den på olika, från olika håll va. Så det är just att kunna vrida och vända på vad man vill egentligen. För att följa upp något som har skett egentligen då och betalningar, produkter som förändras, kunder som flyttar och beteenden så att säga. Det är mycket så att det är controllers och ledningen inom företagen här, bankerna då som, ser på trenderna och kan se då vad, vad det går emot och man också se då genom att uppdraget vad som är svagheten, vad man säljer lite av och kan satsa mer på det området och rikta sig kanske då mot vissa områden som man ser då, här säljs det dåligt på dom här kontoren då eller vissa handläggare som inte sköter sig.

Och din definition på extern data?

Ja (kort paus) där ser väl vi som vi jobbar då så har vi extern data från externa enheter som till exempel BANK 3, det kan vara SBAB, det kan vara Stockholmsbörsen, där vi får in egentligen filer som dom skickar till oss, datafiler som vi läser in då som hamnar i tabeller i datalagret och sen i kuberna också då, det som vi tar in där. Alltså externa företag egentligen.

Då går vi in där på identifieringsfasen och så tar Markus över lite här.

Ni nämnde några källor som ni tog in extern data ifrån, finns det några mer än dom tre som du nämnde?

Ja, det finns det ju. Kasella har vi ett företag som har fonder, bankerna jobbar mycket med Kasella, och sen har vi också mycket fondsparande som kunderna har som vi ju då får från dom här företagen eftersom dom håller i datat egentligen då, vi bara begär att få det och dom skickar över till oss vad vi vill ha efter att vi har kommit överens med dom då (kort paus) Stockholmsbörsen var jag inne på, det finns också något som heter Finess, ett system som är kopplat till Stockholmsbörsen där man tankar ner mycket avslut då från börsen. Alltså Backofficedata för aktieaffärer och fondaffärer hamnar i det här systemet och då får vi ut information från det här Backofficesystemet till oss så vi kan följa upp också i våra affärer. Vad kunder har gjort då. Man kan se på kundförsäljningen eller på fonder och aktier.

Finns det någon viss anledning varför ni hämtar från just dessa källor. Finns det några andra källor som skulle kunna spela samma roll?

(Lång paus) ja det är i så fall direkt från Stockholmsbörsen eller där man handlar så att säga. Men det är lättare då att handla det mesta då i ett slutsystem här eller ett fåtal system. Det är lättare att ta ett gränssnitt från det till oss.

Hur identifierade ni dessa källor? Hur hittade ni att ni ska hämta data från dom?

Ja det är väl ganska (kort paus). Affärsmässigt har ju bankerna då kontakter via avtal med olika företag då har ju dom sagt att det kan ni bäst få från oss via det här systemet för att oftast så har ju dom här stora aktörerna SBAB, BANK 3 med flera andra kunder också ju med liknande gränssnitt till andra kunder också.

Finns det några andra källor som ni känner till, som ni inte använder?

Ja det finns en massa som man säger framåt sett vill ha in då va. Vi får ju nya kanske (kort paus) till exempel nu så ska man börja med något som heter Nordnet också, där man handlar aktier då och så kommer dom att starta upp det här i maj och då kommer vi vilja ha information från det här systemet till vårt datalager. För det är ju någonting som hela tiden pågår då, det kommer nya källor hela tiden beroende på vilka avtal man sluter med olika leverantörer.

Skulle du vilja säga att det är svårt att hitta leverantörer av extern data?

Nej inte om man pratar om just de här som vi sluter avtal med, eller bankerna som vi pratar med så blir det naturligt. Det blir ju dom som, det brukar vara svårt att, det är mer att det kostar pengar och tar tid. Det är det som är problemet. Att man får prioritera då. Men det är så att den här informationen får prioritera helt enkelt då och det viktigaste går alltid först då och läggs mest pengar på, är det någon information som inte är så där jätteviktig, som kanske bara ett fåtal kunder använder och det kostar för mycket att få in

informationen så tar man ju inte med det. Vi har ju sådana källor som vi vet också att här har dom data men vi struntar i det för det är så liten mängd och det kostar för mycket helt enkelt. Så det är alltid en prioritering hela tiden. Så det finns massa som vi inte använder som vi egentligen kanske ville ha.

Framtiden har du nästan svarat på lite men finns det några mer källor som du tror kommer att användas i framtiden än dom du har nämnt?

(kort paus) Ja, menar du, det kommer ju alltid att komma till nya källor det gör det ju, det utvecklas ständigt nya kontakter med leverantörer så det blir det hela tiden.

Ok, kan du ge något exempel som du tror kommer att ske snart?

(Lång paus) Ja, säkert med EU här kommer det säkert att vara information som finns någonstans, kanske i Bryssel eller någonting, det kommer säkert bli gränssnitt därifrån och nya gränssnitt från, ja va det nu kan vara. (kort paus) ja.

Ok, då kommer jag in här på anskaffningen igen. Du nämnde förut att BANK 3 och dom här hade någon fil som ni laddade ner. Är det enda sättet ni hämtar in data på eller finns det andra tillvägagångssätt att få in datan?

(kort paus) man kan säga att vi har då en huvudleverantör av datatjänster här som driftar våra system och datalager bland annat då, som heter TJÄNSTE-LEVERANTÖR 1 och den leverantören då, egentligen så bryr inte vi oss så mycket om hur dom gör det här utan vi köper ju en tjänst då från det här IT-företaget TJÄNSTELEVERANTÖR 1 som då gör det här sedan. Så att vi vill bara ha en funktion, vi vill bara ha den här informationen in i datalagret vid den här tiden till en viss kostnad då. Då får dom lösa det egentligen bäst dom vill. Men det blir ju så att dom har en standardlösning som vi känner till då. Dom vill ha ett visst sätt via filöverföring, kanske då krypterade filer till deras driftcentral en viss tid på dygnet och det ska komma då i ett visst format den här filen och så vidare. Det finns en standard för att dom ska kunna använda sina mottagningsprogram som de har redan då. Det blir också billigare för oss om vi använder oss så mycket som möjligt av det som finns.

Vet du hur mycket av den data ni inhämtar som är skräddarsydd eller det låter som ni menar att det mesta är skräddarsytt efter era behov, är den det?

Nja, det kan vara antingen eller, i många fall är det så va, om det är så att till exempel om vi tar BANK 3 och vi får information från dom så kanske dom har ett gränssnitt mot en annan bank och då blir det oftast billigare för oss om vi tar samma gränssnitt, samma innehåll i filerna som andra får då finns det redan färdigt hos BANK 3 så att säga och då kan vi få det till en billigare kostnad alternativt då är ju att man kanske inte har någon annan kund, vi är första kunden, då kan ju vi mer själva säga till att det här vill vi ha, i det här formatet och med den här informationen, så får de bygga den från början. Då skräddarsys det ju.

Men dom har möjlighet att göra det då alltså?

Ja det finns ju alltid. Det är bara pengar som styr egentligen. Vi kan ju skräddarsy vad vi vill men oftast är det ju en högre kostnad så det är alltid en avvägning.

Kan ni begära från era leverantörer att dom iordningställer annan extern data efter ert behov egentligen om ni skulle behöva?

Ja det är pengar som styr oftast, bara de har möjlighet att lösa det. Det kanske tar längre tid då om dom ska göra det då så det är beroende på projektet, om det är tidskritiskt projekt eller inte och kostnadskritiskt så att det mesta går ju att lösa då.

Ser ni några nya sätt för inhämtning av data i framtiden? Eller det funkar bra så med on-line nedladdning då?

Jag tror att det finns en del, det finns säkert en del förbättringar här man kan göra för filöverföring. Det är egentligen lite föråldrat och det kommer säkert nya tekniker här nu.

Du vet inget exempel på det eller?

(tvekan) Det kommer säkert här något som har med (kort paus) sådant här direkt, alltså att man är direktuppkopplad då kanske via Internet eller någon säker Internet-koppling. Att man då bara direkt så att säga, att vi läser egentligen direkt i någon databas hos någon extern. Det finns sådana här förslag som vi har fått från leverantörer att vi ska uppgradera och ha sådan lösning istället. Det finns ny teknik, säker teknik som gör att man via Internet kan ha säker kommunikation och med nycklar och så här, och privata nycklar kan man få informationen snabbare och säkrare egentligen än via filer, så att det kommer. Vi har diskussioner om det.

Ska vi se om du kan svara på detta också, det blir lite mer tekniskt, det är vilka tillvägagångssätt använder ni er av vid integrering av den externa datan till företagets datalager?

Ja vad är du ute efter då?

Hur ni integrerar den egentligen alltså, hur ni kombinerar den med er andra data.

Ok.

Lägger ni den i en egen dimension eller i ett eget data mart?

Vi har egentligen, i datalagret har flera kuber, men just i själva datalagret (kort paus) vårt datalager är egentligen väldigt uppbyggt efter våra controllers egentligen och säljare på bankerna, hur dom vill ha det. Man tittar på dels på produkter, alltså vad man säljer då, det kan vara (kort paus) Bank-kort, det kan vara VISA-tjänster och det kan vara olika produkter som kunder har va. Tittar man på dom. Eller också tittar man på konton kallas det då, kontonivå på kunder som också går upp till kontor och så vidare. Eller är också är det på, har vi också en basdimension eller baskub, där man har mer då totala försäljningsnivåer, total utlåning, total inlåning, volymer, total försäljningsvolym för banken, kontoret eller för en handläggare. Så att det integreras på så sätt i datalagret.

Ok. Känner du till andra tillvägagångssätt än det här ni använder er av?

(Lång paus) Jag kan ju tänka mig att (kort paus) Vi har ju på gång också att egentligen ha andra datalager nu än (kort paus) alltså mer att titta på statistik som egentligen inte just nu finns i datalagret, att titta på kundstatistik till exempel i datorbanken då som vi har, en sådan där Internetbank. (kort paus) Då bygger man också kuber då på den informationen som man då får ut ifrån det här systemet då, det är ett annat sätt.

Hur lagras den externa datan i datalagret?

(tvekan) Ja egentligen på samma sätt som interna datan om man säger så. Det lagras i databas, själva datalagret är en databas, SQL-serverdatabas där det ligger tabeller, uppdelat då på, ja en kundtabell, en kundfondtabell, en kundvolymstabell och så vidare då. Sedan har vi byggt då ett stjärnschema som det kallas för, eller stjärntabeller som egentligen är ett förberedande steg för de här kuberna och sedan när man laddar kuberna så, så laddar man då från det här stjärnschemat, för att få snabbare laddning av kuberna för att det är stor volym här på datan.

Ok, tror du att ni kommer öka integreringen av extern data?

(lång paus) Menar du att vi kommer ta in mer extern data in i datalagret eller?

Ja eller om ni kommer integrera den mer kanske man skulle kunna säga?

Jag kan nog tänka mig att vi kommer integrera kanske datalagret med andra system kan jag tänka mig. Det har vi inte gjort idag. Så att det finns ju, man märker mer och mer att det som finns i datalagret är det fler och fler personer som är intresserade av. Då finns det ju sätt att till exempel sprida ut det via webbgränssnitt kan man ha mot datalagret eller så att säga man kan distribuera ut informationen från en kub ut (kort paus) så att man kan läsa det, ja till en portal egentligen då och så kan man då gå in i den här portalen och se då en vy utav datalagret ju utav kuberna. Det finns ju sådana tekniska lösningar som man kan bygga upp då, att rikta information då mot olika typer av användare.

Då har jag en avslutande fråga på denna delen, om du ser några tydliga trender i tillvägagångssätt vid integrering?

(lång paus) Det är väl, just det här med (kort paus) intranät och portaler att man (tvekan) som webbgränssnitt är ju väldigt användbart och lätt att använda, alla kan använda det, det kostar inget, har man en webbläsare så ser jag att man kan sprida det mer nu från datalagret ut till ett webbgränssnitt för att kunna titta på det där. För det är väldigt höga licenser också för användare utav de här kuberna, att titta på dom i dom verktygen som finns, dom licenserna är väldigt kostsamma. Men att titta på det via ett webbgränssnitt är mycket, det kostar ju mycket mindre.

Ja då kommer vi in i fasen användning. Inom vilka tillämpningsområden använder ni extern data kombinerat med erat datalager?

(lång paus) Ja det är ju för att följa upp försäljning, att (kort paus) få en vägledning in i framtiden, vilka trenderna är för kunderna, hur de beter sig, används också lite för att (kort paus) ja egentligen underlag för årsbokslut kan man säga lite grand också, man kan ju (kort paus) behöver ju inte bara titta i kuberna utan man kan ju i datalagret via egentligen (kort paus) dom program vi använder oss av, som heter Impromto, kan man söka ut, kombinera själv ihop vilken data man vill titta på och få fram det i datalagret där finns ju det mesta samlat. Så där använder man en del underlag för och (kort paus) som underlag för årsbokslut och annat. Men den största delen är ju för att se trender och att planera framåt.

Det här när ni ska se trender, vad ger extern data för möjlighet och vad är bidraget för just detta?

(kort Paus) Ja, då ser man ju för de externa (kort paus) leverans (kort paus) kunderna så att säga leveransföretagen, BANK 3 och SBAB och allt vad det kan vara, hur kunderna beter sig med dom här, man kanske ökar sitt sparande i externa enheter så att säga, eller minskar och man kan se att vissa kontor inom banken använder (kort paus) vilka källor ska man sälja in, Bank 3:s försäkringar, vissa lyckas mindre, då kan man så att säga rikta kraften mot att öka försäljningen vid det ena eller andra kontoret och se till att dom säljer mer. Man kan se att det går upp och ner. Och man agerar utifrån det då. Det är mycket företagsledningen som tittar på det här ju.

Finns det några problem inom just detta med att se trender och så med extern data? Som du ser det?

(Lång paus) Nä, det är väl mer att (kort paus) man är ju beroende av en extern leverantör där och det händer ju att dom har stopp i sina körningar och får problem i sina miljöer och fel, fel data in till oss, och det händer ju titt som tätt, att dom har brister i sina system, och då blir man ju drabbad, man kan inte själv påverka det så mycket ju. Det gäller ju att ha bra avtal med dem och bra kontakter så att man kan få det ordnat så fort som möjligt, och sen kan, dom ändrar ju sina system och vi kan vara ovetande här om att nästa månad kanske de gör en stor förändring som påverkar oss, som vi missar då eller dom missar och vi får en extra kostnad kanske för att förändra hos oss för att dom ändrar hos sig. Man har ju ett beroende till dom, och det är ju en risk då.

Finns det andra användningsområden som ni avstår ifrån eftersom att det är för stora problem eller att ni har otillräckliga resurser eller något sådant?

Ja det finns det ju hela tiden, det är som jag sa en prioritering hela tiden. Det finns ju många som vi egentligen (kort paus) mycket information vi vill ha in från externa som vi inte har prioriterat då.

Tror du att användandet av externdata kommer att öka i framtiden? Att externdatans roll blir större så att säga?

(tvekan) Ja, hos oss tror jag att det är så, att det kommer att öka

OK, då har vi bara några avslutande frågor kvar då. Vilka generella fördelar ser ni med användningen av extern data? (kort paus) för att sammanfatta det hela.

(tvekan) Ja det är ju för att få en helhetsbild av en kund så att säga, hur dom agerar, man har ju inte bara så att säga produkter internt på banken man har ju också kanske fonder, aktiehandel som ligger då externt då får man in en total bild av varje kund och det måste man ha för att kunna få en total uppföljning på kundens agerande. Alla affärer sker ju så att säga inte bara på banken utan det är ju (kort paus) på Stockholmsbörsen och andra som vi har så att säga, dom externa hos oss är ju partner hos oss egentligen eller partners som vi jobbar med, då ser man ju också vad kunden gör hos dom.

Vad ser ni för generella nackdelar med extern data?

Det är väl det här beroendet man får till dom, till dom här externa att dom sköter sig och att deras systemförändringar påverkar oss oftast då negativt i kostnader och att det blir kanske barnsjukdomar, om dom kör igång ett nytt system så får vi problem hos oss och det tar alltid tid att få det rätt igen. Det är väl det som man märker.

Använder ni även externdata till något annat utanför datalagret så att säga?

Ja det händer att man lyckas (kort paus) få någon ytterligare användning där som är bra för andra också att ha va. Till exempel i den här Internetbanken som vi har så har vi samma information ibland i datalagret som i vår Internetbank till exempel då information beträffande BANK 3, det visar ju för kunderna i deras datorbank och där finns det mycket fördelar.

Hur stor del av den totala mängden externdata som ni har i erat företag används och integreras i datalagret? (kort paus) på ett ungefär.

Ja, det är svårt att säga men det är kanske hälften som vi har prioriterat och fått in men resten finns ju kvar så att säga då som inte, som vi inte använder.

Någonting ni har att tillägga för övrigt?

(kort paus) nej, kan väl säga att det vi har gjort tekniskt här (kort paus) det är så här att vårt system byggdes 1999 kan man säga då, datalagret också då och även banksystemen med och det var så att dom prioriterade då att få igång själva banken här så att det var en konvertering från en banksammanslutning till fristående bank som dom har gjort här då. Och då fick vi prioritera själva banksystemet att få igång det till ett visst datum och datalagret kom så att säga i andra hand då och man fick väl, då fick vi problem med att dom byggde det alldeles för hastigt och dom som byggde det här hade ett väldigt pressat tidsschema och då blev det så att säga (kort paus) systemet blev inte bra från början så vi lider fortfarande av det nu tre år efter, fyra år efter att vi håller på att bygga om systemet eftersom det blev fel från början så att vi håller på nu att bygga om det successivt och det som jag var inne på med stjärnschema och det här har vi inte haft tidigare men det är en ny teknik, det är något som vi gör just nu då. Det är hela tiden det här att det är väldigt pressat i tid och pengar då som styr, man får prioritera så att det är (kort paus) Man får lida för det väldigt om man inte kommer rätt från början. Det kostar ju väldigt mycket då att bygga om, det är frustrerande för användarna som inte har det från början.

RESPONDENT 8

Namn?

NAMN

Och din tjänstetitel?

Chef för IT-systemutveckling.

Företagsnamnet?

FÖRETAGSNAMN

Typ av företag?

Försäkringsbolag

Dina ansvarsområden?

Jag ansvarar för en liten grupp systemutvecklare som arbetar mest med analyser och beslutsstöd av olika slag inom företaget, IT-relaterat.

Har du några tidigare erfarenheter av datalager?

Nej det fick jag när jag kom hit.

Hur länge har du varit där?

Sedan 98.

Har du jobbat med datalagret hela tiden?

Ja

Ok. Vad är din definition på ett datalager?

En mängd vad ska jag säga, (tvekan) definition på datalager det är en mängd tabeller, en stor mängd information som man kan plocka ifrån och plocka delar ifrån i olika riktningar, beroende på vilket svar man behöver så kan man alltså ta från olika hyllor kan man väl kalla det för. Precis som ett lager och då är det inte statiskt på det sättet utan att det är lätt att ändra, ställa en annan fråga, hämta ett annat data.

Ok. Din definition på externdata?

(Lång paus) Extern data, då tänker jag på (tvekan) nu ska vi se, fast data, alltså data som är från staten eller från inköpt data i ett visst format. Tycker jag då.

Då kommer Markus med några frågor om identifieringsfasen här.

Hej. Vilka källor inhämtar ni externdata ifrån?

Ja det var ju frågan om definitionen på extern data då. Vi hämtar data ifrån ett servicebolag först då där källsystemen finns, på ett sätt är det ju extern data men jag tänker mig extern data som att man hämtar från en fixed mängd data som är enligt vissa regler, rutiner då från typ SPAR och så vidare. Så vi får ju, vi får ju data från personregister och bilregistret och kan även då köpa in viss mängd data för kampanjer,

men det kommer alltså via vårt servicebolag då. Det kommer liksom inte direkt till mig om du förstår.

Varför hämtar ni just från de här källorna? Finns det några andra källor som skulle kunna ersätta dessa?

När man jobbar på försäkringsbolag så måste vi följa lagar och avtal va. Så då måste vi ta in visst data som är säkerställd så det måste vi då hämta från dom, befolkningsregister och bilregister och så vidare enligt dom lagarna som gäller då.

Hur identifierade ni källorna? Hur hittade ni att eller förstod att dom hade den info ni ville ha?

Ja dels då att vi måste följa lagar och förordningar så vi måste hålla oss ajour med lagarna och se vad, vilket register vi kan få ta ifrån så att säga, som vi måste ta ifrån. Så är det. Och i annat fall, ja det kan ju vara en sak till och det är att man hämtar en viss mängd, delmängd data i kampanjer och så och då, då går man till, men det är också sådana där statliga register som fastighetsägarregistret till exempel och det, att man vet om dom, det är då att man måste hålla sig ajour med, med dom statliga registren kan man säga, som finns.

Hämtar ni även från SPAR eller?

Ja det är det vi gör, så jag gör ju inte det direkt utan det hämtas från servicebolaget, servicebolaget hämtar det och så får jag min mängd som tillhör det här lokala bolaget i STADENS NAMN län då.

Finns det några andra källor som ni känner till som ni inte tar in i era system?

(Lång paus) Nej, nej, vi måste ju ta in dom som finns då för att vi ska hålla oss inom lagar och avtal.

Tycker du att det är svårt att hitta leverantörer av extern data? Allmänt så

Allmänt. (Lång paus) Nej alltså vi kan ju inte sväva ut då så mycket eftersom vi då är ett försäkringsbolag. Så det blir lite sådär stelt då. Vi kan inte göra så mycket annat än hålla oss efter dom, det som riksdagen har sagt så att säga. Så att, men det är klart att man, vet inte extern data men det är väl mer då att någon skulle kunna ställa samman data på ett visst sätt som jag vill ha till mig kanske, såna det är ju lite svårt att, det finns en uppsjö av företag som kan sälja det där men dom, man behöver känna att dom verkligen vet hur man jobbar med datalager och det kan vara lite svårare, om man går upp på den nivån då.

Hur ser framtiden ut då tror du att andra källor kommer att användas också? Att ni kommer ta in andra källor?

Hur menar du andra källor?

Att ni tar in, nu tar ni ju in från de här olika registren och SPAR och så vidare, att ni även tar in från andra?

Ja alltså man skulle kunna ta in delmängder om man ser att man (tvekan) hur ska jag säga det då (kort paus), om det öppnas upp mer till exempel att man får kolla upp olika kundsegment och så, så då kan vi hämta in sådan information, det blir väl mer och mer vanligt och i framtiden också att alla företag försöker se vilka kundgrupper man vill rikta

in sig på och man ser att här finns det människor som har två bilar och en motorcykel och då har dom ju säkert flera barn eller, då kanske man kan sälja det och det och då kan det inte bara vara försäkringar utan det kanske någon annan som säljer leksaker eller vad som helst. Att företagen börjar inrikta sig mer och mer i segment så man ser en kundgrupp som man säljer mer på. Det tror jag blir mer och mer vanligt. Men sedan är det då när man är på försäkringsbolag då så får inte vi, vi får inte hämta in vad som helst. Vi måste följa varenda förordning som finns.

Då kommer Calle här igen.

Ja hej igen, då ska vi se, då tar vi, kommer vi in på anskaffningsfasen här. Hur hämtas datan in från era externa datakällor? Är det alltså, har ni någon prenumerationstjänst, är det on-demand att ni känner att ni behöver mer data och då säger till eller har ni något avtal med dom?

Om vi pratar om dom här, typ SPAR och SBR, statens bilregister, det är ett avtal som vårt servicebolag har ja.

Och då prenumerationstjänst då? Hur ofta får ni in det då?

Ja det är ju varje vecka och sedan så, dom stora datamängderna och sedan kommer sådana där tillägg varannan dag tror jag att det är.

Ni har inget sådant att ni hör av er och säger att sådan här data skulle vi vilja ha och så köper ni in det?

Jo man kan köpa tillfälligt, alltså vissa delar kan man göra. Men det är ganska dyrt, för att då behöver vi ha en större mängd av oss lokala bolag. Vi är 24 stycken då som i så fall ber servicebolaget köpa in det.

Ok, hur mycket av den datan ni inhämtar är skräddarsytt efter ert datalager?

(kort paus) Ja alltså, skräddarsytt.

Förstår du frågan?

Nej inte riktigt.

Behöver ni göra mycket transformeringar och sådant på den, alltså får ni den i just det formatet ni vill ha eller får ni bara en stor kanske klumpfil som ni fixa till själva?

Ja det är något mitt emellan. Ja vi har ju då bett från, ja som jag sa då så får ju inte jag direkt från SPAR till mig då va eller något. Utan det kommer först till servicebolaget och in i dom försäkringssystemen. Sedan har jag bett dom om min information till det här FÖRETAGET då, STADESNSNAMN information så att säga, och då har jag bett att få det utdraget då och dom fält som jag tycker att vi ska ha här och sedan så ligger det då i batch så det går ju dagligen sedan. Så att på det sättet har jag modifierat det därifrån genom att be dom ta ut det som jag behöver men sedan gör inte dom något mer. Sedan måste jag, när det kommer hit då, göra ett antal programmeringssteg för att lotsa in det i datalagret. Jag måste tvätta och fixa.

Det är det du får göra. Är det mycket sådant arbete med transformering och sådant?

Ja det är det. Därför att det är så många olika system, det är massor av olika försäkringssystem och våra försäkringssystem består av en mängd olika tabeller. Det är ett jättejobb, men det är det som är vitsen med datalager att man ska kunna ändra källsystemen egentligen hur som helst utan att det ska bli så himla beroende i datalagret.

Det kommer vi nog in på lite senare just den biten, men i vilken utsträckning kan era leverantörer skräddarsy datan efter era behov, vet du det?

(kort paus) Alltså dom har ju begränsade resurser, ju man ber dom om, ifrån dom, desto mer kostar det

Så allting styrs så egentligen av pengar?

Ja, men dom skulle alltså kunna göra det mesta men å andra sidan då handlar det så långt ner på prioriteringsskalan därför att dom måste syssla med dom stora försäkringssystemen. Så då blir det liksom att ett lokalt datalager som jag har blir lite längre ner på listan då i en sådan här stor koncern som det här är.

Tror du att du skulle kunna begära från din leverantör att dom iordningställer annan data om ni skulle få nya behov?

Ja, jag kan begära det. Sen beror det väl på när dom har tid att göra det. Jag kan säga vad som helst, jag kan be att få det där och där och där och där. Det är inga problem men, det kommer dom att göra någon gång kanske, år 2350. That's real life liksom. (skratt)

Vilken distributions väg använder ni er av vid inhämtning av datan?

(kort paus) ja menar du, vi har ju då file transfer protocoll.

Jaha, ni kör alltså online distribuering då eller man kan ju få in det på CD-skivor och sånt också.

Nej, nej det är alltså så där med protokoll, via protokoll från stordatorerna och via, i ett fall, och i ett fall, största mängden data tar vi in via en, vi har gjort en ambrovinkel, det är så att vi gör om datat till en virtuell lista och sen har vi ett program som kan exekvera listor och så lurar vi den att tro att det är en lista men det är en IP adress egentligen och inte en skrivare alltså och så får vi då en fil på det sättet som laddas in och det här programmet också, det kan, jag kan lägga upp en tid, och säga att just den här listan kan hämtas då och då. Fast egentligen är det stora tabeller då som den kör. Och sen så, kommer den i in i våran behandling på det sättet. Så man får fippla lite och hitta lösningar.

Ser du några nya sätt för inhämtning eller funkar detta bra, för framtiden alltså?

Det funkar ju bra men jag skulle kunna tänka mig att man kommer att jobba mer med data transaction server eller någonting sådant där och (kort paus) kanske jobba med någon form av tivoli finns ju till exempel, en sådan här router vara som kan hämta in och skicka runt och ut data igen, så att, men problemet är ju att jag har, vi, det här servicebolaget är det ju som gör alla dom här sakerna som har källdatat och allting och då sitter vi liksom på ett lokalt nätverk så att vi måste hitta en teknisk möjlighet för oss att jobba precis som vi jobbade i det företaget. Så att hittills har jag bara gjort så som jag har

gjort nu då och som jag sa till dig. Så att det ena heter FTP och det andra är CYKLOP, heter den program varan Så att i framtiden skulle jag vilja se någon slags transaktionsserver som man kan jobba med, mera så att jag har en riktig lina och så där va.

Då kommer vi in på integreringsfasen sen då. Vilka tillvägagångssätt använder ni er av vid integrering av den externa datan?

Ja, vi jobbar med (kort paus) vi har byggt en Visual Basic program, batch program, som tvättar och renar datat och laddar tabellerna, det är så vi gör.

Så ni använder er alltså av det programmet, och så gör den då i stort sätt allting då?

Ja, men innan dess så har den här cyklopen då gjort om dom här filerna till, från en fiktiv lista eller FTP, FTP-filen går också in via den här cyklopen som omvandlar de här då till sekventiella filer och skickar in det i via, ner på SQL server och så tar Visual Basic, då triggar det igång Visual Basic programmet då.

Det här programmet då, det gör transformering och sånt här då också?

Ja

Är det mycket, vilken sorts transformering är det som är vanligast? Vet du det?

(kort paus) ja, tvätta, rensa från skräp. Det är det som är det stora.

Är det bara det sättet ni använder er av?

Ja, det är det väl. Det är ganska Basic, det är inte så avancerat. Det blir bra ändå.

Känner du till några andra tillvägagångssätt än dom ni använder er av?

(lång paus) nej inte så där, det gör jag inte. Dom flesta, nej man kan ju ha, man kan jobba med olika programmeringsgränssnitt som DB2 tabeller och så kan man jobba med och det är väl mest det. DB2 och det, jag vet inte om DL1 förekommer så mycket. Men du kan ju ha dom här stordatortabellerna i grunden och jobba med i så fall och sen skicka data upp till, till, göra tabellerna och sen så skapa metadata mot andra slags tabeller, det beror ju på vilka källor som det här analysverktyget klarar också.

Hur lagras den externa datan i datalagret?

Ja, alltså, på SQL server.

Ja lagras den alltså integrerat i tabellerna eller har den egna tabeller?

Ja den ligger i den övriga informationen.

Så man kan inte utläsa vad som är externt och internt när det väl kommit in i warehouset kanske?

Nej det kan du inte göra, därför att då är det ju i form av då är det som försäkringsposter och då finns den informationen från SPAR till exempel i där då. Det är ju nya adresser till exempel, då behöver man ju inte säga att det kommer från SPAR det vet ju alla att dom riktiga adresserna kommer från SPAR så att säga och sedan har vi, finns det ju i och för sig en markering man kan använda som, i försäkringssystemen som säger om det är en manuellt inlagd adress men den gäller ju bara två dagar eller i en dag gäller den, för sedan så på natten går det ju mot SPAR igen. Så att liksom vissa saker är egentligen helt

värdelöst att veta om man säger (skratt) det behövs liksom inte när man vet att det kommer från SPAR. För ni pratar mycket om det här med extern data.

Ja det är egentligen det vårt arbete är inriktat på (skratt)

Ja jag förstår. Nä men om man säger så, att om du tar extern data från sådana här stora som SPAR och CBR, och så lägger du in det i källsystemen sedan så ligger det där. Sedan är det upp till själva datalagret egentligen och välja vilken information dom vill ha och är det då att syftet är att jobba med försäkringar så behöver du faktiskt inte veta om det är, att det just står SPAR eller CBR, man vet ju att det är en bilförsäkring så är det ju, då är det ju från CBR vissa delar.

Men det är alltså så ni gör egentligen när ni får det integrerat, ni lägger det i källsystemen först och sen sugs det med den interna datan? Är det så det funkar?

Ja, om inte, jag måste ändå ha en hel del kännedom om källsystemen för att kunna bygga ett datalager som består av kanske en femtio olika källsystem, det är liksom olika försäkringskattegorier, dom är ju liksom uppdelade i massor av system. Och visst måste jag ju ha en viss aning om var kommer det in externt ifrån men huvudsaken är ju då att veta hur försäkringssystemen behandlar informationen, alltså att jag vet, vad är det jag vill ha egentligen, vad är det jag vill visa med mitt datalager, om det kanske är mera produkt inriktat då så att det inte är så mycket om kundinformation då väljer jag produkten och då tänker jag inte i form av externt och data som finns i dom källsystemen jag ska hämta ifrån om du förstår hur jag menar. (kort paus) kanske hela ert arbete blev förstört här nu.

Absolut inte, det här blir jättebra, det är rätt begränsat nämligen hur alla använder det och hur dom använder det så all information är relevant för oss. Tror du att ni kommer öka graden av integrering av extern data? Eller det kanske inte ens går?

Det var som jag sa att det beror ju på, kommer det fler lagar och avtal som försäkringsbolag måste följa måste vi plocka in den informationen, det finns ju viss information som vi måste följa, vi måste följa person, personlig information och bilar också då med mycket, enligt, PUL finns inte så mycket längre men ändå. Vi måste fortfarande följa allting, alla förordningar som finns. Och då är det något mer vi måste veta så plockar vi in det också. För det är ju inte så kul för folk om dom får, det kan ju vara en skyddad adress, vi får inte skicka ut någonting till den här personen till exempel, då måste vi stoppa det. Det är inte kul att få hem det till exempel, eller om någon har avlidit så måste vi ha koll. Allt sånt där kommer från SPAR.

Ser du några tydliga trender i tillvägagångssätt för integrering annars? Förut gjorde alla så här men nu gör alla så?

(kort paus) det är väl det att man försöker börja, istället för att göra små datalager här och var i ett stort företag kanske man försöker börja göra ett stort detaljlager och sen ta mera marts, datamarts ifrån det. Lite mer åt det hållet. Det kan vi väl kalla en integrerings, det är ett annat sätt att tänka och det är för att man verkligen ska få en sanning av datat

Då kommer Markus med några frågor här beträffande användning.

Hej igen. Då kommer vi in på användning då som sagt, inom vilka tillämpningsområden använder ni extern data i erat datalager? Du nämnde något om kampanjer förut vet jag.

Ja just det, kampanjer. Kampanjer som utskick då då, men sedan också att man som säljare till exempel får en delmängd information om en viss kundgrupp som man ska gå ut och göra något arbete kring då till exempel. Eller (lång paus) fast det är inte mycket externt, det är ju det där med extern data som ställer till det lite. Ja man kan ju (lång paus) jag vet inte, ja det var ju det mest om man rent extern extern data för annars ligger ju extern datan som sagt var i försäkringsposterna och då annars använder vi det i form av att se hur mycket man säljer av en produkt då och en produkt är ju en försäkringsform då, ett försäkringsavtal. Man ser hur, till exempel man följer upp om man ger rabatter, hur utvecklas kunders beteende då. Köper dom mer av oss till exempel och vad köper dom mer av då, det är ju så, man kan försöka inrikta sig på olika grupperingar, segment av kunder som man jobbar vidare på då av olika slag men sen kan man också jobba med produkter och då är det hemförsäkring till exempel och så lägger man till olika som jag sa antigen testar man rabatter eller så tittar man på om man utvecklar en produkt och gör den lite bättre, vilken kundnöjdhet får man då osv. osv. men i grunden gäller att vi vet liksom var kunden bor och ja, kan vara bra att veta hur stor (kort paus) hur ska jag säga det då. Om det finns villor till exempel i en kommun i, vilka adresser, vilka postnummer ligger det inom och sådana där saker. Och sådan information får vi från SPAR ju, fast det kommer in via försäkringssystemet först och sedan till oss så det är aldrig någon direktlinje till SPAR

Ok, så det är ganska nödvändigt för att göra sådana analyser då?

Ja, det är det ju, vi måste ju ha upp, vi måste ju veta var kunden, presumtiva kunder bor för någonstans, om det hänt någonting, har någon avlidit ska vi ju inte gå på den personen då och skicka någonting och har dom flyttat så är det bra också att veta, kanske flyttar inom länet, kanske man kan erbjuda en annan produkt.

För dom här kampanjerna då, vad ser du att externdata ger för möjligheter där? Vad är själva bidraget för externdatan?

(kort paus) ja det kan ju vara att, ja, (tvekan) dels är externdata till för att vi ska veta var folk bor för någonstans annars hamnar vi helt fel när vi väljer folk som vi tror bor i villa men nu har ett helt annat postnummer, ja och så vidare. Så så är det ju. Likadant med bilar, att man kan välja ut en viss sorts bilägare till exempel och iden är väl att man erbjuder dom i till exempel i en kampanj en rabatt eller under visst tillfälle att man då får en ökning av försäkringarna helt enkelt.

Och det är där ni använder den datan då från bilregistret antar jag?

Ja, ja, att man får, har dom en bil så vill dom, då bor dom ju någonstans men vi verkar inte ha hemförsäkringen här till exempel, då skulle man kunna ta en kampanj och inrikta sig på dom och fråga är du intresserad av en hemförsäkring också eftersom du har en motorcykel bl.a. bl.a. bla. Så gör man.

Ser du några relaterade problem med någon utav dom här användningsområdena, dom här kampanjerna till exempel? För just externdatan då?

Nja, det är ju det där då att det kan vara lite gammalt data om någon person flyttar väldigt fort eller har sålt sin motorcykel dagen efter eller, då kan det ju bli lite pinsamt om man skickar ut information. Det gör det ju, men i det stora hela gör det ju inte det för just den här informationen vi får måste ju vara så exakt som möjligt det kan ju vara med ett annat företag som köper in information från sådana här mellanparter, dom kanske kan råka värre ut va men statliga myndigheterna måste ju skicka rätt information.

Finns det användningsområden som ni avstår ifrån om ni har otillräckliga resurser eller om det är för stora problem med någonting?

Användningsområden som vi avstår ifrån. (kort paus)

Som ni skulle vilja använda externdatan till men inte kan. Om det är några lagar eller någonting som sätter stopp för det, något?

(kort paus) Ja det finns ju, fast då handlar det inte om SPAR då, utan om man till exempel köper in det här fastighetsregistret, då är det ju, det är på kunder som vi inte har men vi vet att det bor en massa folk i villa och (kort paus) så kan man skicka ut ett utskick till i alla dom i den villan men sen, man får inte till exempel personnummer och sådant då va så man vet inte om dom redan är kunder hos oss till exempel så då kan det ju bli det. Man kan ju checka av det i och för sig va men det är inte säkert att det stämmer i alla fall och sådana grejjer. Visst det skulle vara kul om man kunde få all information om ickekunder men det får man inte, så att, men det får vi aldrig göra.

Ok, tror du att användandet av extern data kommer att öka i framtiden?

(lång paus) Alltså det finns ju möjligheter, det gör det ju, men jag tror inte att, vi måste följa så mycket lagar att det är lite stelt här. Man får inte göra vad som helst. Så för mig är det inte det men för andra mindre företag kan jag tänka mig att det gör det för det finns ju så stora möjligheter när man kan plocka in så mycket data från olika håll va.

Men inte för eran organisation då?

Nä, man får ju inte.

Ok, du ser inte att det kommer släppa på ngt sätt heller med lagarna eller på, att det förenklas på något sett för er som försäkringsbolag?

Nej det tror jag inte.

Jaha, då har vi väl egentligen bara avslutande frågor kvar då. Vilka generella fördelar ser ni med användning av extern data med erat datalager?

(kort paus) Ja det är att man får en säkrad källa på en gemensam grund för många användare, alltså bilinformation och SPAR information, man är ju då säker på den källan

Om vi då vänder på det, vilka generella nackdelar ser ni med extern data?

(Lång paus) Ja det är väl i sådana fall, det är ju samma sak då va, man litar ju på den då. Så att om den inte stämmer, om det är så att det inte vore nu statliga myndigheter utan någon annan extern källa då, man känner ju att, man vet ju inte tillräckligt mycket om en extern källa egentligen, kanske. Det kanske du aldrig kan få. Det är ju en viss osäkerhets procent där i då.

Appendix 2 – Transcribed material

Använder ni externdata till någonting annat i företaget som inte kommer in i data lagret?

Nej.

Ingenting?, Allt kommer in i datalagret?

Ja.

Är det något du har att tillägga?

(lång paus) Nej, inte om extern data men om data lager kanske som sagt (skratt)