Data Warehouse – An Outlook of Current Usage of External Data

HS-IDA-EA- 02-407

Marcus Olsson (a98marol@student.his.se)
Institutionen för datavetenskap
Högskolan i Skövde, Box 408
S-54128 Skövde, SWEDEN

Examensarbete på det dataekonomiska programmet under vårterminen 2002.
Handledare: Mattias Strand

Data Warehouse – An Outlook of Current Usage of External Data

Submitted by Marcus Olsson to Högskolan Skövde as a dissertation for the degree of B.Sc., in the Department of Computer Science.

2002-06-05

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: _____

Data Warehouse – An Outlook of Current Usage of External Data

Marcus Olsson (a98marol@student.his.se)

Abstract

A data warehouse is a data collection that integrates large amounts of data from several sources, with the aim to support the decision-making process in a company. Data could be acquired from internal sources within the own organization, as well as from external sources outside the organization.

The comprehensive aim of this dissertation is to examine the current usage of external data and its sources for integration into DWs, in order to give users of a DW the best possible foundation for decision-making. In order to investigate this problem, we have conducted an interview study with DW developers.

Based on the interview study, the result shows that it is relative common to integrate external data into DWs. The study also identifies different types of external data that are integrated, and what external sources it is common to acquire data from. In addition, opportunities and pitfalls of integrating external data have also been highlighted.

Key words: Data warehouse, External data

Table of contents

1 Introduction

Many organizations are today operating within increasingly competitive business environments. For them, to be able to face new challenges originating in the changing business and technology environment, it is of great importance to adapt to new requirements and market conditions in a rapid and efficient way. The deployment of IT and the usage of different kinds of information systems (IS) to support management functions has now become a necessity. New kinds of powerful systems have emerged for decision support: one type of IS that has attracted many organizations are the data warehouse (DW). A DW is a system, which integrates large amounts of data from several sources internal, as well as external to the organization (Singh, 1998). Internal data is data that is gathered and managed internally to a corporation's systems and usually begins its life as a by-product of transaction processing. Internal transaction data includes such things as sales data, shipment data, order data, and so forth. External data is data that is gathered outside of the corporation. External data might be gathered by looking at the sales at the cash register of a grocery store and on other occasions external data may be collected and stored in an aggregate form, such as the movement of the Dow Jones average throughout the month (Inmon, 1999). A DW is implemented with the goal of providing the organization with an integrated, consistent and subject-oriented view of its data, which can be used when performing various kinds of analyses and forecasting, for decision support.

Unfortunately, many organizations have experienced problems when implementing DWs and investments have become costly, without satisfying requirements or living up to expectations (Barquin & Edelstein, 1998). One problem that has been highlighted is that data warehouse development focuses too much on traditional internal data, ignoring the potential value of external data (Barquin & Edelstein, 1998). The majority of data that was interesting to an organization had, until recently, its origin within the own business. The interest in external data was relatively insignificant (Devlin, 1997). However, Inmon (1996) claims that it is important to pay attention to the inclusion of external data when developing DWs. Such inclusion may provide with novel information and novel insights that is not impossible to render from the internal sources. External data that is interesting and important to an organization can provide some level of market-share information and could for example be data that includes economic forecasts, political information, consumer demographics, supplier information, and competitive and purchasing trends (Singh, 1998). Still, although many authors highlight the importance of external data, there is not much literature describing what external data that is important and from which sources it is acquired.

Therefore this final year project intends to investigate the current usage of external data and its sources for integration into DWs, in order to give users of a DW the best possible foundation for decision-making. The work is based on an interview study that has been conducted with DW developers. The results of the interview study shows that external data is relatively frequently used in DWs and the interest will probably increase in the future. However, it was also shown that there still exist problems that need to be solved, e.g. immaturity of organizations usage of DW and the management of internal data, before external data may become a natural component of modern DWs.

The structure of this dissertation continues with chapter 2, which is aimed to give an introduction to data warehouses. The chapter provides the reader with the evolution and definition of data warehouses, a comparison between DW and OLTP systems, data warehouse architecture, and description of internal/external data, and lastly, a brief overview of the DW development process are given. In chapter 3, the problem definition of the work is described. Furthermore, delimitations are also presented. Chapter 4 includes a description of relevant methods for answering the research problem. In addition, chapter 4 also describes and motivates the method chosen. The report continues with chapter 5, which gives an account of the implementation of the method chosen. In chapter 6 we synthesize the material that has been gathered during the interviews. Chapter 7 uses chapter 6 as input to analyze the material and information collected in the empirical survey. This is done on basis of the research problem for the dissertation. Result and conclusions that can be drawn from the examination is presented in a concise way in chapter 8. The report is concluded in chapter 9, by discussing the contents in this thesis in a wider context. This chapter also includes our own experiences during the work and finally, suggestions for future research is given.

2 Data warehouse

This chapter gives an introduction to the area of data warehousing and defines central concepts for the dissertation. The intention is to provide the reader with relevant background information to the problem area. Firstly, the evolution of the data warehouse concept is described. Then the definition of a data warehouse is discussed, along with a description of characteristics of the data stored in a DW. The next section includes a comparison between data warehouses and online transaction processing (OLTP) systems. After this a description of data warehouse architecture is provided, followed by a declaration of the difference between internal and external data. Finally, the development process of a data warehouse is described briefly.

2.1 Evolution of data warehouse

A typical organization has numerous operational systems, but those systems are not designed to support strategic decision-making. However, in order to be able to regain competitive advantage, organizations have become more and more focused on incorporating novel ways to use operational data to support decision-making (Connolly & Begg., 2002). There are several reasons why existing operational systems could not meet these needs. Singh (1998, p.16) mentions the following:

- They lack on-line historical data
- The data required for analysis resides in different operational systems
- The query performance is extremely poor which in turn impacts performance of operational systems.
- The operational database designs are inadequate for decision support.

The concept of the data warehouse has evolved to solve these problems and to meet the requirements of a system capable of supporting decision-making, to be able to make comprehensive analysis of the organization and its business, and future trends. To facilitate this type of analysis, strategic decision makers require access to, not only current values in databases but also historical data, and multiple data sources, wherever located. To handle the needs that Singh (1998) points out, a data warehouse stores the current and historical data, that business decision maker's need, from secluded operational systems and external sources into a single, consolidated system. This system could access the data needed without interrupting the on-line operational workloads (Singh, 1998). The challenge for an organization is to turn its archives of data into a source of knowledge and to present a consolidated view of the data to the user (Connolly & Co, 2002). This idea of presenting a unified view of the organization data, which may be presented to business users, is the main motivation for implementing a DW.

2.2 Defining data warehouses

It is hard to find an unambiguous definition in literature, which makes it difficult to clearly define the concept. Therefore this section will provide with several definitions to show on different views of what a DW actually is. We will not argue in favor of, nor choose, any definition to use as a basis for this dissertation. The reason is that we do not attach great importance to a specific definition of a DW, with respect to the

focus and problem definition of this dissertation. Still, it was considered as important to show on the richness of different definitions.

There are numerous definitions of the data warehouse. The focus with earlier definitions was on the characteristics of the data held in the warehouse. Other definitions widen the scope of the definition to include the processing associated with accessing the data from original sources and the delivery of the data to decision makers (Connolly & Begg, 2002). According to Poe & Co (1998, p 18), *"A data warehouse is an analytical database that is used as the foundation of a decision support system. It is designed for large volumes of read-only data, providing intuitive access to information that will be used in making decisions"*. The purpose of a data warehouse is to ensure that the appropriate data is available to the appropriate end user at the right time (Poe & Co, 1998).

Another definition of what constitutes a data warehouse is stated by Devlin (1997, p 20): *"A single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use in a business context"*. Devlin (1997) says that achieving completeness and consistency of data is far from an easy task. In the context of the business, this means understanding the business strategies and the data required to support and track their achievement.

Barquin & Edelstein (1997, p 5) focus more on the data warehousing process than on the data warehouse itself, and suggest the following definition: *"Data warehousing is the process whereby organizations extract value from their informational assets through the use of special stores called data warehouses"*.

Chaudhuri & Dayal (1998, p 65) describe data warehouse as follows: *"Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions"*.

One of the primary pioneers in the development of the data warehouse concept is W.H. Inmon, who is frequently referred to in literature. In his work he offers a constantly recurring definition of the data warehouse: *"A data warehouse is a subject oriented, integrated, non-volatile, and time variant collection of data in support of management's decision"* (Inmon 1996, p 33). These characteristics are clarified to make it easier to understand the definition:

- *Subjet-oriented* means the data warehouse is logically organized around major key subjects of the corporation, e.g, customers, sales or items produced (Inmon, 1996).
- *Integrated* is the second prominent characteristic of the data in a DW; this is the most important aspect (Inmon, 1996). The data is integrated because of the combination of source data from different enterprise-wide applications systems, derived and aggregated data, historical data and data from external sources. Inmon points out that the source data often is inconsistent, using different formats, for example in naming conventions and variables. As the data enters the warehouse, the integrated data source must be made and stored in a consistent format in order to be able to present a unified view of the data to the users (Connolly & Co., 2002).
- *Time-varying* because the data in a data warehouse contains a time dimension, and is only accurate and valid at some point in time or over a time interval.

- *Non-volatility* is the last important characteristic of a data warehouse and means that the data is read-only. It should not be updated, modified or changed by end users. The data is loaded and accessed, but it doesn't change thereafter (Inmon, 1996). New data is not added as a replacement, but added as a complement and integrated incrementally with the previous data (Connolly & Co., 2002).

Many authors' remark that "data warehousing" is a broader term than "data warehouse". It is used to describe the whole process and includes extraction, creation, maintenance, use and the continuous refreshment of the data in the warehouse (Watson et al., 2001., Singh, 1998., Barquin & Edelstein, 1997). The precise definition of a data warehouse is debatable, though it means different things to different people. However, it is agreed by researchers, academics and vendors alike that data warehouses are built to support business decisions (Arun & Varghese, 1998). Söderström (1997) means that the concept data warehouse is used on many counts and has different meanings in various contexts. Therefore it's an important matter for individual businesses to declare what data warehousing means within their own organization (Söderström, 1997). We agree with Söderström. It is important for an organization to really make sure what it intends to achieve and what the objectives are, when implementing the ideas of the concept data warehouse. Nevertheless, it is not of great significance for the aim of this dissertation to clarify our own view of specific data warehouse definitions. What is important, with respect to this work, is that regardless of data warehouse definition adopted there is an agreement that data from several internal and external sources is acquired and integrated.

2.3 Comparing Data warehouses and OLTP systems

This section will provide (see Figure 1) a comparison of the major characteristics of online transaction processing (OLTP) systems and data warehousing. The aim is to explain how data warehouses differ from operational systems, from what they contain to how and when they are used. Also, characteristics of data warehouse data will be discussed. This can also be related to how Inmon (see section 2.2) defines the characteristics of the data to be included in a DW. We find it important to stress these differences in a comparison to traditional OLTP-systems, given the focus of this report.

To discuss data warehouses and distinguish them from OLTP systems calls for an explanation with respect to what set of requirements each type of system is designed to meet. OLTP systems are designed to support and maximize on-line transaction processing capacity, which includes insertions, updates and deletions, while also supporting information query requirements (Connolly & Begg, 2002). These systems are optimized to automate business operations and must be efficient with transactions that are predictable, repetitive and update intensive. They are organized around business processes and the transactions they perform, such as entering orders or updating inventories (Barquin & Edelstein, 1997). By contrast, data warehouses are designed to support efficient processing and presentation for analytical and decision-making purposes (Elmasri & Navathi, 2000). DWs provide information to users so that they can analyze a situation and make decisions (Poe et al., 1998). A data warehouse holds data that is current, historical, detailed and summarized to various levels. Apart from being supplemented with new data, the data in a DW is rarely subject to change (Connolly & Begg, 2002). Another difference is that the number of

users served by a DW tends to be smaller than for OLTP systems. In part, the smaller number of users is derived from the nature of a transaction in an OLTP system, and from the way that a DW is designed (Barquin & Edelstein, 1997). The DW is designed to support relatively low numbers of unpredictable transactions that require answers to queries that are unstructured, heuristic and *ad hoc*. The data in a warehouse is organized according to the requirements of potential queries and supports strategic decisions of managerial users (Connolly & Begg, 2002). Furthermore, the time horizon for holding the data in a data warehouse is significantly extended compared to that of OLTP-systems. Normally the time horizon for a data warehouse is 5-10 years, whereas an OLTP system holds its data 60-90 days (Inmon, 1996).

Figure 1 is provided to show the comparison of OLTP system and DW in a more comprehensive way.

OLTP systems	**Data warehouses**
Holds current data	Holds current and historical data
Stores detailed data	Stores summarized and detailed data
Data is dynamic	Data is largely static
Repetitive processing	*Ad hoc*, unstructured, and heuristic processing
Transaction-driven	Analysis driven
Application oriented	Subject-oriented
Predictable pattern of usage	Unpredictable pattern of usage
Supports day-to-day decisions	Supports strategic decisions
Servers large number of operational users	Serves relatively low number of managerial users

Figure 1: Contrasting OLTP systems and data warehouses (Connolly & Begg., 2002)

To summarize this section, OLTP systems have their own databases and are used for transaction processing; a data warehouse is a separate system and its used as a support for decision-making (Watson et al., 2001). The fundamental distinction is that OLTP systems drive business operations on a day-to-day basis, whereas a decision support system like the data warehouse determines the outcome of the decisions in the business environment related to customers, products, suppliers etc., and the timing of exchanges between them and the firm (Agosta, 2000). However, one must remember that while OLTP systems are radically different from data warehouses and are built with other purposes in mind, these systems are closely related in that the OLTP systems provide some of the source data for the warehouse (Connolly & Begg, 2002).

2.4 The Data warehouse architecture

In this section, a description of the architecture and major components of a data warehouse is given. The purpose is to present the technology behind the concept of the data warehouse. The typical architecture of a DW is shown in Figure 2. The figure is based on Chaudhuri & Dayal (1996) and gives a good overview of the DW environment and the DW as a whole. In this work, the focus is not on the technology behind data warehouses, but it is appropriate to outline the general architecture, since it thereby positions the external data in the DW environment.

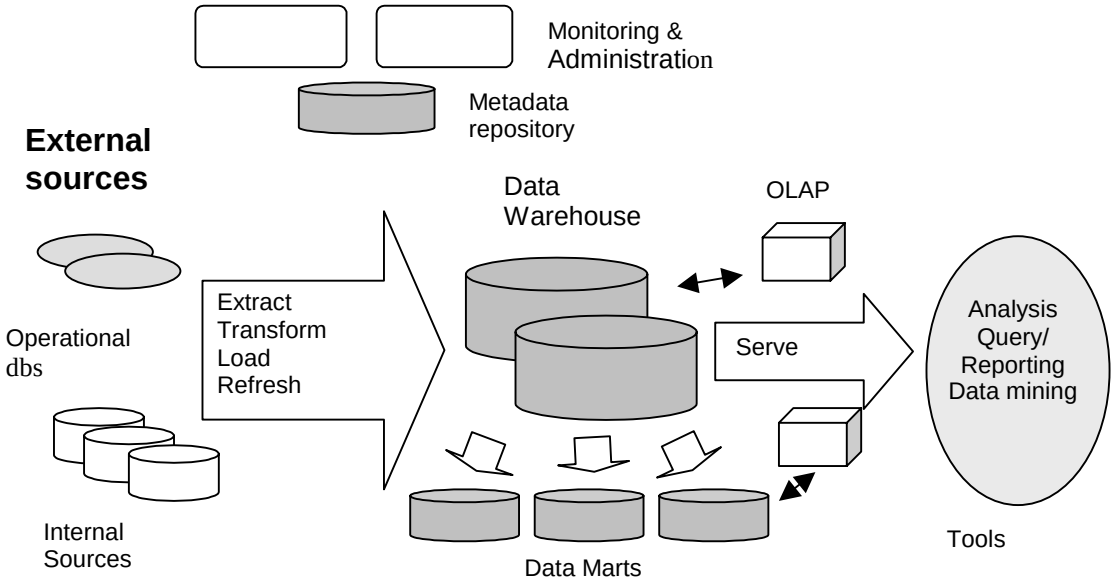


Figure 2. A typical data warehouse architecture (from Chaudhuri & Dayal, 1996, p.66)

Architecture is a set of rules or structures providing a framework for the overall design of a system or a product (Singh, 1998). Poe et al. (1998) state that there are networking architectures, client-server architectures, architectures for specific products as well as many others. Further, the authors describe a *data architecture* as providing a framework by identifying and understanding how the data will move throughout the system and how it will be used within the corporation. A data warehouse has a primary component for the data architecture, which is a read-only database. Different architectural alternatives exists, Figure 2 (Chaudhuri & Dayal, 1996) shows a possible DW data architecture. It shows how data from several sources, operational systems as well as external sources, are integrated into a data warehouse. It also illustrate that in addition to the main warehouse, there may be some departmental data marts. The data is managed by servers, which present multidimensional views to different kinds of analysis tools. As shown in Figure 2, there is also a repository and tools for monitoring and administering the system.

As system architecture, the data warehouse firstly includes tools for extracting data from multiple operational databases and external sources (Chaudhari & Dayal, 1998). These tools, often referred to as *backend tools/components* or, alternatively, *the load manager*, performs all the operations associated with the extraction and loading of data into the warehouse (Connolly & Begg, 2002). Chaudhari & Dayal (1998) point to and associate the following operations with these extracting tools:

- cleaning, transforming and integrating the selected data; preparing the data for entry into the warehouse
- loading data into the warehouse
- periodically refreshing the warehouse to reflect updates at the sources

The data will be extracted from the source systems and converted to the data warehouse. Often, source data comes from operational databases (day-to-day processing), but note that the data may also come from outside the organization, for instance from companies that specialize in selling data to other corporations. The

distinction between internal and external sources will be explained in section 2.5. According to Connolly & Begg (2002), the majority of data for the data warehouse comes from sources held in first generation and network databases. Connolly & Begg also advocate the use of data supplied from external systems such as the Internet, commercially available databases, and databases associated with an organization's suppliers or consumers.

The central component in a data warehouse system is a separate database designed specifically for decision support (Poe et al., 1998). In addition to this database, there may exist departmental data marts (Chaudhari & Dayal, 1997). They are built to focus and meet particular needs, within for example a specific region, department or function of an organization (Barquin & Edelstein, 1998). Singh (1998) feels that a data mart is a subset of the enterprise-wide data warehouse and that the organization may build a series of data marts over time, as part of an iterative development process of a data warehouse. To store and manage data in data warehouses and data marts, one or more warehouse servers are used. These servers present multidimensional views of data to a variety of data access and retrieval tools. These tools are referred to as front-end tools and consist of query tools, report writers, analysis tools, and data mining tools (Chaudhari & Dayal, 1998).

According to Singh (1998), one of the most important components of a data warehouse is metadata, defined as data about data. Connolly & Begg (2002) state that metadata is used for a variety of purposes. Firstly, in the extraction and loading processes meta-data is used to map data sources to a common view of the data within the warehouse. Secondly, in the warehouse management process, meta-data is used to automate the production of summary tables. Lastly, as a part of the query management process meta-data is used to direct a query to the most appropriate data source. In short, meta-data plays an essential role in the data warehouse environment.

Furthermore, in order to access the DW data in an efficient way, it is necessary to understand what data is available and where that data is located in the warehouse. Metadata helps to locate desired data and provides a catalogue of data in the data warehouse and the pointers to this data (Singh, 1998). The data warehouse architecture includes a repository for storing and managing all the meta-data associated with the warehouse. This enables the sharing of meta-data among tools and processes for designing, operating, using and administering a data warehouse (Chaudhari & Dayal, 1998).

There are different approaches and methods for analyzing the data in a data warehouse. Given the scope of this work, these tools referred to as front-end tools, will not be described in detail. However, we will briefly describe some of the more relevant techniques of front-end tools that are used to interact with the data warehouse. Connolly & Begg (2002) categorize these front-end tools into five main groups:

- Query and reporting tools
- Application development tools
- Executive information system (EIS) tools
- Online analytical processing (OLAP) tools
- Data mining tools

Query and reporting tools are mostly used to track day-to-day operations to support tactical business decisions (Singh, 1998). *Production reporting tools* are best suited for retrieving operational data and generating operational reports using different formats and layouts. *Query tools* are only efficient for less complex situations, and are designed to manage elementary SQL statements to query data stored in the warehouse (Connolly & Begg, 2002). *Application development tools* are more advanced than the previous category, and can be used where query and reporting tools are inadequate. For example, they can be better graphical data access tools. The third group in Connolly & Begg's categorization is *Executive information (EIS) tools*. These tools were originally developed to support high-level strategic decision-making but have evolved to offer support for all levels of management. These systems are aimed at presenting a high level of user friendliness and providing advanced functionality, while hiding the complexity of the underlying systems and data structures from the users. *On-line analytical processing (OLAP)* is a term for multidimensional analyzing tools. OLAP allows users to view and analyze data across multiple dimensions, hence the term multidimensional analyzing. The tools associated with this group allow users to analyze and slice and dice data across multiple dimensions such as time, market, and/or product category. According to Singh (1998) OLAP tools are most suited to analyzing and forecasting trends and to measuring the efficiency of business operations over time. The last category, *Data mining tools*, helps in extracting new correlations, meaningful patterns and trends, which prior to the search were not known to exist or were not visible, in large amounts of data. Statistical, mathematical, and artificial intelligence techniques are used to achieve this. Bichoff (1997) describes data mining in simpler terms: data mining asks a processing engine to show answers to questions we do not know how to ask.

2.5 Internal and external data

It is important, according to the scope of this work, that we define what we mean by internal and external data, and how this dissertation intends to use the terms, in order to prevent confusion.

External data refers, in this dissertation, to data that originates from outside the organizational boundary. By this, we mean that external data is data that is sourced from outside databases and services within the corporate environment. This data is not generated from the corporation's own systems, but is considered useful enough to be included in the data warehouse. Internal data is data selected and obtained from existing business systems within an organization. Singh (1998) states that internal source data comes from on-line transaction systems, which are deployed by the enterprise.

Devlin (1997) is of the opinion that the majority of data of interest to an organization in the past has originated from the organization. Data that derives from inside the environment of an organization is referred to as internal. Devlin defines external data as follows: *"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."* (Devlin, 1997, p. 135). In practice, Devlin states, that external data must be subject to a formal acceptance process before it is integrated and being used within the company. The author contends that the management of external data differs from that of internal operational data in some respects. This is because the receiving company usually has less control over the structure or content of data

obtained from outside the organization than it has over its own data. However, Devlin remarks, that the real problems are not technical. Rather, it is primarily an organizational issue when a data warehouse looks like an overstuffed drawer; this is undesirable. Business users must be willing to sacrifice some freedom of choice in obtaining external data, in order to capture only data that is designed for innovative uses and to ensure the overall consistency of such data within the company. There is a need to understand the consequences of arbitrarily including external data.

Inmon (1996) claims that, external information doesn't say anything directly about a company, but can give a lot of valuable information about the universe that the company must work and compete in. When comparing internal data to external data one of the most useful things is to compare the types over a period of time. The comparison allows management to *"see the forest for the trees"* (Inmon, 1996, p. 272). In other words, the ability to gain insights not possible without contrasting personal activities and trends against global activities and trends. The comparison must be made on a common key. There are many diverse types of data that come from external sources; Inmon (1996) points out some typical sources of interesting data to include the following: *business newspapers like Business Week and the Wall Street Journal, industry newsletters, technology reports, reports generated by consultants specifically for the corporation, competitive analysis reports, marketing comparison and reports, sales analysis, and new product announcements.* Bischoff & Alexander (1997) also state that there is a real need to purchase external data from outside sources. These authors argue that external data and internal data need to be merged, in order to answer individual queries. They mention several different types of external data. A business may extract data to enhance customer information, such as demographic data, life-style data, and data in response to questionnaires circulated by outside vendors, which makes sense to include with customer profiles. Another type of external data comes from government bodies or industry data providers (Bischoff & Alexander, 1997). Nevertheless, according to Inmon (1996), external data is harder to acquire, systemize and manage in comparison to internal data. This is because external data usually enters the corporation in an unstructured, unpredictable format.

Another factor is the form of the external data, in order to fit in the warehouse it must be reformatted and transformed into an internally acceptable and usable form. Unlike internal data, there is no real pattern of appearance of external data. The problem with this unpredictable frequency of appearance is that constant monitoring must be set up to ensure that the right data is captured. External data may be available and come from practically any source at almost any time. These factors make external data harder to systematize and manage.

According to Singh (1998), one of the clear goals with a data warehouse is to free the information that is locked up in the internal operational databases and to mix it with information from other external sources of data. The author further advocates that organizations increasingly acquire additional data from outside there own databases. Generally this includes some level of market-share information and could for example be information that includes economic forecasts, political information, consumer demographics, and competitive and purchasing trends. The Internet is one factor that is providing access to more and more data resources every day. The development of the Internet as a distribution- and exchange channel of information has increased both the quantity and the quality of various data sources. This has led to an increased amount of external suppliers of data and new selection possibilities (Singh, 1998).

Another issue concerns the storage of the external data. Inmon (1996) states that external data may be stored in the data warehouse if it is desired, convenient and cost effective to do so. But the author also points out that in many cases it will not be possible or economical to store all the external data in the DW. The alternative is to make an entry in the metadata of the warehouse, describing where the actual body of external data can be found. In this way the external data is stored elsewhere, where it is convenient, in for instance a filing cabinet, on magnetic tape, and so on. Bischoff & Alexander (1997) also point to the fact that it is not a good idea to intermingle some types of external data and internal data in the same tables in a database. The quality of external data may be questionable, or dynamic and subject to change on a regular basis.

2.6 Data warehouse development process

This section intends to give a brief overview of the data warehouse development process. In this work, the focus addresses early stages in the process, where the selection of relevant external data occurs. Therefore, this section will mainly discuss and concentrate on source system analysis and how to identify and select valuable external data, and only provide a brief survey on other tasks and later stages. We will introduce a generic methodology that describes one possible way to build a useful DW (Figure 3). With methodology we refer to a formal definition of the processes required to bring an IT solution from an initial idea to a useful result (Bischoff & Alexander, 1997). Experts say that data warehousing is *"a journey, not a destination"* in order to point out to that the development must be an iterative process and also to emphasize it's constantly evolving nature (Watson et al., 2001). In addition to the requirement that DW projects must be iterative, business needs must continually be reflected. There are numerous descriptions of data warehouse development processes in the literature. Even though we present a brief overview of the development process, it is important to recognize that this is a generalization that doesn't describe every company's experience or every book in the literature. We here outline the major steps in the general flow that occur during a data warehouse implementation. However, some of the activities (as shown in Figure 3) are happening concurrently and we do not attempt to reflect an absolute project timeline for the tasks.

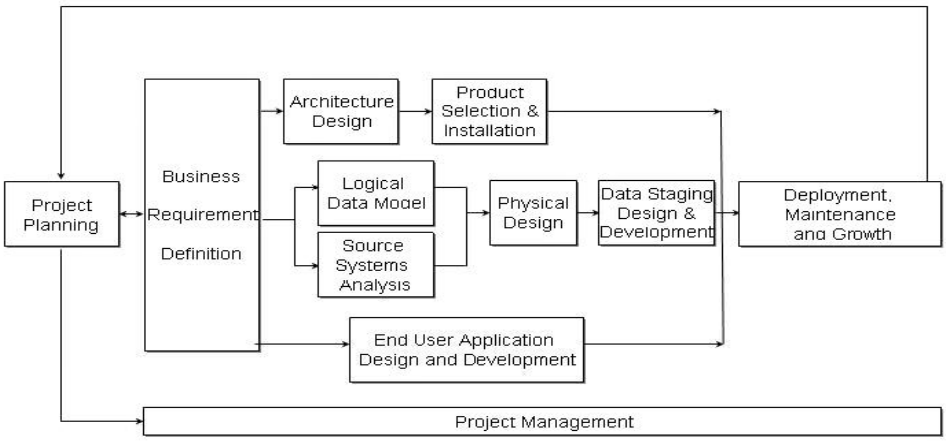


Figure 3. A generic development process of a data warehouse (from Kimball et al., 1998, p.33)

Data warehouse planning and project initiation
The planning activity is always an important aspect of a project, and the key to initiating the project is classifying the project. This includes focusing on resource and staffing requirements, coupled with the selection of the relevant tasks and activities (Kimball & Co., 1998). One of the first things to be clarified in the development is the business objective of the data warehouse. The business objective should be clear and will effect what gets developed, what data goes into the data warehouse, and the parameters for success (Poe & Co., 1998).

Business requirements definition
The next steps entail definition of business requirements. Developers must effectively determine business requirements and decision-maker needs, and translate them into design considerations. Reviewing the required data elements leads to the identification of information requirements. The definition of these requirements determines the data needed to address business users' analytical needs. With a sound understanding of the business end users and their requirements, a data warehouse's likelihood for success is greatly increased. The business requirements establish the foundation for the focus of technology, data, and end user applications (Kimball & Co., 1998).

Data warehouse logical data model
If not already made, a subject area analysis is accomplished to refine the information requirements. This means that the level of detail required in the subject area is identified, the content of the subject area is verified, and the DW data model is initiated. A subject area usually covers a particular aspect of the business: for example, sales information, or customer information (Bischoff & Alexander, 1998). The data warehouse logical model is developed from the subject area data model. The data model is the starting point for the design of the data warehouse environment and the data model acts as a roadmap for development.

Source systems analysis
As the data model is developed, a more detailed data analysis of relevant source systems and any special considerations related to them are conducted. This step is referred to by Bischoff & Alexander as source system analysis. It determines where the data will come from. A common problem with data warehouses is discovering the end-user's requirements, wants and needs. When conducting interviews with end-users, one of most general problems is that they don't really know what they want. A common answer is that they want "everything". However, all data are not equally valuable. For this reason there needs to be a process that examines candidate data sources from the perspective of their business value (Inmon, 1996).

Many large organizations have long been data rich and information poor (Barquin & Edelstein, 1997). This implies that, as an organization develops a data warehouse, it is of great importance to consider carefully what to put into the DW and the sequence for doing so. According to Kimball & Co (1998), two levels of analysis must occur. First, there must be a clear understanding of the data that has been requested by the business to be available in the data warehouse. This helps to select the data sources. Second, an in-depth understanding of each of the data sources that are to be included in the data warehouse must be gained. Candidate data sources should be listed with the user requirements. Systems and files that are candidates for providing data are evaluated in terms of quality, integrity and operational problems. Many sources may

contain the right data element names, but not the elements desired (Bischoff & Alexander, 1997). In some cases, more than one system will offer a viable candidate. Each source system is rated as to the risks and advantages of its use; the quality, accuracy, and timeliness of the candidates is weighed to complete the evaluation. Bischoff & Alexander (1997) also point to the fact that new attributes, not originally required, may be available in the source system and be considered as valuable. There may also be attributes that are required that cannot be found in the candidate source systems. This means that the data model must be updated accordingly. Kimball & Co (1997) says that a small number of actual data sources must be identified as the primary focus for the first phase project. The authors do not recommend tackling too many at once. They prefer to begin with the primary data sources, which are often driven by the business. This is the recommended place to start, but the decision must also be made as to from where the data will be extracted.

As stated above, it is important to consider what to put in the DW; the task here is to determine what to extract from internal and external systems. As described in section 2.5 it is advisable to seek for external sources of data, at the Internet or examine relevant external commercial databases, which becomes more and more important to gain a competitive advantage. Inmon (2002) remarks that the following issues should be considered if a source is external data:

- What external data will be accessed and what volume of that data can be expected?
- Will the external data have to be integrated with other data residing in the warehouse?
- How much will the external data cost?
- Are there intellectual property or security restrictions on the use of external data?
- Will the external data have to be edited for quality prior to entry into the data warehouse?

Architecture design
Developing the DW architecture definition (if the general warehouse architecture has not already been developed) establishes the technical and application infrastructure for the data warehouse. The choice of architecture will determine many aspects of how the system is developed, such as the tools, platforms, databases, communications, training and so on (Poe & Co., 1998). Data warehouse environments require the integration of numerous technologies. Kimball & Co. (1998) state that the establishments of the technical architecture design is done simultaneously with the following factors – business requirements, the current technical environment, and planned strategic technical directions.

Physical design
In this step the DW database is physically specified and set up to support the logical database design. This includes creating the tables and setting up the database environment.

Data Staging Design and Development
Once the source systems are identified the next project task is to get data from the source to its destination, while maintaining accuracy and integrity. The data staging design and development process has three major steps: extraction, transformation, and load (Kimball et al., 1998).

Product Selection and installation
By using the technical architecture design as a framework, specific architectural components such as the hardware platform, database management system, data staging tool, or data access tool are evaluated and selected. The products will then be installed and thoroughly tested, to ensure an appropriate data warehouse environment.

End User Application design and development
At this point the development process moves to end-user tools and the types of end-user access to the DW are specified and laid out. The development of end user applications follows requirements and specifications that are defined by the development team and business users (Kimball et al., 1998).

Deployment, Maintenance and Growth
Deployment requires extensive planning to ensure that all puzzle pieces fit together; it represents the convergence of technology, data, and end user applications accessible from the business users' desktop. Before any business users have access to the warehouse they must be educated, and user support and communication should be established. Deployment should be deferred if all pieces are not ready (Kimball et al., 1998). After the initial deployment of the data warehouse there is still a need to focus on business users by providing them with ongoing support and education. It is also important to ensure that processes and procedures are in place for effective ongoing operation of the warehouse. Kimball et al. feel that a DW is bound to evolve and grow and that processes must be established to deal with this business user demand for evolution and growth.

Project management
As in all projects, management is important to ensure that the development process remain on track and in sync. Project management activities occur (as illustrated in Figure 4) throughout the development process, and focus on monitoring project status, issue tracking, and change control to preserve scope boundaries (Kimball et al., 1998).

3 Problem description

This chapter describes the problem area and the research problem of this dissertation. Delimitations are also included.

3.1 Problem area and justification of research problem

Technology is not the only matter when developing a DW. Many organizations have been stunned by the impact of poor data quality or badly performed data integration. The value of a DW comes from the use of the data stored and therefore it is critical that the optimal sources of data are identified and used. Those sources may be internal or external to the organization (Bischoff & Alexander, 1997). There is nothing new about external data and internal data. But what is new is the need to integrate these two forms of data. With data warehousing, and in particular exploration processing inside a data warehouse, the mixing of external data and internal data becomes very inviting. When a decision-maker is doing analysis of critical factors for the corporation, it may be very helpful to be able to compare internal numbers to external numbers. There is then a real reason behind needs to compare and analyse external data along with internal data when it comes to analysis at a strategic level (Inmon, 1999). Until recently, organizations have merely extracted and integrated internal data into their DWs and they have thereby missed the opportunity to fully exploit the potential of the data warehouse (Devlin, 1997). It is agreed upon, amongst DW managers that for organizations to be able to fully exploit the potential of their data warehouse there is a need to include external data, since it may provide with insights that is not possible to render from only the internal data (Inmon, 1996).

Still, even though many researches and books argue for the importance of integrating external data, there is not much literature elaborating upon external data usage and the descriptions of external data are on a general level. In addition to this deficiency, the literature that considers the usage of external data is solely originating from US. It is a common fact that development and research about DWs in the US has been around for a longer time and is more mature, in comparison to the Scandinavian market. In Sweden, we have not found any published material about the problem, and the current usage and integration of external data in DWs has not devoted much attention. Therefore, research focusing on the acquirement and the usage of external data may contribute with new valuable insights.

3.2 Research problem

In this section, we specify the aim and the objectives for this work. The comprehensive aim of this dissertation is to:

Examine the current usage of external data in data warehouses.

With the concept usage we refer to the acquirement and integration of external data into DWs, and not in what way external data is used to present analyzing data for decision-making. To be able to reach this aim, it requires us to identify to what extent that external data is acquired for integration, and what external data, if any, that is integrated. In order to reach the aim we will also investigate the types of external domains that are most interesting to organizations, and what external data sources that are most common to acquire data from. In a comprehensive survey of the current

usage of external data in DWs we also found it important to get insights to opportunities and pitfalls with integration of external data into DWs. In order to fulfill the aim stated above is our objectives therefore to answer the following four part questions:

- *To what extent is external data acquired for integration into DWs?*
- *What external data, if any, is integrated into DWs?*
- *What external sources are the most common to acquire data from?*
- *What are the opportunities and pitfalls with integration of external data into DWs?*

3.3 Delimitation

The focus of this dissertation is to review the current usage of external data for integration into DWs. The study is delimitated to what external data, if any, that is integrated and what types of external sources that are most common, and will not consider how the selection of what internal data in a data warehouse occurs. Furthermore, technical aspects and the actual integration of external data into DWs are outside the scope of this work.

4 Method

In this chapter, we describe different methods that may be suitable to collect information, in order to answer the research problem and fulfill the objectives stated in chapter 3. To be able to answer the problem, it is necessary that the working process is based upon a method. A method is a procedure that gives a description on how to tackle a problem and may be used to collect, process and to summarize information to acquire knowledge in a certain subject field (Andersen, 1998). This chapter begins to describe possible approaches for this work. Then, a discussion of different techniques that may be appropriate to collect data will be presented and described. The chapter is concluded with a description and a motivation for the approach and techniques chosen.

4.1 Different approaches in research

The type of research influences the choice of data collection technique, and in addition this is also based on different approaches in research. There are, according to Andersen (1998), two main forms of approaches within scientific research. These are generally characterized into either quantitative or qualitative approaches. These approaches differ in the way collected data is processed and analyzed. The choice of which way to use, is based on how the problem is specified and how the collected data should be processed (Patel & Davidson, 1994). Despite that the type of approach in research mainly is dependent on how data and information is processed, we have the opinion that it is important to have this clear from the beginning to characterize the whole research process.

The quantitative approach is referred to research that uses statistical analyzing methods. There is a clear guideline on how to put research into practice and it makes frequent use of statistics, mathematics, and arithmetic formulas (Andersen, 1998). A quantitative study is usually based on large amounts of numerical data that is more precise, with the goal to achieve a general result. In contrast, the central point in the qualitative approach is to create a deeper comprehension of the problem area (Andersen, 1998). This approach requires the usage of verbal analyzing methods (Patel & Davidson, 1994). Qualitative methods do not use numerical data, though this type of data cannot meet the main purpose of this approach. That is, to exemplify and achieve a deeper understanding (Andersen, 1998). The qualitative and the quantitative methods are often referred to as opposites. For example, it is a general opinion that a measurement is of secondary matter when using a qualitative approach. However, to be more precise it is almost impossible to avoid data given by measurements or number likewise in usage of qualitative methods (Repstad, 1993). Much of today's research is to be found somewhere in between these ends. Research that primarily targets towards quantitative aims often includes qualitative features, and vice versa (Patel & Davidson, 1994).

For this work, an approach that is mainly qualitative will be used, since the research problem is concerned with interpreting and understanding DW developer's experiences in developing DWs. The applicability for adopting a qualitative approach for this kind of problems is also pinpointed by Patel & Davidson (1994). Furthermore, the aim of the research is not to give statistical answers or to achieve any measurable results, but to describe the ins and outs of integrating external data into DWs.

Therefore, a quantitative approach was not considered as appropriate. In addition, considering the research problem, no numerical values will be collected and this implies that a verbal analyze is most appropriate. Nevertheless, quantitative aspects may occur, since it is hard to strictly apply one approach without being influenced by the other.

4.2 Data collection techniques

In research, there are different ways to collect data and information. Several techniques may be considered as useful for this research. The technique that is chosen is the one that appears to have the best opportunity to answer the problem. However, in relation to the technique(s) chosen, it is also important to consider time and resources that are given (Patel & Davidson, 1994). The techniques that have been estimated and possible to use from our point of view, with respect to our research problem and resource limitations are the following:

- Interviews
- Questionnaires

In the following sections, we will describe these data collection techniques. The benefits and disadvantages of each technique will also be discussed.

4.2.1 Interviews

The appliance of interviews, to acquire information, is a technique that is based on questions and is suitable to obtain qualitative material. This as, it allows the interviewer to receive exhaustive and considerable answers from the respondents, and may provide a great breadth of data (Denzin & Lincoln, 2000). There are different ways to conduct interviews. An interview may be performed personally or through phone calls. Common benefits of either way are that the interviewer could elucidate the questions for the respondent and also have the possibility to ask additional questions to achieve more comprehensive answers. A personal interview has the benefit that it is easier to collect more detailed answers and it is also a chance to get better insights in the respondents work. A large benefit with interviews by telephone is that more respondents could be interviewed in shorter time and also that there is no geographical limitations (Andersen, 1998).

Two important aspects to take into consideration when formulating interview questions, is the extent of structuring and standardization that should be used. It is possible to control the interviews by leave more or less flexibility for the respondent to answer within, which is referred to degree of structuring. There is very little flexibility in the way questions are asked or answered in a structured interview setting (Denzin & Lincoln, 2000). In contrast, in an unstructured interview the questions leave maximal space for the respondent to answer within. How the questions are elaborated and in what order they are asked, is the degree of standardization (Patel & Davidson, 1994). Interviews with low degree of standardization is performed when the questions are formulated during the interview, and asked in suitable order for a specific respondent. In totally standardized interviews, the interviewer asks all respondents the same series of pre-established questions in the same order. A

qualitative approach is distinguished by interviews that are performed unstructured and with low degree of standardization (Patel & Davidson, 1994).

In our opinion, interviews are of current interest to achieve our objectives in this dissertation. This as it could describe the current situation of how the usage and integration of external data into DWs looks like in companies that develop DWs.

4.2.2 Questionnaires

A questionnaire is another technique that collects information by using questions. This technique has some common parts with interviews, but there are also things that differ. The difference lies in that a questionnaire is usually sent to the respondents by, for example, post to be answered in written form, whereas the questions in an interview are asked directly to respondents (Patel & Davidson, 1994). Questionnaires are usually used in a quantitative approach and are formed like a standardized interview where the questions are asked in exactly the same order to all respondents (Andersen, 1998). The benefit is that this is a cheap way to conduct a research on, especially in situations where a large number of respondents are included and the questionnaire could be sent to respondents for a relative low cost. Disadvantages are that questions cannot be explained and that it cannot be assured that all questions will be answered. There is also a greater risk for a severe loss of answers, which is another drawback (Dahmström, 1996). To conclude, questionnaires offer a possibility to perform interviews with many respondents in a short time, which gives a broad basis on which the problem in this dissertation could be answered.

A questionnaire could contribute with a great selection of responses to answer our research problem. The risk is low answer frequency and very brief answers.

4.3 Research approach chosen

In order to answer the aim and objectives stated in chapter 3, we have chosen to base our research approach on interviews. The choice was made in consideration to the alternative of conducting a material collection, based on questionnaires. The main reason to select interviews is the possibility to ask additional questions during interviews and also the ability to sort out unclear points concerning the questions and the answers. In addition, based on the possibility given above, we also believe that interviews will result in more detailed answers than questionnaires. A questionnaire does not give any chance to ask additional questions and leaves little room for alternatives of answers, which implies the risk of not so detailed and not exhaustive answers enough to perform a qualitative analyze. Furthermore, the need for detailed and exhaustive answers gets even more urgent when considering the amount of literature covering the topic. There is not much written concerning external data in DWs and the few references existing only gives indications of usage and benefits on a high level. Therefore, the interviews will almost solely generate all the material that this report will be based upon.

The main aim of the interviews is to generate insights into how external data is used in practice. We consider it to be essential to get an empirical anchorage of the material, to be able to answer the research problem in a satisfactory and scientifically

interesting manner. Therefore, we have chosen to conduct interview with experienced DW developers, which may contribute with exhaustive and relevant answers.

Since, we decided to use a qualitative approach (section 4.1) for our research; this implies that the interviews should be conducted with low degree of structuring. By this we want to leave a large room for the respondents' subjectivity and a chance of collecting a greater breadth of data. Even though, according to Patel & Davidson (1994), the qualitative analyze should use low degree of standardization we will conduct the interviews in a relative standardized way. This, as a certain degree of standardization is required, to be able to compare different answers from several respondents. However, we still want to ask the questions in the order that are most suitable and try to perform the interviews more like a dialog, which leave us with relative low standardization. There is a possibility that the mutual order of the questions will vary during the interviews.

The interviews will mainly be conducted by phone calls. The working conditions of this dissertation, in form of timely and economically resources influence this choice. This means no geographical boundaries, which in turn decreases the time frames for the interviews. The cost of performing interviews by telephone is considerably low in comparison to personal interviews, which involves a lot of traveling since the respondents are geographical dispersed. This also implies that personal interviews would take much more time.

Finally, to give an outline for the interview study, we have chosen to give the activities that will be performed. These activities are:

- Select the respondents
- Draft questions for the interviews
- Conduct the interviews
- Evaluate the interviews
- Experiences from the interviews

These activities will be more thoroughly described in chapter 5. However, an observant reader may still grasp the order of sequence of the activities and use them as a guideline for further reading.

5 The interviews

This chapter gives an account for how the interviews, as we presented as our data collection technique in previous chapter, have been implemented, in order to answer the problem of this dissertation. Firstly, we present the selection of the 12 respondents that have participated in the interviews. Then, we describe how questions for the interviews were drafted. After that, we will explain how the interviews have been conducted. Lastly, the interviews are evaluated and finally our own experiences are pointed out.

5.1 Selecting the respondents

Our goal was to interview between 10-15 respondents, which was considered as reasonable to achieve a relative broad material. The fundamental characteristic of the respondent was that they had experience from development of data warehouse(s). Our objective was to interview DW developers that had experience from earlier phases in the development process, with background in business activity development and project management, rather than on the technical level like programming etc. We felt that we could not be too choosy in the selection of respondents. Therefore, has the respondents that have participated in the study been contacted relative randomly.

The easiest way to get in contact with designers of data warehouse was to contact companies in the IT industry, which possible have experience of development of DWs. The first attempt to find suitable IT companies was made on a career day for IT students, where many companies were represented. This showed to be a convenient method to get in contact with respondents. The people that represented the companies were helping with reference to appropriate contact persons in their organization. These persons were contacted by phone or email and resulted in that 6 respondents agreed to take part in the research. This was complemented with search for suitable respondents on the Internet, which resulted in that 8 additional persons were contacted by email (see Appendix 1). This email contained information about the aim of our research and how the respondents' contribution should be elaborated. In addition, the e-mail also gave the approximated time frames for the interviews (20-40 minutes), and that it was DW developers with experience from earlier stages in DW development that were most interesting. Patel & Davidson (1994) pinpoints the importance of these aspects for being able to clarify the aim of the research and how the respondent going to contribute. Here we were lucky to have a good response from people that agreed to participate. Out of the 8 persons that were contacted 6 respondents agreed on participating. We do not know the reason why the others could not participate, since they did not reply the inquiry. The arrangement of point in time for the interviews was scheduled either by telephone or email. It should be mentioned that it required quite a lot of e-mailing before suitable respondents were scheduled for interview.

Some of the respondents that agreed to participate expressed, before the interview, that they do not work anything with external data. Still, we considered it interesting that these persons participated in the research. This partly since, one primary research objective was to investigate to what extent external data is integrated into DWs. In addition, this could give reasons for not using external data in DWs and enable

different views on integration of external data among developers, which in turn may be compared and analyzed.

5.2 Draft questions for the interviews

As our chosen technique to collect information was interviews we had to draft questions (see Appendix 2). The questions were aimed to collect a qualitative material, and were drafted unstructured and with relative low degree of standardization. Firstly, thinkable questions were listed spontaneously. After that, discussions were held with the supervisor, and some questions were cut out and some re-drafted. Furthermore, the mutual order of the questions was changed so similar questions were worded in a more logical sequence. The questions were sequenced in consideration to a technique, which is referred to as *"funnel-technique"* (Patel & Davidson, 1994). The *"funnel-technique"* implies that an interview should start with open questions to gradually go over to more specific ones. The aim was to activate the respondents, in a sense that they felt free to verbalize free in the beginning. Moreover, this is a way for the interviewer to show interest in the individual respondent (Patel & Davidson, 1994).

The questions were divided into 3 parts: general/initial questions, main questions & possible questions, and lastly concluding questions. The initial questions were aimed to collect background facts about the company, the respondent and his/her experience of data warehousing, and a chance to show interest for the individual respondent. It was considered as important, for being able to put the respondent answers in relation to their experiences. Additionally, we found it important that the respondent defined data warehouse, and internal as well as external data, before we approached the main questions. The main questions were drafted to answer the research problem and to fulfill the aims of the dissertation. Possible questions were elaborated, to be asked in consideration to earlier answers, whether external data has been integrated in the development of DW. Concluding questions were aimed to get the respondents view of integration of external data in DW in the future, and left room for additional important remarks that has not come to light earlier in the interview. Main part of the questions was open and the intention was that respondents' should discuss the questions from experiences in the problem area and with freedom to formulate the answers.

5.3 Conducting the interviews

As stated previously, the interviews were conducted by telephone, with only one exception, where the interview was conducted in personal. This interview took longer time than the others, since the respondent demonstrated their work with a presentation. Several respondents have also offered the possibility of personal interviews, but due to inadequate resources we decided to conduct them by telephone. When an interview was scheduled: the interview questions were sent in advance, in order to give the respondent a chance to go through and reflect upon the questions. This had a positive effect and in most cases the respondents seemed to be prepared, which made the interviews easier to perform.

There are basically two ways to record answers: by taking notes or to use a tape recorder. The advantage to record the interviews on tape is that the interviewer could concentrate on the questions, and the answers that are given by the respondent.

Another benefit is that the answers are recorded exactly, which means that we can go through the recording many times. To record an interview demands the respondents' agreement. The disadvantages is that the process to print downs the interviews is time-consuming and that presence of a tape recorder may have an effect on the answers that are collected (Patel & Davidson, 1994). Since, we did not have any access to a tape recorder, we have chosen to take notes in this work. Even though, this may have a negative effect on how well we were catching the answers we believe that any unclear points have been sorted out afterwards. This as we in all cases had the possibility to contact the respondent again for additional questions or to complement the notes. Directly after an interview was conducted, the collected material were compiled and sent to the respondent to be approved before the inclusion in the report. By this, the respondent also could make possible complements.

Every interview begun to make sure that the respondent had received information about the aim of the research and the questions in advance. All the respondents were given a chance to decide whether the interview should be handled confidentially. Given that one of the first respondents wanted to stay anonymous we decided to treat all the interviews the same. All interviews were conducted according to the same sequence of pre-established questions. In some situations it was a necessity to further elaborate on the answers given, with follow-up questions and explanations.

5.4 Evaluation of the interviews

In some interviews, discussions have arisen on what is meant with internal and external data. This had an effect on the continuous interview and what follow-up questions that has been asked. To guarantee that the respondents had adequate knowledge to discuss the questions, was the responsibility placed on the respondents by explaining the approximate questions before the interviews were scheduled. This meant that they had a chance to decide whether they consider themselves to be able to contribute to answer the questions. In three cases, has this caused that interviews has been forwarded to other persons in the same company.

When putting together the material gathered in the interviews, it did not appear to be so clearly as it seemed in the present interview. Nevertheless, information has come to light in the interviews that could be used to answer the research problem and the objectives of this work. Firstly, information about whether external data is integrated in data warehouses has been gathered. In those cases where external data is integrated: information about what external data that is used and what types of external data sources have also come to light. In some cases, there has been a problem to get specific answers about what external data that is integrated into DWs. The answers have been a little bit lucid, e.g. statistics, which has brought up questions like what kind of statistical data has been integrated. Secondly, opportunities and pitfalls with integration of external data in DWs, both from a current and a futuristic viewpoint, have been identified to a certain extent.

Lastly, in general the information is considered to be representative for the specified research problem. Even though, it cannot be claimed that the research is exhaustive for the whole DW industry, the research consists of 12 interviews with DW developers that working for leading companies in the Swedish IT industry. All of the respondents are currently working in DW development projects, and some of them have many years of experience in the area.

5.5 Our experiences of the interviews

We do think that the information mail, in addition to that the questions, were sent to the respondents a couple of days in advance, helped to keep the interviews more focused to the problem. It also meant that the respondents were prepared and more distinctly answers could be collected. We got in contact with respondents in early phase of the work, which was lucky since it in many cases took some time before a definite occasion for the interview was scheduled. Nevertheless, as we were in good time, the interviews could begin as planned and were carried out in relatively short period of time.

In some cases it felt like the respondent left a comprehensive answer, but afterwards it has been discovered that the answer did not reply the question. This meant that the respondent begun to discuss other matters that is to be found in the same area, but did not answer the question directly. However, in most cases has the questions been answered in an appropriate way. There are also other sources of error that could be identified in the nature of the task of interviewing: for instance the sequence of wording the questions. Another source of error is the interviewer, whose characteristics or questioning techniques can impede proper communication of the questions (Denzin & Lincoln, 2000). Since, we do not have much experience of interviewing this may have affected how the interviews has been conducted. The experience was that it became better the more interviews that were conducted. We have tried to play a neutral role, and never interjecting the own opinion of a respondents answer.

6 Information presentation

In this chapter, we summarize the information resulting from the interviews. Firstly, the respondents that have participated in the research will be presented. Thereafter, a compilation of information that is relevant for answering the research problem is given. More detailed answers from the interviews are described in appendix 3.

6.1 The respondents

In this section, the twelve respondents that have participated in the investigation are presented. This information is based on question 1-5 of the interview questions (see appendix 2). Furthermore, the Internet has been used to collect information about the companies where the respondents are employed. This information has been collected from the companies' own web sites. The respondents represented eleven different companies, and a range of different experience in the data warehouse area. The interviews are treated confidentiality and therefore we will refer to the companies and the respondents with numbers, according to the order in which the interviews was conducted.

Respondent 1

Respondent 1 is employed on a global management and IT consulting firm. The concern is established in management consulting, system transformation, information systems management and outsourcing. Their goal is to assist companies to continue to grow, and to create benefit by new technology. Respondent 1 has been working 3 years for this company, and has been involved in several DW projects.

Respondent 2

Respondent 2 is working in an IT-company within one of the largest Swedish local governments. It is a knowledge organization with focus on IT and business process performance for the local government. Their primary customers are public administrations and government services in the region. They are 300 employees, of which 100 are working with development. The respondent has been working in this company for 15 years. Respondent 2 started to work with development of data warehouses 2,5 years ago.

Respondent 3

Respondent 3 runs his own company and through a broad network; consults and competent personnel are engaged in different projects. The business concept in this company is to use suitable tools to acquire and present the most valuable in companies databases with respect to customers and behaviors, seen from the perspective to strengthen customer relations. The company is active in IT-oriented services, mainly with CRM-applications with the aim to work as a support for customer relations and marketing analyzes. Respondent 3 started this company one year ago, and have experience of data warehouses since 10 years back in time.

Respondent 4

Respondent 4 is employed in a large concern that supplies IT services in Europe. The company provides consulting, systems development and integration, operation and support, product development services for customers, and software services. Their aim is to be a strategic IT partner to its customers. Respondent 4 has been working as

a IT architect for 4 years in this company, and has been involved in the earlier stages in 2 data warehouse development projects. The respondent works in the business area: public sector.

Respondent 5
Interview 5 was conducted on a company that today has 25 employees, equally distributed among systems analysis and design, and education. Their business concept is based on user-driven business analysis and process and object oriented system development. The company is working with development of information systems, databases and also data warehouses. Respondent 5 has been employed in this company for 3 years. The respondent has been working with data warehouses and analyzing tools for a couple of years. This has involved different development projects among hospital and country council, SSAB industries, and Preem petroleum.

Respondent 6
Respondent 6 is working on a global company that develops a enterprise application solution. They develop, implement and take responsibility for the on-going support of their own system. The respondent has been employed in this company for 2 years. The experience of data warehouses is 1,5 year and the respondent has been involved in one larger development project. Right now the respondent manages and administrates 10 data warehouses.

Respondent 7
Interview 7 was conducted on a business intelligence (BI) technology services/consulting company focused on helping enterprises with their information needs. The company is specialized in delivering business intelligence solutions: better reports, better analysis and data warehousing. Respondent 7 has been employed for 5 years in this company, and works as a project leader, sell support and architect in BI and DWs. The respondent has been in contact with data warehouses since he was employed in this company 1997.

Respondent 8
Respondent 8 is employed in a company that supplies IT-related services. The company business vision is to be a leading European supplier of IT-related services, with the Nordic region as their home market. The number of employees is amount of 7000, of which 4000 in Sweden. Respondent 8 has been working for this company for 1,5 years, and have 2,5 years of experience as a project leader in DW projects. The respondent has been involved in 6 projects, of which 2 larger ones.

Respondent 9
Respondent 9 is working in a IT-company with 60 employees with focus in strategies, processes, and IT supporting customer relationships. Their heritage is in Business Intelligence, and today the company focus much on the area of Customer Relationship Management – CRM. Respondent 9 has been working in this company for 12 years, and currently the respondent as a consult within analyze and development of CRM, where DW is an integral part. The experience is 12 years with development of decision-support systems and data warehousing. This has involved around 30 projects.

Respondent 10
Respondent 10 is employed in a company that develop and supplies component-based business applications for medium and large enterprises. The Applications, which is

based on web and portal technology, offers 60+ enterprise application components used in manufacturing, supply chain management, customer relationship management, financials, engineering, maintenance and human resource. The company has more than 3,200 employees, with sales in 43 countries. Respondent 10 has been working in this company for 3 months. The respondent have experience from 1995 with development of databases and data warehouses, and have technical as well as management experience of data warehouses. This has involved in about 10 different projects.

Respondent 11
This respondent is working for the same company as respondent 4, but at a different department in another city. Respondent 11 have been working since 1973 (29 years) in the IT-business, and the company that the respondent used to work for got purchased of the company where I am now employed. The respondent have worked in development of economics and personal systems, and in connection with this developed decision-support systems. For 10 years has respondent 8 been working with data warehouses and this has involved development of small and larger systems. Have participated in 7-8 projects, and this projects has only been in the public sector.

Respondent 12
Respondent 12 is working on a consult firm in the IT industry and in the present situation they are 40 employees. The company business concept makes the direction clear: *"To be a technical knowledge company that deliver IT-consulting services for companies and organisations in the Göteborg region. This comes about by competent consults that puts customer loyalty and quality in the forefront"*. Respondent 12 has been working 4 years in this company. The respondent has 2 years of experience of data warehousing, which has involved 2 projects.

6.2 Study findings from interviews

In this section we report what the respondents answered to the main interview questions 6-15. All the respondents, regardless of whether they have experience of working with external data, responded to the questions concerning their perception and experience of integration of external data into DWs. To present the answers more clear, they are arranged from whether external data is integrated in data warehouses or not. In section 6.2.1, we present answers from the respondents who have experience of working with data from external sources. Further in section 6.2.2, responds from respondents that do not have any experience of external data are presented.

6.2.1 Respondents working with external data

One of the objectives of the study was to ascertain if any and to what extent external data is integrated in data warehouses. Respondent 1, 3, 5, 8, 9, 10 and 12 have been working with, or are currently working with integration of external data in data warehouses.

Respondent 5 states that how external data is defined in relation to a data warehouse is conclusive to whether external data is used or not. They are working with data that originates from external sources, yet this data is first stored in internal systems before inclusion in the DW. In this dissertation, we consider external data as data that

originates from outside an organization. Therefore, we consider that respondent 5 have experience of working with external data. Respondent 12 feels that the answer got to be yes and no, since they acquire external data in form of ordered files from an external system in another company, yet inside the same concern. The respondent states that this data is external in the sense that they do not have any control over it. Considering how we have defined external data in this dissertation, we interpret this as respondent have been using external data.

Some of the respondents that have been using external data spoke in terms that it was not that common. Respondent 9 states that it is not so commonly used, and respondent 8 adds that one has to be aware of the fact that 70%-90% of the data stored in a data warehouses, is acquired from internal systems.

6.2.1.1 External data integrated into DWs

For those respondents who have used and integrated external data into DWs, a question were asked concerning what external data that have been integrated in projects where the respondents have been involved. Here we present answers from each of the respondents 1, 3, 5, 8, 9, 10 and 12.

Respondent 1 experience that the following external data is most common:

- *Enterprise address information*: address register to all company offices, for instance in a certain branch.
- *General enterprise information*: like for example economical information about a company, the number of employees, what types of customer they have.
- *Branch codes*: Codes that describes in what branch a companies is competing.
- *Individuals information*: The respondent knows that this has been used in other projects, but do not have any own experience of integrating data on individual level. Credit rating is one example that could be used.

The address information have more exactly been:

- Address (street/box, postal number, post district)
- Telephone number
- Fax number
- Name of highest managerial position
- Title of highest managerial position

In terms of the what external data that has been integrated, respondent 3 declared the following:

- *Public documents & company documents:* For example, economical data like annual account information.
- *Branch codes:* Different levels that is more or less foreseeable. It is a SNI-code that is used.
- *Risk class:* a credit evaluation in the form of credit rating and risk forecast and a recommended credit limit. Means that companies are classified.

Respondent 5 mentions the following data that originates from external systems:

- Exchange rates
- Oil prices
- Customer information “SNI-codes”: Index and codes that describes branch belonging. There is a great demand for this kind of branch index in DW.

- Files and information from other companies, for example from electricity companies to get a correct follow-up of electricity.
- Geographical data: for example zip codes
- Follow-up definitions: how accounts should be grouped for follow-up. It describes how you want to view the information.
- Algorithms: to estimate calendar functionality. Different algorithms to calculate public holidays.

Respondent 8 states that there are a lot of external data that is used:

- Economy data
- Municipality data: for example community analyses
- Branch organisation data: for example, a common organisation for grocery stores. Here is address data most common to collect.
- County council data: this data is free and could for example be, population or age groups
- Target groups
- Population statistics
- Education
- Age groups
- Age groups and pattern of movement: for example, how many and in what ages people are moving in a certain area
- Customer groups: for example, how many individuals exist in a certain customer group.

Respondent 9 declares that there are companies that offer services where you can send your customer database, and they will add up and update with additional information of addresses. This also washes your own data so you assure the quality of the data, and verify that you have the right information. It is possible to subscribe services so that you have access to this information on-line. Respondent 9 brings up the following external data:

- Demographical data[1]
- Credit report information.
- Customer register, mainly address register.
- SNI-codes and line of business codes

Respondent 10 defines external data as data, which is not collected from the ERP-system that they develop. The respondent mentions external data from cash registers in shops, and also other statistics and reports. Another example is freight companies, where data about delivery dates could be integrated to analyze delivery precision. This implies that these companies open up their systems for others to collect information.

Respondent 12 have experience where information has been acquired about apartments and locals from different real estate systems. Other information has been marketing investigations and statistical information. In one of the projects this has been age groups and age categories that are in need of rental apartments.

[1] Demographics describe what an individual is like and here you start from: age, gender, income, education, accommodation, hobbies, marital status, etc.

To summarize what type of external sources and domains that organizations are most interested in, three main areas of external information were identified among most of the respondents:

- *Customer information*: e.g. to relate customers geographical position to sales places, target groups (how do they think?), address register.
- *Company information*: e.g. economical information about companies, address registers.
- *Market and competitive information*: e.g. geographical data, statistical data.

In addition, respondent 1 and 3 consider information from different branches to be important. This area means access to branch codes, and to market information that are specific for a certain branch. Respondent 3 and 9 also states credit report information as an important area to have access to. Information about suppliers and deliveries is another area that respondent 10 mention as important to have access to.

6.2.1.2 External data sources

This section takes up answers that were collected when asked about what external sources that are most common to acquire data from. In terms of what external data sources that are most common was the following sources identified among respondents 1, 3, 5, 8, 9, 10 and 12.

- *UC* – Upplysningscentralen. An organization that provides information about Swedish companies. It is a business and credit information agency owned by the Swedish banks. At UC, there is a possibility to establish a subscription, which means that data could be collected when needed. It is also possible to order information about specific areas and branches. Respondent 3 are working exclusively with UC when acquiring external data. Respondent 9 also states UC as a useful external source that supplies a lot of information.
- *SCB* – Statistics Sweden: Offer different kinds of statistics. Respondent 1, 8, 9, 12 consider this as a common external source.
- *PAR*: This source in contains address information about all addresses in Sweden. It also provides data about companies, the branch they are active in, and so forth. Respondent 1 states this as a useful external source.
- *Spar*: This is a public person address register that contains data about private persons. Also stated by respondent 1 as a useful external source.
- *Dun and Bradstree (D&B)*: Provides more global economical information about companies. For example credit refining on companies. Respondent 3 mentions this one as a source to acquire global information. D&B offers a large database, with information on 70 million companies worldwide – for credit, marketing and purchasing decisions. The aim is to help businesses to reduce credit risk, find profitable customers and manage vendors efficiently.
- *Branch organizations:* that offer statistics specific for a certain branch. Respondent 8 and 9 mention this as a possible source. However, respondent 8 also has experience that it has been hard to get hold of this kind of data.
- *County council and community:* Respondent 8 states that external information could be acquired from public organizations like county council and communities.
- *Internet*: Respondent 9 means that the Internet may be used to look after competitors, by watching what they offer and to what prices. Respondent 3 also

believe there is a lot of useful external data on the Internet that could be useful if it could be found in all that information that is available.

- *Marketing investigators*: Respondent 12 mention different marketing investigations and statistical information. Respondent 9 also states that consuming statistics could be acquired from different investigations etc.
- *Cash registers:* Respondent 10 mention this as a possible external source.
- *Other enterprise systems:* Data from other businesses in a concern, or data from suppliers systems is mentioned by respondent 10.

6.2.1.3 Perceived opportunities/pitfalls of using external data

Most of the respondents stated that the integration of external data always is related to the business objective of the DW, and that users requirements is decisive to what external data to integrate.

Respondent 1 consider the inclusion of external data as valuable to become more effective, and to be able to ask sharper questions, and to achieve a better hit-rate. One useful area is in customer campaigns, that demand that external data is bought from external sources, like for instance addresses. The premier problems are difficulties in definitions, when the supplier and the buyer of external data sometimes define data different.

Respondent 3 states that the usage of external data is important, and creates a greater value, as more data and information about the market and the customers is acquired. The only disadvantage is that it cost money, and is expensive to buy. Concerning the quality, it is important to use reliable sources.

Respondent 5 means that integration of external data is totally dependent on the aim of the data warehouse. It is important if needed to satisfy user requirements. The disadvantage is the less control of external data, and also that external data may use other keys that have to be washed together to fit in with internal keys, which could imply some extra work.

Respondent 8 feel that external data provides more possibilities for analyzes, which in turn strongly arguments the inclusion there of. Data quality could in some cases be a disadvantage, since the risk that external data could be misinterpreted in comparison to the internal data.

Respondent 9 states that external data may be used to complement internal data and to generate new reports. Another field of application is to wash internal data through an external source. This could save costs by for example keep better track of customers and their current addresses. There are no disadvantages in using external data, but it may exist difficulties with matching external data to internal data.

Respondent 10 agrees with respondent 5, that external data is important when there exist needs to integrate external data, to be able to fulfill the customers requirements. In terms of disadvantages, there are difficulties to match the external data with the internal. A data type could be defined in different ways, which requires additional data wash before integration with internal data.

Respondent 12 means that external data may provide tracking of relationships and analyses, and this in turn could imply more effective steering of the business. A solid validating process is required to guarantee data quality, though there lays a risk in that we do not have any control over it.

6.2.1.4 Integration of external data in the future

Generally speaking, the respondents considered that the usage/integration of external data in development of DW would increase in the future. This is again related to the aim of the DW, and that it today no longer exist any technical limitations, put in performance words.

Respondent 1 have the opinion that the increase of external data is driven by the fact that today is the integration of internal sources performed in a good way. When the internal picture is clear, there is a chance to become more powerful by complement with external data.

Respondent 3 feels that the possibilities of large benefits that could be drawn will increase the usage of external data. Today it is technical possible to manage large amounts of data, and at the same time as the competition gets tougher, it become more and more important to know as much as possible about the market where you are active.

Respondent 5 believes that there are great opportunities with external data, and that the usage will increase in projects where there exist demands. As long it is possible to assure the quality of external data, it would be important to consider in developing DWs, and would also influence the development.

Respondent 8 thinks that the usage of external data will increase a lot in the future. Wholesaler sites of data will pop up, with companies that are specialized in collecting data and perform the heavy work.

Respondent 9 means that in general the use of external data will increase in the future, as a complement to internal data. Yet, the situation is often first of all, to structure and compile the internal data, and that many DW projects are not ready to integrate external data.

Respondent 10 consider that the integration of external data will increase. Primary because of the fact that the Internet is more used for distributing data nowadays, but also because the increased openness in business-to-business environments. This, in turn simplifies the integration of data from other businesses.

Respondent 12 states that internal data naturally will have greater priority, and that it is not until everything internal is prepared that new possibilities may be drawn by the inclusion of external data. But, the respondent has the opinion that the needs of external data will increase as soon as the internal data is structured in an appropriate way. Finally, respondent 12 also claims that the integration of external data is primarily a matter for the private sector.

6.2.2 Respondents not working with external data

The respondents 2, 4, 6, 7, 11 have no experience of integrating external data into DWs. They have only been working with internal data acquired from internal sources.

Respondent 4 and 7 do not have any own experience of integration of external data, but knows that it has been planned and used in other DW projects where the respondents have not participated. Respondent 6 states that they only collect data from internal source systems, but there exist external data in the internal systems. In this case, we do not consider that the respondent have used any external data, as the respondent do not have any own experience where external data has been required, and only know that it exists in the internal systems. Respondent 7 is treated in the same way, since the respondent does not have any experience where external data has been integrated straight from external sources into the DW.

6.2.2.1 External data sources

For those respondents who do not have any experience of working with external data, questions were still asked about what type of external sources and areas that would be interesting for organizations. We also asked if the respondents were aware of any external sources that supply data. Answers to these questions are presented in this section.

Since, respondents 2, 4, 11 do not have any experience of external data, they meant that they do not have any knowledge of what type of external sources and areas that would be interesting for organizations. Respondent 4 do not have any experience of external sources that provides data. Respondent 2, 11 know that SCB provides statistics and demographically data.

Respondent 6 states that things that would be interesting to include in a data warehouse is: branch codes, soft KPIs (e.g customer satisfaction), and customer addresses. The external sources that the respondent has knowledge about are SCB, UC and Smelink.

Respondent 7 means that the external domains that are most interesting are customer information (customer analyzes and surveys), market research, and information about competitors. The respondent knows that this kind of information have been integrated in some data warehouses, but not in the projects where the respondent has been involved. However, respondent 7 knows that there are different research-groups that supply with this kind of data, for example the Gartner group.

6.2.2.2 Perceived reasons/opportunities/pitfalls of not using external data

In general, the respondents states that the reason why they have not worked anything with external data, is based on the fact that there has not been any requirements or needs to acquire data from outside the own organization.

Respondent 2 states that the data models and cubes they have created have consisted of data from the own organization. With external data it could be harder to assure that data is defined in the right way, and that the same definitions is used. This complicates the integration of external data.

Respondent 4 states that there have not existed needs to integrate external data. The focus has been on data that is produced by internal processes. One advantage by not using external data is concerning with data quality. It is harder to assure the quality of external data, though you do not know how the data has been validated, which implies that it has to be washed to a greater extent. A disadvantage, by not integrating external data, is that analyzes could not be rendered how external sources influence the internal processes. The respondent believes that it is more common to use external data in development of DWs in the private sector.

Respondent 6 is of the opinion that you always should start with the own internal data and sales. It is not until this is fully developed, that there is a possibility to start look for external sources. The respondent means that immaturity in development of DWs is one of reasons why they do not work with external data. The disadvantage is that there is no possibility to analyze internal numbers against external. To work more with external data would imply improved analyzes.

Respondent 7 indicates that there are enough problems to structure and integrate the own information from different internal systems. It is difficult to know and to assure the quality of external data. The disadvantage by not integrating any external data is that the chance of a wider perspective is lost.

Respondent 11 means that there have not been any needs for external data. The respondent believes that it is more common with external data in DWs that are implemented in the private sector. The respondent has only participated in projects in the public sector. One possible benefit by not using external data may be less data wash before integration in the DW.

6.2.2.3 Integration of external data in the future

This section provides answers about the integration of external data in the future, given by the respondents that do not have any experience of external data. In general, the respondents believes that usage/integration of external will increase in the future.

Respondent 2 believes that the inclusion of external data lies in the future. Respondent 4 agrees and means that the usage of external data will increase in the future. This as organizations gets better structure in their data warehouses, which makes it easier to integrate external data. Respondent 4 consider external data to have a market value and that there exist needs for external data. Especially in the private sector, but not that much in the public sector.

Respondent 6 and 7 feels that development of DWs is immature and that the interest for external data will increase in the future. Respondent 7 also means that the market is probably not very consciousness about what possibilities there are with external data. In relation to that new needs emerge, the needs for external data will increase, but it requires an easy way to integrate external data into an existing DW.

Respondent 11 also consider that the use of external data will increase, above all in the private sector. This as companies more and more wants to compare their own business to others in the same branch.

7 Analysis

In this chapter, the material collected in the empirical study is analyzed. In the analysis, the respondents' different and similar answers will be compared and the main emphasis will be on the material collected during the interviews. In addition, the material will also be compared to information that came to light in literature study. In cases where there exists a connection to information that is documented in chapter 2 (Data warehouse), this will be taken into consideration. The analysis is divided into four parts, which reflects the objectives of the study. Firstly, the integration of external data in DWs will be analyzed together with a discussion about the external data that has come to light in the interviews. Thereafter, external data sources that are used among DW developers will be analyzed. Then, we extend the analysis with opportunities and pitfalls with the integration of external data in DWs. Finally, we discuss issues when integrating external data in DWs, from a futuristic perspective.

7.1 The integration of external data into DWs

When the empirical study was compiled it showed that 58%, 7 out of 12, respondents had experience of working with integration of external data into DWs. In some cases it has not been obvious, whether the respondent should be considered as they have worked with external data. This uncertainty lays in the definition of what data that should be considered as external. We assume that external data is data that is sourced and originates from outside the own corporate environment. However, when evaluating the answers, it was shown that two respondents were to consider as questionable, from an external data perspective. One of them is respondent 5, who states that he is working with data that originates from external sources, but this data is pre-stored in the internal operative systems. The respondent means that this data that is informational extern, but systematically intern. In this case, the respondent remarks that there have existed requirements to acquire data from external systems, and therefore we consider the respondent to have experience of working with external data. Respondent 6 also remarks that this is a question of definition. This respondent knows that there exist external data in internal systems, but that the development of DWs has only acquired data from internal source systems. In this case, we have assumed that the respondent does not have any experience of integrating external data. This could seem contradictory to consider these respondents as different. Still, we mean that respondent 5 could be regarded as having experience of external data, and not respondent 6.

The responses from DW developers in this study indicate that integration of external data in DWs is relatively common in Sweden. Even though, it cannot be claimed that the research is exhaustive for the whole DW industry, it consists of 12 interviews with DW developers, working for leading companies in the Swedish IT industry. All of the respondents are currently working in DW development projects, and some of them have many years of experience in the area. Therefore, we consider the outcome both scientifically valid and interesting. DWs are given a lot of attention, both by academics and practitioners, and the amount of DW literature describing different aspects of data warehousing is increasing. Still, we have not found any studies to compare our results with, neither American nor Swedish.

Naturally, the specific project and user requirements are always of vital importance and decisive whether external data is used or not. This is a unified view, which all the respondents' remark. A tendency is that most of the respondents means that there are great opportunities with external data. However, it is only a little more than half of the respondents that are using or have any experience of integration external data in DWs. Some responses points to the fact that this is a question of maturity; that external data is not integrated to a greater extent. In some cases there are enough problems to structure the internal data, and it is not until this is fully developed that integration of external data comes into question. Even though defined business demands and user requirements always determine what data to integrate in a DW, it is at the same time essential that the users know what possibilities that exists, for being able to fully exploit the value of their DW. From our point of view it is the developer that should inform the customer (users) about the possibilities of integrating external data, and what additional analyses and reports that possible could be generated from this. It is difficult for the users to specify requirements, and needs do not arise before you are informed. It should not be a question of maturity of the customer.

When a data warehouse is developed it is natural that the integration of data from several sources starts with internal systems. This is a unified view among the respondents', which also is supported in the literature. Some of the respondents that had experience from integrating external data stated that it is not that common to acquire data from external sources, and the majority of data in a DW are always acquired from internal systems. This is consistent with DW literature (Inmon, 1996., Singh, 1998) which advocates that mainly internal sources are used in DWs. Since, the aim of integrating external data usually is to complement the internal data, and provide possibilities to compare internal information to external, it is natural that internal data forms the base of a DW.

The external domains identified as most desirable and interesting for organizations are: *Customers, competitors, business partners, and business market*. This external information is directly linked to explicit and important aspects of organizational performance, e.g. customer satisfaction, competitive positioning, etc. In other words, there are important points where external data can add value. Among the respondents with no experience of working with external data, some of them agreed that customer information, and market and competitive information are important areas where external data might provide with valuable insights. Nevertheless, the majority meant that, as they have no experience of external data they did not know what domains that would be most interesting for organizations. This seems to indicate that there has not been any needs to look outside the own enterprise for information, and also that they do not know in what aspects that external data may provide additional value.

In terms of what external data that is integrated in DWs, the study has provided with various examples from the respondents. These were implicit answers from developers that have been working with external data. The interviews indicate that branch codes are a type of external data that is common to integrate. Many respondents have mention that they have worked with SNI (Svensk Näringsgrenindelning) codes, which is an index that describes branch belonging. This index divides the market into different levels, and could be used to sort out step by step in a search for e.g. companies. In several interviews address information, concerning customers as well as companies, come to light as desirable external data. This address information has more exactly been address (street/box, postal number, post district), telephone

number, and fax number etc. The address information is used to collect information about possible new customers, and also addresses to companies. This is commonly used in more targeted campaigns, and to help an enterprise to manage customer relationships by providing access to current customer information. The data warehouse works as company-wide database were all information related to the customers is stored. The aim is to avoid “badwill” and costly returns of payment notifications, member information, and DR (direct advertising) to consumers. This is achieved by keeping the address register updated by external address services.

Another type of external data that came to light in the interviews was integration of different kinds of statistics and marketing investigations in DWs. This type of information included data concerning: target groups, population statistics, age groups, pattern of movements, education, and demographical data. Our point of view is that much of this data is integrated to fully understand customers, current as well as future. This data provides a possibility to have a comprehensive understanding of customers, and much of this data is available in external sources. The DW could be an enabler of one-to-one marketing and customer relationship management. Something that also has been mentioned is the use of data concerning economical classification of companies. In this issue, annual accounts information, risk class, credit rating and risk forecast has been stated as data that was acquired. Credit ratings provide information about the business risk of a company or a sole trader business. Risk forecast includes information about the risk of future insolvency. In companies newly started is data about both the accounts and about the board of directors plays an important part. The liability of a company is important to integrate in a DW, as a support in making strategic decisions about for example what companies to start partnership and trading with. The risk class data covers both credit ratings and risk forecasts, and is rated from 1-5, where 1 implies the highest risk and the lowest risk of becoming insolvent within a two-year period (UC, 2002).

The remaining external data that has been mentioned in the interviews is information that changes frequently over time like exchange rates and oil prices. Also information and files from other companies, where electricity companies has been used to get a correct follow-up of this cost, or delivery dates from freight companies, are examples that has been mentioned.

Furthermore, another fact that has been noticed in the study is that is seems to be differences between integration of external data in the private sector in relation to the public sector. Several respondents mean that it is more common and that more requirements exist for external data in the private sector. Moreover, this has been a viewpoint of both a developer that have worked with external data as well as developers that have not. One of the respondents even indicates that the reason for not working with external data is that the respondent only has experience from projects in the public sector. However, this raises some questions, e.g. what makes the difference? The only answer, which has come to light in the interviews, is that there are more requirements and needs in an organization that act in the private sector. From our viewpoint, a more competitive market often challenges an organization in the private sector. Therefore, it becomes more important for executives in the private sector to look outside their enterprises for information.

7.2 External sources

The external domains that were stated as most important and interesting in previous section, is consistent with some of the knowledge domains identified as valuable by organizations (e.g. customer and competitive information) in studies conducted by Alavi & Leidner (1999). Since, some sort of external sources or external data providers must be accessed to supply these external domains, this section will discuss the external sources that came to light during the interviews.

One thing that recurs in several interviews was the useful area of external data for address verification. As mentioned before, capturing accurate information at customer acquisition reduces undeliverable and cuts the costs of promotional campaigns. The sources that were expressed in this context were PAR, SPAR and UC. PAR contains information about addresses, and other data about companies like for instance what branch they are active in. This could be used to assure that all addresses are coupled to correct postal number, and that all changes in localities are updated. SPAR is a state public person- and address register. SPAR address themselves to organizations that want to keep their customer and member register that consists of private persons updated. From this it may be concluded that these kinds of sources is used mainly in DWs that is developed to be used for campaigns and as support in CRM-systems[2]. It is hard to see how this data would provide additional insights to support strategic decisions. Instead, it is easier to distinguish benefits like: avoid costly returns, keep the contact with customers, and to get more out of marketing for the same amount of money.

UC AB is another external source that has been mentioned in the interviews. The Swedish banks own UC AB, which is Sweden's largest business and credit information agency. They offer reports to support credit and commercial decisions, credit monitoring, and qualified financial analysis. The reports are produced from a database, which contain information on all enterprises registered in Sweden and all individuals over 16 years of age living in Sweden. This external source supplies a lot of information that could be used to support decision-making, like for example to uncover opportunities for business partnerships. It is an easy way to get hold of data that would take lots of work to collect and compile. Information and reports could be acquired according to different levels that is appropriate to specific requirements and needs. Another interesting issue is the possibility to access UCs database online via the Internet. Subscribing an on-line membership, which means that the information could be acquired when needed, performs this.

Furthermore, several respondents mention SCB (statistics Sweden) as a common external source. SCB offers a lot of different official statistics in many fields e.g. the labour market, the economy, trade and industry, prices, population and welfare, housing and construction. SCB also take on commissions for specially processed data, data collection and consulting. These commissions range from undertaking special processing of existing statistics to providing complex information systems like data warehouses. To meet special needs Statistics Sweden is generally able to provide tailor-made services (SCB, 2002). This external source appears to supply a lot different information and statistics, which may be integrated in a DW to enhance the information from internal systems with external information. Often, changes in a

[2] Methodologies, software, and systems that help an enterprise manage customer relationships in an organized way by providing access to all customer touchpoint information through a data warehouse.

turbulent market occur from the outside, and the type of information that is possible to acquire from SCB might provide basic data, which support decision-making in an organization. This could for example be used in segmentation and target marketing. Two of respondents also remark that there exist other marketing investigators, which supplies statistical information and consuming investigations etc. Yet, the responses did not name any more organizations of this kind. Another external source that should be mentioned in this context is that some branch organizations offer statistics specific for a certain branch. This could be useful to acquire data that is more specific for the branch, where the organization is acting. However, some respondent have experience that it is sometimes difficult to get hold of this data.

Moreover, information from other enterprise systems likes for instance suppliers or business partners are other thinkable external sources. One of the respondent points out cash registers in stores as an external source, from where data is acquired for integration in DWs. An example of this could be to look at the sales at the cash register of a grocery store. Furthermore, public organizations are another source that supplies information, and the benefit is that this information often is free to collect. This could for example be county councils and communities. The benefit is that this information often is free to collect. In terms of more global organizations has Dun & Bradstreet been mentioned as an external source, that provides global economical information about companies. D & B offers a large database, with information on 70 million companies worldwide, and the aim is to help businesses to reduce credit risk, find profitable customers and manage vendors efficiently.

Lastly, some responses have pointed to the Internet as a possible external source. Respondent 3 believes there is a lot of useful external data on the Internet that is valuable if it could be found in all that information that is available. Respondent 9 agrees and means that the Internet may be used to look after competitors, by watching what they offer and to what prices. The web technology is already used extensively as the delivery mechanism for warehouse data, but no one has seriously considered using Web content as input to data warehouses. This, as many question the reliability of Web content (Hackathorn, 2000). Clickstream analysis and web analytics are sources where information can come up. Clickstream data, a trail of behaviour left behind as users navigate the web, may provide businesses with a rich vein of information for better understanding the needs of their most valuable constituents (Capps T, 2002). Capps (2002) states, when combined with other enterprise information, the corporate world is able to perform sophisticated analytical operations that deliver immediate benefits and serve as practice for estimating the future.

In the interviews, it could also be concluded that the reliability of the source is of great importance. This statement was paid attention to by several respondents that have been acquiring data from external sources. Many responses pointed to the fact that they only acquire data from well-known companies, whose only business objectives are to collect, compile, and sell data. These kinds of companies are considered as professional on what they do, and therefore considered as reliable, from a data quality perspective. One example is SCB, which four respondents emphasize as a useful external source. If the answers in this study are considered on a large scale: there are only a minority of different external sources that is relied on when it comes to integration of external data in DWs. The responses shows that developers trust in experts, companies that are specialised on what they are doing, mainly, compiling data and selling it to other organizations.

Among the respondents with no experience from working with external data, the answers have showed that the greater part has knowledge that there exist external sources where data may be acquired. Again SCB is mentioned as a provider for statistics and demographic data. UC and different kinds of research-groups are also named in this context. However, since these respondents have not worked with external data is their implicit knowledge on external sources limited.

7.3 External data – opportunities and pitfalls

The respondents that reported experience from working with external data expressed several opportunities with integration of external data in DWs. Firstly, it should be mentioned that they all agreed that integration of external data always is related to the business objective of implementing a DW, and that the users needs is decisive to what external data to integrate. Among these respondents, it could be interpreted that the main opportunities to integrate external data into a DW is that it provides additional possibilities for analyses, and that it creates a greater value with the DW, as more information about the customers and market are acquired. One of the respondents means that external data is valuable for becoming more effective, to be able to ask sharper questions, and to achieve a better hit-rate. These opinions are supported by similar responses from developers that have not been working with external data. One of these respondents' means that the disadvantage, by not using external data, is that there is no possibility to analyze internal numbers in contrast to external. Other responses were that the chance of a wider perspective is lost, by not integrating any external data, and that analyses could not be rendered on how external sources influence the internal business processes.

However, some of the respondents, which have not worked with external data, hold different views of the matter. These respondents did not mention any benefits or opportunities with external data, but seemed to look more at the quality issue with integration of external data into a DW. One opinion that was pointed out by several of these respondents was the fact that there have not existed any needs to acquire data from outside the own organization. This is related to the importance of customer requirement.

Another positive effect with integration of external sources that has been noticed in the empirical study is the usage of external sources to assure the correctness of the own data. This means that the internal data is "washed" through an external source and implies another field of application than the organizational benefits that has been stated earlier. Yet, this serves a different aim, and is assumed to primarily have its usage in DWs that is developed for various kinds of campaigns and CRM-systems. By capturing accurate information at customer acquisition reduces undeliverable and cuts the costs of promotional campaigns. Undeliverable can take two meanings. In one sense, it refers to the mail pieces in a promotional campaign that come back due to incorrect information. In the second, it relates to excellent customer service. You cannot achieve excellent customer service if an order is incorrect or delivery to the wrong location. From an enterprise strategic management perspective the purpose would be to increase the marketing effectiveness and the customer loyalty.

The opportunities with external data that came to light in the empirical study have similarities to viewpoints stated by some researchers in the area. Professor Peter Drucker, has admonished IT executives to look outside their enterprises for

information, and remarks that it is a big challenge to organize outside data because change occurs from the outside. He means that the majority of data warehousing efforts result in enterprise focusing inward, while the enterprise could more keenly alert to its externalities, and that the internal information from internal systems must be enhanced with external information. Drucker further states that it is the synergism of the combination that creates the greatest (Oglesby, 1999). Inmon (1999) agrees and means that there exists a need to compare and analyze external data along with internal data, when it comes to executive level.

When considering disadvantages and pitfalls of integrating external data into DWs, the main opinions have concerned the quality of external data, and difficulties in data definitions. Among the respondents that have not been working with external data is data quality pointed out as an important issue to consider. This as it is more difficult to assure how external data has been validated. This is also agreed upon amongst some of the respondents that have been working with external data. They all point to less control of external data, which in turn may require a solid validating process to guarantee data quality. This could imply some extra work. However, if concerning the data quality respondent 3 claims that it is a matter of using reliable sources. The respondent means that it should not be a problem when using well-known external sources, which are professionals on what they do. As respondents question the reliability of external data, the question of how many that analyzes the reliability to any depth also arise.

Furthermore, another problem is that external data, even if it can be collected and/or acquired, does not easily and naturally mix with the internal data. Inmon (1999) means that in many ways external data and internal data are like mixing oil and water, which he says is a shame because there is real value in being able to look at external data and internal data holistically. This is a difficulty that has been highlighted mainly amongst developers with experience of integrating external data into DWs. The difficulties are stated to include matching problems between internal and external data, and that the supplier and the buyer define data different. One problem with a holistic perspective of internal and external data is that the keys, the content of the keys, the meaning of the key, do not match. An example, given by Inmon (1999), of a key structure mismatch could for example be: external data has customers keyed by an account number, and suppose that the internal system of a corporation use social security number as a basis for identifying customers. The external data must be converted from the account number key structure to a social security key structure, in order to achieve compatibility between external and internal data. Two different data providers could also define data in different ways. For example, one external data vendor states that there exist a certain number of companies within a specific branch index in Sweden, whereas another data vendor states that the number is totally different. No one of the vendors has necessary wrong, but there may exist differences in definitions.

As mentioned before, developers without experience of integrating external data have talked about immaturity, problems to validate the quality of external data, and that there has not existed any requirements, as reasons for not integrating external data into DWs. Except for the quality issue, there are not many pitfalls that could be interpreted among these respondents. Therefore, it could be assumed that it is not difficulties and problems that is the main reason why developers do not integrate external data into DWs. It is more likely to be a question of maturity, and that

development firstly focus on internal processes and make sure that all this data is in order and integrated into the DW, before external data comes into question.

One of the respondents' express that the only disadvantage with external sources, which the respondent has been working with, is that it cost money and is expensive to buy. Oglesby (1999) means that these kinds of external vendors sell their data to a reasonable price if the data warehouse is large enough to overcome minimum-order size. However, if you are small to mid-size company, establishing relationships with these vendors can be time-consuming and not very cost effective. These vendors are eager to sell their data, so the only real hurdle is cutting the check (Oglesby, 1999). From this point of view, the main pitfall of using external data is: a question of costs.

From the answers collected, regarding what the respondents consider about the future usage of external data into DWs, it could be concluded that it is a general opinion that it will increase. This is a unified view of both developers that have worked with external data, as well as developers, that have not. New opportunities and the chance to develop more powerful DWs are mentioned as driving factors. In addition, in various business environments are organizations challenging with more and more competition. As competition gets tougher, it becomes more important to know as much as possible about the market where you are active and companies want to compare their own business to others in the same branch. It may be predicted that some DW projects today miss information, which in situations could give a more valuable DW.

Today's technology and development of new technology makes it possible to manage large amounts of data and will enable easier ways to integrate external data into DWs in the future. As organizations get better support by technology, to structure their DWs, it will in turn increase the usage of external data. If companies experience that integration of external data implies a lot of extra work, it could be assumed that the pros and cons will have to be weighed up carefully before integrating external data. The Internet and Internet-based technologies may also play a key role in the inclusion of external data in the future. In particular, the Internet is already used an exchange and delivery channel for various types of data, and most organizations today already have required technology available. Respondent 8 believes that more Internet sites that are specialized in collecting and compiling data will pop up In other words: companies that perform the heavy work. A cost-effective way of gaining access to external data sources is through the Internet. In this scenario, the Internet maintains the direct relationships with data vendors and allows end users to access their online databases via the Internet. The immense information resources of the web are largely untapped by data warehousing systems. However, it should be remembered that there are some problems associated with this. Here we would like to state that data quality is crucial, and it is not an easy task to find valuable data in the enormous quantities of information available on the Internet. We agree with Strand (2000), that it does not matter how much data you have access to when performing trend-analysis or forecasting, if the quality of the data is low. The key is to have large amounts of high quality data (Strand, 2000).

8 Conclusions

In this chapter, we will highlight the conclusions of this work. The conclusions should answer the aims and objectives of the research problem stated in chapter 3.2.

8.1 Integration of external data into DWs

The responses of the study shows that slightly more than half of the developers that have participated, 7 of 12, have been working with and integrated external data into DWs. However, all of the respondents, which have or have not worked with external data, show on the importance of integrating external data into DWs. In cases where external data not have been integrated, two major causes are mentioned. Firstly, the customer organizations have not expressed any need for such integration and therefore have none external data been used. Secondly, many organizations are still struggling with their internal data and have not been mature enough, for also integrating external data.

The result of the interview study also gave at hand that there is a difference between companies acting in the private sector, compared to organizations acting in the public sector. It was clearly shown that the maturity for integrating external data was much higher in the private sector. The reason for this was not verified in the study, but a reasonable explanation is that organizations in the private market are more exposed to competition. Therefore, for them it is more urgent to acquire external data, as a complement to the internal data, in order to be able to sustain competitive edge.

Along with the investigation of the extent of external data usage, the study was also targeted towards the identification of important external domains, which were crucial for organizations to acquire data about. The most important external domains that have been identified are:

- Customers
- Competitors
- Business partners
- Business market

To acquire data about these domains are not trivial and there are a lot of different categories of external data, covering these domains. Below, we have included the categories identified in this work:

- *Branch codes*: an index that describes branch belonging.
- *Address information* (both customers and other companies): used in target campaigns, and also to keep the address register updated by external address services.
- *Statistics and marketing investigations*: e.g. population statistics, age groups and demographical data. Integrated to fully understand customers.
- *Economical information and company classification*: e.g. follow-up definitions, annual accounts information, risk class, credit rating, and risk forecast. Used as a support in for example making strategic decisions about what companies to start

partnership and trading with. A follow-up definition describes how accounts should be grouped for follow-up (how you want to view the information).

- *General enterprise information*: e.g. number of employees, what type of customers they have etc.
- *Frequently changing data*: e.g. oil prices, exchange rates for currency and stocks.
- *Data from other enterprises*: e.g. delivery dates from freight companies, data used to get a correct follow-up of costs. Integrated to analyze delivery precision

8.2 External sources

In section 8.1 important domains and categories of external data are presented. In this section, we will give some example of external sources available. Most of the sources are well-known companies, whose only business objective is to collect, compile, and sell data. Below, these external sources are presented.

UC AB
UC is a large business and credit information agency owned by the Swedish banks. They offer reports to support credit and commercial decisions, credit monitoring, and qualified financial analysis. The reports are produced from a database, which contain information on all enterprises registered in Sweden and all individuals over 16 years of age living and Sweden. This external source supplies information to support decision-making, e.g. uncover opportunities for business partnerships.

PAR and SPAR
PAR (Postens basregister) offers address services for companies that want to match their address register against a reliable source. The aim is to increase the quality for effective and professional contacts with customers, and it could for example be used to assure that all addresses are coupled to correct postal number. SPAR is a combination of PAR and the state public personal- and address register. SPAR addresses themselves to organizations that want to keep their consumer- and member-registers of private persons updated. Our conclusion is that these sources are used mainly in DWs that are implemented to support CRM systems.

SCB (Statistics Sweden)
SCB supplies official statistics in many broad fields, e.g. the labor market, the economy, trade and industry, prices, population and welfare, and housing and construction. SCB also take on commissions that range from undertaking special processing of existing statistics to providing complex information systems like data warehouses.

Other enterprise systems (business partners and suppliers)
Other enterprise systems are also thinkable sources. This could for example be data that is acquired from business partners or suppliers.

Public organizations
Public organizations often supply data that is free to collect. These sources could be county councils and communities, which for example supplies municipality data or community analyzes.

International companies
There are also more global companies that supplies and sell data. Dun and Bradstreet is an example, which provides global economical information about companies with the aim to help businesses to reduce credit risk, find profitable customers and to manage vendors efficiently.

Internet
Internet is a possible source if external data that is valuable could be found in all the information that is available. The Internet may be used to look after competitors, by watching what they offer and to what prices.

8.3 External data - opportunities and pitfalls

The result of the study shows that the main opportunities of integrating external data into DWs are:

+ Being able to compare internal numbers with external numbers
+ Becoming more flexible to changes in the marketplace.
+ Getting a more holistic view of the organization.

If considering pitfalls related to the integration of external data into DWs, the following were identified:

– Difficulties in assuring the quality of the external data.
– Difficulties in data definitions and matching problems between internal and external data.
– The costs for acquiring the external data.

To conclude, the study shows that the integration of external data will increase in the future, primarily because of the fact that new technologies constantly appears, along with an increased maturity on managing and using DWs. Finally, above all this, the constantly increased competition will force organizations to also integrate external data in a greater extension than previously, since it becomes more and more important to have information about the marketplace, business partners, customers, and competitors. Without the external data, this is unachievable and this, in turn, may lead the organization to the edge of its ruin.

9 Discussion

In this chapter, we will discuss our own experiences during the work. The dissertation and its results will also be evaluated in a wider context. Finally, suggestions for future work will be given.

9.1 Own experiences during the work

The working process has given positive experiences in planning and implementing a work on my own. This has given insights to how much time that actually is required to perform a satisfactory research result. Despite, that a project plan was established before the actual work started, it has in some periods been short of time for dealing with everything to a desirable extent. Afterwards, we have realized that the planning in some cases have been a little bit optimistic. But, there has always been a comfortable margin to complete the work before handing in different parts of the report. Another valuable experience is that we have realized the importance with a clear research problem, and to have an unambiguous idea of what the work should achieve. This means that we know in what direction the work is aimed at and contributes to accomplish a more fulfilling result. However, there have existed some difficulties to state a well-formulated research problem, which has caused adjustments. Likely, the reason for this is probably that knowledge has been gained throughout the working process, which has put the research problem into new light. Still, we do not believe that this has affected our work, and generally has the working process been smooth running.

The selection of interviews for data collecting has afterwards showed to be an appropriate technique. We have received good material that has answered our research problem and its objectives. However, it should be mentioned that it required a lot of work with e-mailing to get in contact with appropriate respondents, and to schedule a suitable time for each interview. The choice of conducting the interviews by telephone has made it easier to carry out a greater number of interviews. Many of the respondents have been geographical dispersed, which means that personal interviews would have been time consuming. It may also be assumed that many consults travel a lot in their work, and presumably it would have been more difficult to schedule personal interviews. Furthermore, it should also be mentioned that several respondents offered personal interviews. With personal interviews it would probably have been more appropriate to limit the number of respondents. A drawback by conducting the interviews by telephone was that the possibility to visit companies and the chance to get an insight in their environments was lost. One interview was carried out in personal, which was really interesting and gave a chance to see how the company worked in practice with development of DWs.

It should be stressed that interviews are much about different understandings and interpretations on the basis on the own frame of reference. Therefore, it could be difficult to assure that the interpreted understanding exactly agrees with what is stated. To cover against these potential fallacies, all the respondents have approved the interview compilations before they were included in the study. Thus, we do not believe that this has affected the results.

As indicated above, the respondents have played a central role in this work. Their agreement to take part and willingness to cooperate during the interviews has contributed to that the research problem could be answered. The selection of respondents could have been made different. One alternative could have been to only interview developers that have experience with external data. This would have implied more work to find suitable respondents, but could have meant that we had achieved more detailed answers on what external data that is integrated into DWs, and also a broader result on what external sources that is most common. The drawback with such a selection would have been that no conclusions could be drawn, concerning how common it is to integrate external data into DWs. As this was one of our primary aims in this work, we found it more appropriate to make the selection more randomly.

To conclude, this work has given insights on how to perform larger study on the own. The benefit was that we could, without any consideration into someone else, plan the work. The primary difficulty was to keep disciplined during the work. This as nobody else is dependent on the progress of the work. Finally, worth to notice is that several respondents comment that they consider it to be an interesting research problem, which has been a motivation during the work.

9.2 Evaluation of the dissertation in a wider context

When considering the presented results and conclusions have all the objectives been achieved to a certain extent. Of course, it is difficult and debatable if the interviews could constitute a common outlook for DWs. However, our assessment is that the respondents' competence in the area is adequate to give representative answers, and the result has fulfilled expected outcomes. At the end of the day, the study was not intended to build or test theory, but does offer some insights into the current usage of external data in DWs.

In most cases, the result described in the study contains no greater surprises. External data that has come to light could for the most part easily be derived to usage for support in organizational decision-making. On the basis of received answers about external data, it is difficult to state whether this could be considered as general data to integrate into DWs. It is, as mentioned before, always the requirements in the individual project that is decisive for what external data to integrate. It could with greater certainty be assumed that external domains identified as most interesting for organizations as more general. We dare to consider this conclusion as more general, since identified domains from interviews also are motivated in some literature (Alavi & Leidner, 1999).

However, the fact that many responses were greatly customer focused was a bit unexpected. It seems like a lot of data warehousing is turning its focus to customers, which in turn implies that designers of DWs must turn their focus to input that comes from external sources outside the company. This also means that the DW gets a slightly different aim. Instead of supporting decision-makers at a strategic level, the DW is aimed to supply the right information to the customer care team in an organization, by providing them with relevant and timely information. This means that data is integrated and summarized in DWs to support CRM systems. The aim with those systems is to improve the customer data with the company. In order to gain this knowledge, data must either be collected directly from the customer or purchased

from outside suppliers. Furthermore, in this case the usage of external sources also gets more important for quality checking of the own data, for instance for address verification before a larger campaign or in target marketing to certain customer groups. Instead of doing this work by using resources in the own organization, it could save a lot of time to use an external supplier. To fully understand customers, it is necessary to integrate additional demographic data about customers, which is available in external databases. It is only by integrating this data in the DW that it is possible to have a comprehensive understanding of customers. Apart from the external sources that have come to light in this study, it could be conceivable that information about customers could be collected by, for example web forms, click stream data, and call centers.

Moreover, another important discussion that has come up during this work concerns what data that should be considered as internal and external in relation to the DW. This has, to a certain degree, also affected the result of how many respondents that have been concluded to have experience with external data. For this dissertation we have considered external data as data originating from outside the own corporation. However, this has also been the general viewpoint of the respondents. The discussions have arisen when considering whether data originating from external sources, but is acquired to the DW from internal systems, where it first is stored, should be considered as external data. To put it plainly and according to the definition this should be considered as external data. Nevertheless, it is first used in internal systems, and when integrated into the DW it is acquired from internal systems. Therefore, we find it complicated to discern if a respondent that knows that some data integrated into the DW originates from external sources, but the same respondent has not used any external sources and only acquired data from internal systems. In these cases we have decided to make our own interpretation how the current respondents should be regarded. We have made our assessment on the basis of whether there have existed any requirements for external data, or if the respondent only knows that it exist external data in internal systems. The important thing is if there has existed any reason, on the basis of requirements of the DW, to integrate external data to be compared and analysed along with internal data when it comes to executive level. To exactly unravel what data that is external or not in the specific case is out of the scope of this work. Still, we wanted to discuss this issue since it has affected the result, and also with the motivation that certain confusion has occurred during our work.

In the interviews, responses on how the developers and their organizations define a DW have been collected. The answers shows that separate definitions goes a little bit apart, but the general aim and basic view of a DW is unanimous amongst the respondents. We found it important to elucidate what is referred to when speaking about DWs when we conducted the interviews. Even if the exact wording is diverse among different respondents, they all seem to agree that a DW is integrating data from several sources, and that the primary aim is to use it as a foundation and support for analysis and decision-making.

As for technology, much has been made and it is evolving every day with new kinds of tools for developing DWs. Of course, it is very important to keep up with new technologies, but to derive business value of a DW could in many cases be more related to organizational culture and access to the right data. In order to increase the business value of a DW, this study has showed on opportunities with external data, which in some cases likely could give a greater chance to obtain business payoffs.

The use of external data could enable to assess the impact of forces outside the own organization, which in turn could improve the usage of a DW to predict the need for business change or to give opportunities to expand product mix. Yet, it should be remembered that a greater quantity of data and information does not necessarily lead to enhanced knowledge creation and a better chance for improved analyses. The data must be demanded by the users and of high quality.

On the basis of the information that came to light in the interviews it cannot be stated that integration of external data is crucial for a successful usage of DWs. The current practice of data warehousing is in many cases already fulfilling its promises of delivering real business benefits, without integration of external data. Nevertheless, it may be assumed that the inclusion of external data will challenge with deeper issues. The integration of external data will and would in many cases evolve the data warehouse into a better system of knowledge management for the enterprise. Even though, the development of a DW is strongly correlated to user needs and to the aim of the DW, it is still the people that have access to the best information that may have an advantage. Moreover, businesses are not isolated from the effects of natural, political, and economic events occurring throughout the world. External perspective on the business of the enterprise seems like a chance to provide additional opportunities for implementations of DWs. However, this study has also showed that developers cannot afford to ignore the fact that integration of external data in DWs comes along with some difficulties and pitfalls

To conclude, the empirical study shows that companies in Sweden to a certain extent do not possess the maturity that is required for more extended integration of external data in DWs. On the basis of the study conducted in this work, it cannot be claimed that companies who use external data systems have a strategic advantage over those who do not. However the possibilities is out there, and it could be assumed that one key factor in implementing a successful DW project of our era will be the companies with the most robust access to different data, internal as well as external.

9.3 Ideas for future research

This report has in a comprehensive way identified and described external sources, which supplies data and information, commonly used by DW designers. A more detailed survey of external data vendors available on the market would have been an interesting continuation of this work. An investigation like that could consider the availability of external data, and therefore investigate what data that external data vendors could offer and which external data that is currently available for purchasing on the market. In addition, a research like this could also be aimed to identify what opportunities those external data vendors consider themselves to offer, in order to develop more competitive and useful DWs in the future.

The quality issue of external data has been given attention, as one of the pitfalls, in this work. No one wants to make business moves based on the assumption that the external data is 100 percent correct. Therefore, an interesting factor to research is how to evaluate the external data that is considered for integration into a DW. Such research should be investigating the reliability and quality of external data. The research would also survey how the external data vendors validate the quality of data, and how reliable the external data must be, for being integrated into the DW.

This study has verified that external data to some extent is integrated into DWs. The investigation has also elucidated what external data that is integrated, and common external data sources where this data is acquired. An interesting area for future research would be an investigation that is aimed at the usage of external data when it is integrated with internal data into DWs. How is the reports generated when internal data is compared with external data and what additional reports could possible be generated?

Finally, it would also be interesting to conduct a similar survey that focus on the users of DWs, to investigate the users' point of view and survey to what extent they are aware of the possibility to acquire external data. A study like that could give a more complete account of the situation, and may be compared with the results of this work.

References

Agosta, L. (2000) *The Essential Guide to Data Warehousing.* New Jersey: Prentice Hall.

Alavi, M. & Leidner, D. (1999) Knowledge Management Systems: Emerging Views and Practices from the Field, *Journal of data warehousing*, Vol.4, No. 1 pp. 2-7.

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

Barquin, R. & Edelstein, H. (1997) *Planning and designing the data warehouse.* New Jersey: Prentice Hall.

Bischoff, J. & Alexander, T. (1997) *Data warehouse: a practical advice from the experts.* New Jersey: Prentice Hall.

Capps, L. (2002) *Toward Pervasive Computing: Web and Customer Analytics: You Click, They Learn.* DM Review. Faulkner & Gray. http://www.dmreview.com/master.cfm?NavID=68&EdID=5017, printed from DMReview.com, 2002-04-23.

Chaudhuri, C. & Dayal, U. (1997) An Overview of Data Warehousing and OLAP Technology, *SIGMOD Record*, Vol.26, No. 1, pp.65-74.

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

Dawson, C. W. (2000) *The essence of computing projects: a student's guide*. London: Prentice Hall.

Dahmström, K. (2000) *Från datainsamling till rapport: att göra en statistisk undersökning.* Lund: Studentlitteratur.

Denzin, N. & Lincoln, Y. (2000) *Handbook of qualitative research*, 2nd Edition. Sage Publications, Inc.

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

Elmasri, R. & Navathe, S B. (2000) *Fundamentals of database systems*, 3rd Edition. Harlow: Addison-Wesley Longman.

Hackathorn, R. (2000) *Farming web resources for the Data Warehouse*, DM Review. Faulkner & Gray. http://www.dmreview.com/master.cfm?NavID=198&EdID=2221, printed from DMReview.com, 2002-04-05.

Inmon, B. (2002) *A Data Warehouse Development Methodology*, The Bill Inmon.com library LLC. http://www.billinmon.com/library/library_frame.html, printed 2002-02-03.

Inmon, W.H. (1996) *Building the data warehouse*, 2nd Edition. New York: John Wiley & Sons.

Inmon, W.H. (1994) *Using the Data Warehouse.* New York: John Wiley & Sons.

Inmon, W.H. (1999) *Integrating internal and external data*, The Bill Inmon.com library LLC. http://www.billinmon.com/library/articles/intext.asp, printed 2002-04-03.

Kimball, R., Reeves, L., Ross, M. & Thornthwaite, W. (1998) *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses.* John Wiley & Sons.

Meyer, D. & Casey, C. (1998) *Building a better data warehouse.* New Jersey: Prentice hall.

Oglesby, W. (1999) *Using external data sources and warehouses to enhance your direct marketing effort.* DM Review. Faulkner & Gray. http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=2364, printed from DMReview.com, 2002-01-31.

Patel, R. & Davidson, B. (1994) *Forskningsmetodikens grunder. Att planera, genomföra och rapportera en undersökning.* Lund: Studentlitteratur.

Poe, V., Klauer, P. & Brobst, S. (1998) *Building a data warehouse for decision support.* New Jersey: Prentice Hall.

Repstad, P. (1993) *Närhet och distans: Kvalitativa metoder i samhällsvetenskap.* Lund: Studentlitteratur.

SCB (2002) Home page for Statistiska Centralbyrån (Statistics Sweden), available on Internet: www.scb.se, collected 2002-04-12.

Singh, H. (1998) *Data warehousing: concepts, technologies, implementations, and management*. New Jersey: Prentice Hall.

Strand, M. (2000) *The business value of data warehousing – opportunities, pitfalls and future directions.* Master Thesis, HS-IDA-MD-00-013, Department of Computer Science, Univeristy of Skövde.

Söderström, P. (1997) *"Data warehouse", Datalager: verksamhet, metod, teknik.* Lund: Studentlitteratur.

UC (2002) Home page for UC AB, available on Internet: www.uc.se, collected 2002-04-12.

Watson, H., Ariyachandra, T., Matyska, Jr. & Robert, J. (2001) Datawarehousing: stages of growth. *Information Systems Management*, Summer 2001, Vol.18 Issue 3, pp. 42-50.

Appendix 1 – Interview inquiry

Hej *[namn på mottagare]*,

Mitt namn är Marcus Olsson och jag skriver till dig angående intervjupersoner för mitt examensarbete (20 p) inom data warehousing. Jag såg dina kontaktuppgifter på *[företagets namn]* hemsida, för information om hur ni arbetar med data warehouse. Det jag tänkte höra med dig är om du har möjlighet att ställa upp på en intervju eller hänvisa mig till personer som arbetar med utveckling av datalager hos er på *[företagets namn]*. Det vore kanon om du kan hjälpa mig med detta. Nedan har jag skrivit ihop lite om exjobbets inriktning, och vilken typ av personer som skulle vara intressent. Kan nämna att det rör sig om kortare telefonintervjuer, som jag beräknat kommer att ta mellan 20-40 minuter. Om det ev. finns möjlighet för personlig intervju, ser jag det som positivt.

Exjobbet behandlar som sagt datawarehouse-området, med fokusering på hur om någon extern data integreras i utvecklingen av ett datalager. Arbetet har avgränsats till att undersöka vilken extern data som isfåfall integreras, samt vilka externa data källor som är vanligast förekommande. Med extern data och externa data källor avses data som inte härstammar från system inom den egna verksamheten, utan hämtas utanför den egna organisationens gränser.

För att undersöka vilken extern data används (om någon extern data integreras överhuvudtaget) och vilka typer av externa datakällor som är vanligast förekommande, kommer jag att genomföra intervjuer med utvecklare/designers som har erfarenhet från utveckling av datalager. Med avseende på arbetets fokus är det personer som jobbat med de tidigare faserna i utvecklingsprocessen, som är mest intressent. Även om externa data inte är något som används i någon större grad, är detta också intressanta svar för min undersökning. Vore tacksam om någon person hos Er på *[företagets namn]* kan ställa upp på en kortare intervju. Intervjuerna kommer att genomföras mellan runt 15 mars - 5 april. Om det finns möjlighet för detta, vore det kanon för mig att försöka boka in ett tillfälle så snart som möjligt.

MvH
/Marcus Olsson, Högskolan i Skövde
Tel: 0500 – 487468
Mob: 0736 – 369 557

P.S Jättebra om du kan reply'a mig i vilket fall, så jag vet om du mottagit det här mailet.

Appendix 2 - Interview questions

General/Initial questions

1. In what type of company are you employed?
2. How long have you been working in this company?
3. Your background and experience in developing of data warehouses?
4. How do you define the concept data warehouse?
5. How do you and your company define internal and external data in relation to data warehouses?

Main questions and follow-up questions

6. Are you working anything with external data, and is any external data integrated into data warehouses? Have there been any requirements to acquire data from external systems?

If answer **Yes** to question **6** wills questions 7-10, 13-14, be asked. If answer **No** wills questions 11-12, 13-14 be asked.

7. What external data have been used/integrated in the DW development projects that you have been involved in?
8. How does the selection of what external data to integrate in a DW occur? What factors is behind the selection?
9. Why is external data used/integrated in data warehouses, what possibilities and benefits can be related?
10. Are there any disadvantages in using/integrating external data?
11. Are there any reasons why you do **not** work with any external data in development of DW?
12. Do you see any advantages or disadvantages, by not work with external data and external sources?
13. What type of external sources are the most common and interesting for organizations? What areas do they want access to?
14. What are the most common external data sources that provide data?

Concluding questions

15. What do you consider about the future, regarding the usage/integrating of external data and external sources in data warehouses?
16. Anything else to add?

Appendix 3 – Interviews 1-12

Interview – Respondent 1

General/Initial questions

1. In what type of company are you employed?

Answer to this has been collected on this companies website. It is a global management and IT consulting firm. The concern is established in management consulting, system transformation, information systems management and outsourcing. Their goal is to assist companies to continue to grow, and to create benefit by new technology.

2. How long have you been working in this company?

I have been working here for 3 years.

3. Your background and experience in developing of data warehouses?

I have been involved in a couple of projects in developing of data warehouses. In these projects, I had different roles and participated more or less. At the moment I am taking an active part in a large data warehouse project.

4. How do you define the concept data warehouse?

The idea of DW has been around for a long time, but the term data warehouse has not been around for so long. There are new words and concepts that pop up all the time, and many times there is new terminology, but not new things. I see a data warehouse as a system that is developed with historical data, with the aim to derive advantage of this data in analyzes and reports, to support users to make more effective decisions. The data that is stored in the warehouse can be named as not-transaction data, and a DW has totally different aim in comparison to a transactional system.

5. How do you and your company define internal and external in relation to data warehouses?

Internal data is created within the own company. External data is created by someone else, or has been collected by someone outside the own organization. This could be interpreted in different ways, and the line between internal and external could be drawn different for different people. The juridical aspect is also important. Is the own organization juridical responsible for the data, or is the data for example bought from some other provider. It is of great importance that the juridical map is clear. Especially important is that the user knows about this. What data is our own, and what data has been acquired from external data sources.

Main questions and follow-up questions

6. Are you working anything with external data, and is any external data integrated in data warehouses? Have there been any requirements to acquire data from external systems?

Yes, we have been working and integrating external data in the projects I have been involved in. This is related to the business objective of the DW, and is naturally different from project to project. What external data that has been integrated is dependent on the information needs of the warehouse. In the project that I am currently participate in; there are needs to include external data. Firstly, is data

acquired from internal systems. After that is "required data", which does not exist in internal systems, collected from external systems.

7. What external data have been used/integrated in the DW development projects that you have been involved in?
From my experience has the following external data been used:

- *Enterprise address information*: address register to all company offices, for instance in a certain branch.
- *General enterprise information*: like for example economical information about a company, the number of employees, what types of customer they have.
- *Branch codes*: Codes that describes in what branch a companies is active.
- *Individuals information*: The respondent knows that this has been used in other projects, but do not have any own experience of integrating data on individual level. Credit rating is one example that could be used.

The address information have more exactly been:
- Address (street/box, postal number, post district)
- Telephone number
- Fax number
- Name of highest managerial position
- Title of highest managerial position

8. How does the selection of what external data to integrate in a DW occur? What factors is behind the selection?
It is always related to the business objective of DW. By for example workshops, is requirements gathered, to reach the goal with the DW. What kind of questions do they want to ask? What kind of report will be generated?

9. Why is external data used/integrated in data warehouses, what possibilities and benefits can be related?
Again, it is the specific project that has the decisive influence, whether it is crucial to include external data to create benefits. Today is the development work moving to that direction; the need of external data is increasing. This to become more effective, and to be able to ask sharper questions, and to achieve a better hit-rate. One useful area is in campaigns against customers and other companies, which demand that a lot external data is bought from external sources, like for example information about companies and addresses. Another important reason is to keep the own data updated and complete with current information. For example, it is not for sure that customers inform when they changing address.

10. Are there any disadvantages in using/integrating external data?
The premier problems are difficulties in definitions. The experience is sometimes that the provider and the buyer define data different. It is not always obvious what is meant with a certain definition. This needs to be taking into consideration when matching internal and external data. It is also important to be careful from where the data is acquired. There are organizations that are specialized in providing and selling data, and this are the ones we are using. They are professionals in collecting and

compiling data from several sources. This time-demanded work we get away from, by using well-known organizations that working with this.

11. What type of external sources are the most common and interesting for organizations? What areas do they want access to?
Like I mentioned before (see question 7) it has mainly been company information, branch codes, economical information about companies, customer information, and so forth. Individual person data is also thinkable, like for instance in credit refining.

12. What are the most common external data sources that provide data?
The external data sources that was declared during this interview was:

- *The Post base register*: This source in contains address information about all addresses in Sweden. It also provides data about companies, the branch they are active in, and so forth.
- *Spar*: This is a public person address register that contains data about private persons.
- *SCB:* Statics Sweden
- *Dun and Bradstreet*: provides more global economical information about companies. For example credit refining on companies.

Concluding questions
13. What do you consider about the future, regarding the usage/integrating of external data and external sources in data warehouses?
My opinion is that integration of external data will increase in the future. This is driven by the fact that today it is easier to benefit by internal data and this is performed in a good way. As a result of this, and as tools and technology for data warehouses are developing fast and become more effective, are information requirements to extract from the warehouse increasing. The internal situation is clear, and wants to complement with external data to become more powerful.

14. Anything else to add?
No, not concerning external data.

Interview – respondent 2

General/Initial questions

1. In what type of company are you employed?
It is an in an IT-company within one of the largest Swedish local governments. It is a knowledge organization with focus on IT and business process performance for the local government. Their primary customers are public administrations and government services in the region. They are 300 employees, of which 100 are working with development.

2. How long have you been working in this company?
I have been working here for 15 years.

3. Your background and experience in developing of data warehouses?
I started to work with development of data warehouses 2,5 years ago.

4. How do you define the concept data warehouse?
A data warehouse is a system where a lot data and information is collected. Data is modeled in star schemas and stored in a database.

5. How do you and your company define internal and external in relation to development of data warehouses?
We don't have any definition of internal and external data.

Main questions and follow-up questions

6. Are you working anything with external data, and is any external data integrated in data warehouses? Have there been any requirements to acquire data from external systems?
Data has only been collected from systems within the organization where this company is an integral part. At the same time, they do not use any data from the own departmental company.

As we do not consider data that is acquired within the same organization as external, is the answer on this question no.

7. What external data have been used/integrated in the DW development projects that you have been involved in?
The only data that can be considered as external is demographical individual data, which the organisation has acquired from external systems.

8. Are there any reasons why you do not work with any external data in development of DW?
There has not been any requirements or needs to acquire data from systems outside the own organization. The data models and cubes we have created have consisted of data from our own organization.

9. Do you see any advantages or disadvantages, by not work with external data and external sources?
With external data, it could be harder to assure that data is defined in the right way, and that the same definition is used. This may be more complicated with integration of external data.

11. What type of external sources are the most common and interesting for organizations? What areas do they want access to?
Since I do not have any experience of external data, I do not know.

12. What are the most common external data sources that provide data?
I do not have that much knowledge of external sources that provides external data. I know that RSV supplies demographically data.

Concluding questions
13. What do you consider about the future, regarding the usage/integrating of external data and external sources in data warehouses?
I believe that the inclusion of external data lies in the future.

14. Anything else to add?
No

Interview - respondent 3

General/Initial questions

1. In what type of company are you employed?

It is a one-man company that I am running myself through a broad network, where consults and competent personnel are engaged in different projects. The business concept is to use appropriate tools to acquire and present the most valuable information in companies databases with respect to customers and behaviors, seen from the perspective to strengthen customer relations. The company is active in IT-oriented services, mainly with CRM-applications with the aim to work as a support for customer relations and marketing analyzes.

2. How long have you been working in this company?

I have been working as an IT-consult and marketing economist in this company for one year.

3. Your background and experience in developing of data warehouses?

I have experience of data warehousing since 10 years back in time. Over this time, I have participated in 3-4 larger development projects of data warehouses. Much of this development has been about to develop own environments, but I also have experience from development with software like Cognos and SAS.

4. How do you define the concept data warehouse?

Data warehouse, a database or data warehouse that is created with the starting point from existing data, which often concerns different functions within a company. The data warehouse is organized to be able work up the information in a process called data mining.

5. How do you and your company define internal and external in relation to data warehouses?

There are many benefits to use external data, which is referred to data that is bought from other external suppliers outside the own business. Internal data is acquired within the own company.

Main questions and follow-up questions

6. Are you working anything with external data, and is any external data integrated in data warehouses? Have there been any requirements to acquire data from external systems?

Yes, we are working a lot with external data and there are many needs to use and also to integrate external data to complement internal data, and to match external data against internal systems.

7. What external data have been used/integrated in the DW development projects that you have been involved in?

It is mainly public information and company information that is used. For example,

- *Public documents & company documents:* For example, economical data like annual account information.
- *Branch codes:* Different levels that is more or less foreseeable. It is a SNI-code that is used.

- *Risk class:* a credit evaluation in the form of credit rating and risk forecast and a recommended credit limit. Means that companies are classified.

8. How does the selection of what external data to integrate in a DW occur? What factors is behind the selection?
The requirement of the data warehouse is decisive to what external data to integrate.

9. Why is external data used/integrated in data warehouses, what possibilities and benefits can be related?
External data is important, as well as necessary. By using external data is a greater value created, as you get more data and information about your market and your customer.

10. Are there any disadvantages in using/integrating external data?
The only disadvantage I can see is that it cost money, and is expensive to buy. I do not see any technical disadvantages, and concerning the quality so are we only using reliable sources. What you possible might do is a quality test before you integrate external data in to your warehouse, which implies some extra work.

11. What type of external sources are the most common and interesting for organizations? What areas do they want access to?
As mentioned before it is mainly company information and customer information that are the most interesting areas. Also market information in different branches, competitive analysis, and geographical data are important areas.

12. What are the most common external data sources that provide data?
We are working exclusively with UC (Upplysningscentralen), which is an organization that provides a lot of information about Swedish companies. It is a business and credit information agency owned by the Swedish banks. At UC, there is a possibility to establish a subscription, which means that data could be collected when needed. It is also possible to order information about specific areas and branches. I assume that UC also have competitors that supplies and gather similar information.

Concluding questions
13. What do you consider about the future, regarding the usage/integrating of external data and external sources in data warehouses?
I think that requirements and the usage of external data will increase. This, as large benefits to could be drawn. Before, it was not possible to manage that large amounts of data, put in performance words. Today, these limitations does not exist, and at the same time as the competition gets tougher, it become more and more important to know as much as possible about the market where you are active.

14. Anything else to add?
I believe there is a lot of useful external data on the Internet that could be useful if it could be found in all that information that is available.

Interview - respondent 4

General/Initial questions

1. In what type of company are you employed?

It is a large concern (13 000 employees) that supplies IT services in Europe. The company provides consulting, systems development and integration, operation and support, product development services for customers, and software services. Their aim is to be a strategic IT partner to its customers.

2. How long have you been working in this company?

I have been working as an IT architect for 4 years in this company. I am working in the business area: public sector.

3. Your background and experience in developing of data warehouses?

My interest in data warehouses started when I did my final year project about data warehousing and information selection. I have participated in 2 data warehouse development projects, and mainly been involved in the earlier stages of the process.

4. How do you define the concept data warehouse?

A data warehouse is a compilation of timely historical data from several systems. The characteristics of historical data differentiate from operational data.

5. How do you and your company define internal and external data in relation to data warehouses?

Internal data is generated within the own business processes, and that external data is data not generated in the own business processes. External data originates from systems outside the own organization.

Main questions and follow-up questions

6. Are you working anything with external data, and is any external data integrated in data warehouses? Have there been any requirements to acquire data from external systems?

I do not have any experience of integration of external data in data warehouses. In one of the projects I was involved in, they were planning to use external statistics data in the later stages of the project, in which I did not participate.

7. What external data have been used/integrated in the DW development projects that you have been involved in?

Statistical data, but I do not know what sort of statistical information it was about.

8. Are there any reasons why you do not work with any external data in development of DW?

This is dependent on user requirements and what needs that exists. In the projects that I have been involved there has not existed sufficient needs to integrate external data. Here, the focus has been on data that is produced by internal processes.

9. Do you see any advantages or disadvantages, by not work with external data and external sources?

One advantage is concerning quality; it is harder to assure the quality of external data. You do not know how the external data has been validated. Internal data do not have to be washed in that great extent that external data. A disadvantage, by not integrating

external data, is that analyses could not be drawn how external sources influence the internal processes. But as mentioned before, you always start with the goal of the DW and what requirements that the customer have. I believe it is more common to use external data in development of data warehouses in the private sector.

10. What type of external sources are the most common and interesting for organizations? What areas do they want access to?
Statistical information.

11. What are the most common external data sources that provide data?
I do not have any experience of external sources that provides data.

Concluding questions
12. What do you consider about the future, regarding the usage/integrating of external data and external sources in data warehouses?
I think there is a need for external data and that it has a market value. This especially in the private sector, but not that much in the public sector. Generally I believe that the usage of external data will increase in the future. This as you gets better structure in the data warehouse, which makes it easier to integrate external data. Particularly, if the quality of external data could be validated.

13. Anything else to add?
No.

Interview - respondent 5

General/Initial questions

1. In what type of company are you employed?

Answer to this question has partly been collected on the company website. It is a consulting company that today has 25 employees, equally distributed among systems analysis and design, and education. Their business concept is based on user-driven business analysis and process and object oriented system development. The company is working with development of information systems, databases and also data warehouses.

2. How long have you been working in this company?

I have been working here for 3 years.

3. Your background and experience in developing of data warehouses?

I have worked with data warehouses and analyzing tools for a couple of years. This has involved different development projects among hospital and country council, SSAB industries, and Preem petroleum.

4. How do you define the concept data warehouse?

Data warehouse is defined as a process, where several components collaborate with the goal to present information for the users. This process consists of several steps: where data firstly is collected from several different source systems, to be washed and refined (referred to as staging area) before it is stored in a data warehouse in normalized form. In the next step, is data collected from this warehouse and is washed and refined further before it is stored in a analyzing warehouse. In this database is data stored un-normalized. On the basis of this database are then OLAP-cubes and reports created.

5. How do you and your company define internal and external data in relation to data warehouses?

This is a question about definition, that is important for whether we use external data or not. See next question.

Main questions and follow-up questions

6. Are you working anything with external data, and is any external data integrated in data warehouses? Have there been any requirements to acquire data from external systems?

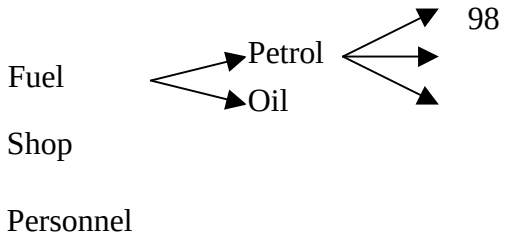
This is dependent on how external data is defined in relation to a data warehouse. We are working with data that originates from external sources. Yet, this data is first stored in internal systems before inclusion in the DW. The answer is that we are working with data that is informational extern, but still systematically intern. So the answer to this question is dependent on what data that is considered as external.

In this dissertation, we consider external data as data that originates outside an organization. Therefore it is assumed that external data is used.

7. What external data have been used/integrated in the DW development projects that you have been involved in?

External data that originates from external systems has been:

- Exchange rates
- Oil prices
- Customer information "SNI-codes": Index and codes that describes branch belonging. The branch index is collected externally, and there is a great demand for this kind of branch index in DW.
- Files and information from other companies, for example from electricity companies to get a correct follow-up of electricity.
- Geographical data: for example zip codes.
- Algorithms: to estimate calendar functionality. Different algorithms to calculate public holidays.
- Follow-up definitions: how accounts should be grouped for follow-up. It describes how you want to view the information. For example, in a case with a petrol station. Firstly there are main groups like fuel – personnel – shop, which can be broken down in different levels.



8. How does the selection of what external data to integrate in a DW occur? What factors is behind the selection?
A selection is always based on requirements and desires of the data warehouse. What data that is needed to satisfy user requirements.

9. Why is external data used/integrated in data warehouses, what possibilities and benefits can be related?
This is totally dependent on the aim of the data warehouse.

10. Are there any disadvantages in using/integrating external data?
One disadvantage is the less control of external data, and also that external data may use other keys that have to be washed together to fit in with internal keys. This could imply some extra work.

11. What type of external sources are the most common and interesting for organizations? What areas do they want access to?
Customer information – customers' geographical position in relation to selling places. This has been important in several projects, for example what customers' lives close to a specific petrol station. Also in projects within hospitals, to know where the patients are living.

12. What are the most common external data sources that provide data?
I do not have any knowledge about external sources, since I have not been involved in buying external data.

Concluding questions
13. What do you consider about the future, regarding the usage/integrating of external data and external sources in data warehouses?

I believe there are great opportunities with external data, and that the usage will increase in those projects where there exist demands. If it were possible to assure the quality of external data, it would also be possible to build services on the Internet where data could be collected. In that case, that would be important to consider in developing of data warehouses and would also influence the development. The more available something is, the greater chance that people will use it.

14. Anything else to add?

No, not concerning external data.

Interview - respondent 6

General/Initial questions

1. In what type of company are you employed?

It is a global company that develops an enterprise application solution. They work with their customers, in the spirit of long-term partnership. The company develops, implement, and takes responsibility for the on-going support of their system.

2. How long have you been working in this company?

2 years.

3. Your background and experience in developing of data warehouses?

I have been working for 1,5 years as a business consult with data warehouses and business intelligence. During this time I have been involved in one larger development project. Right now I am involved in managing and administrating of 10 solutions.

4. How do you define the concept data warehouse?

A data warehouse is a repository for data, on which several analyzing tools are used. It creates a company's information central, and gives a chance to compile information from several sources.

5. How do you and your company define internal and external data in relation to data warehouses?

Internal data is data acquired from the own internal systems, while external data is acquired from external sources.

Main questions and follow-up questions

6. Are you working anything with external data, and is any external data integrated in data warehouses? Have there been any requirements to acquire data from external systems?

No, we collect data from internal source systems. But, this is a question of definition though there exist external data in the internal systems. So, we are working with external data, nevertheless is only internal source systems used to develop the data warehouse.

7. What external data have been used/integrated in the DW development projects that you have been involved in?

External data already exist in internal systems could for example be data from companies in the same concern or sales data from other countries, or exchange rates.

8. Are there any reasons why you do not work with any external data in development of DW?

You always start with the own internal data and sales. Firstly, when this is fully developed, there is a possibility to start look for external sources. This, to be able to compare the own business against competitors in the same branch. I feel that it is immature in development of DWs that is one of reasons why we do not work with external data. However, I am sure that there exist needs for external data.

9. Do you see any advantages or disadvantages, by not work with external data and external sources?

The disadvantage is that we cannot analyze internal numbers against external. To work more with external data would imply improved analyzes.

10. What type of external sources are the most common and interesting for organizations? What areas do they want access to?
Things that would be interesting to include in a data warehouse is: branch codes, soft KPIs (e.g customer satisfaction). Also addresses to potential customers, when developing a system for campaigns.

11. What are the most common external data sources that provide data?
The external data sources that I have knowledge about are SCB, UC (UC-select disc) and Smelink.

Concluding questions
12. What do you consider about the future, regarding the usage/integrating of external data and external sources in data warehouses?
As mentioned before, I feel that development of data warehouses is immature and I think that the interest for external data will increase in the future.

13. Anything else to add?
Another area that I believe will be exciting in the future is cases where the users instead of pulling the information out of the warehouse, get the information pushed to them. This push principle means that different rules are set up against the data warehouse. The data warehouse reacts when for example a specific limit is reached (a rule is fulfilled), the information is then sent (pushed) to the user by e-mail or by SMS to a mobile phone. Another trend is writeback, which means that analyzed information is written back to the DW. Writeback could be very useful in budgeting and forecasting to update the budget with earlier periods outcomes. It could be a budget-loop, and for every round with budgeting – outcome – budget – outcome, the budget gets more and more into shape.

Interview - respondent 7

General/Initial questions

1. In what type of company are you employed?

It is a business intelligence (BI) technology services/consulting company focused on helping enterprises with their information needs. The company is specialized in delivering business intelligence solutions: better reports, better analysis and data warehousing. The number of employees amount to 120-150 consults, and the company is represented in 5 northern European countries.

2. How long have you been working in this company?

I have been working for this company for 5 years (since 1997).

3. Your background and experience in developing of data warehouses?

I have been in contact with data warehouses since I was employed in this company. This has involved 2 whole projects, where I have participated from analyze to implementation. I have also supported several other projects.

4. How do you define the concept data warehouse?

The data warehouse is the repository for the data that is collected from different source systems. Data warehousing include the whole process and is defined like: "*A process that is aimed to strategically decision-making, by supplying information that is subject-oriented, integrated, time-varying and non-volatility*.

5. How do you and your company define internal and external in relation to data warehouses?

Internal data is data that exist in a company's own systems. It may be data that originates from external sources, but as it is stored in within internal systems is considered as internal. External data is acquired directly from external sources.

Main questions and follow-up questions

6. Are you working anything with external data, and is external data integrated in development of data warehouses? Have there been any requirements to acquire data from external systems?

I do not have any experience where external data has been integrated straight from external sources into the DW.

7. What external data have been used/integrated in the DW development projects that you have been involved in?

Not in the projects that I have been involved in. But, I do know about a project, where it is of current interest to acquire external data. This project is concerning a company within the building sector, which compile and provide (sell) information to other companies. Information that should help companies to improve the planning of coming projects.

8. Are there any reasons why you do not work with any external data in development of DW?

Often, there are enough problems to structure and integrate the own information from different internal systems. External data is could often be unstructured, for example in the form of a textual document. This implies that you have to search for information in the text. It also hard to know and to assure the quality of external data.

9. Do you see any advantages or disadvantages, by not work with external data and external sources?
The disadvantage by not integrating any external data is that the chance of wider perspective is lost.

10. What type of external sources are the most common and interesting for organizations? What areas do they want access to?
The external areas that are most interesting is mainly customer analyzes and customer surveys, market research, and information about competitors. I know that this kind of information has been integrated in some data warehouses, but it has not been used in the projects where respondent I have been involved.

11. What are the most common external data sources that provide data?
There are different kinds of research-groups that provide this kind of information, for example Gartner group.

Concluding questions
12. What do you consider about the future, regarding the usage/integrating of external data and external sources in data warehouses?
I feel that this is a question of maturity of the customer. The market is probably not very consciousness about what possibilities there are with external data. In relation to that new needs emerge, I think the interest for external data will increase. This requires an easy way to integrate external data into an existing data warehouse.

13. Anything else to add?
No.

Interview – Respondent 8

General/Initial questions

1. In what type of company are you employed?

It is a company that supplies IT-related services. The company business vision is to be a leading European supplier of IT-related services, with the Nordic region as their home market. The number of employees is amount of 7000, of which 4000 in Sweden.

2. How long have you been working in this company?

I have been working here for 1,5 years. 2,5 years if we include 1 year as orderer against this company.

3. Your background and experience in developing of data warehouses?

I have worked as project leader in DW projects for 2,5 years. My focus is on users of data warehouses. I have been involved in 6 projects, of which 2 larger ones. The work has involved development with products mainly from Microsoft and Oracle.

4. How do you define the concept data warehouse?

Our definition of a data warehouse is changed a bit to reflect customer level of knowledge, but generally could be said:
"*A common storage of internal and external data for analyze, decision support, and information spread. A data warehouse makes it possible to combine and analyze information from large amount of sources, and to see relations that before only were imaginable"*.

5. How do you and your company define internal and external data in relation to data warehouses?

Internal data: data from internal systems. External data: data from external systems.

Main questions and follow-up questions

6. Are you working anything with external data, and is any external data integrated in data warehouses? Have there been any requirements to acquire data from external systems?

Yes, we are working with and integrating external data in the projects where I have been involved. However, most data that is integrated in data warehouses is acquired from internal systems (70%-90%).

7. What external data have been used/integrated in the DW development projects that you have been involved in?

There are a lot of different data:

- Economy data
- Municipality data: for example community analyses
- Branch organisation data: for example, a common organisation for grocery stores. Here is address data most common to collect.
- County council data: this data is free and could for example be, population or age groups
- Target groups
- Population statistics

- Education
- Age groups
- Age groups and pattern of movement: for example, how many and in what ages people are moving in a certain area
- Customer groups: for example, how many individuals exist in a certain customer group.

8. How does the selection of what external data to integrate in a DW occur? What factors is behind the selection?
This is always based on user requirements and how much they want to cover in the world around one. We perform a trial of needs, to find out on what level the customer wants to have on the DW. How much do they want stake in the development of a DW. We are using IRM data modelling technique for data warehouses.

9. Why is external data used/integrated in data warehouses, what possibilities and benefits can be related?
It gives more possibilities for analyzes, which is positive and strong arguments to use external data. You can learn a lot from external data.

10. Are there any disadvantages in using/integrating external data?
One disadvantage could be quality and data security. Insecure in quality of data, could in some cases arise from the fact that investigations may be subjective estimations. The risk is that this could be misinterpreted in comparison to the own data.

11. Are there any reasons why you do not work with any external data in development of DW?
In cases where the customer do not have any needs.

12. What type of external sources are the most common and interesting for organizations? What areas do they want access to?
- Customer information
- Different target groups: how do they think
- Statistics: see answer question 7
- Competitive information
- Economics

13. What are the most common external data sources that provide data?
- SCB
- County council
- Community
- Branch organizations

Concluding questions
13. What do you consider about the future, regarding the usage/integrating of external data and external sources in data warehouses?
I think that the usage of external data will increase a lot in the future. I believe that "wholesaler functions/gathering sites of data" will evolve for subscription etc. This wholesales sites, is companies that are specialized in collecting data and performs the heavy work. I mean that the IT-industry follows the same way as the old industry.

14. Anything else to add?

The importance to search for right data.

Interview – Respondent 9

General/Initial questions

1. In what type of company are you employed?

It is an IT-company with 60 employees with focus in strategies, processes, and IT supporting customer relationships. Their heritage is in Business Intelligence, and today the company focus much on the area of Customer Relationship Management – CRM. In this field, they are working in four dimensions: analytical, interactive, operational, and strategic CRM. In development, they are using standard products from USA.

2. How long have you been working in this company?

12 years. I am working as a consult within analyze and development of CRM, where DW is an integral part.

3. Your background and experience in developing of data warehouses?

I have been working with development of decision-support systems and data warehousing for 12 years. This has involved around 30 projects.

4. How do you define the concept data warehouse?

"A database of collected information from several systems, internal and external, that is used as a foundation for analysis".

5. How do you and your company define internal and external data in relation to data warehouses?

Internal data originates from systems within the organisation. External data is produced somewhere outside the organisation.

6. Are you working anything with external data, and is any external data integrated in data warehouses? Have there been any requirements to acquire data from external systems?

Yes, but it is not so common

7. What external data have been used/integrated in the DW development projects that you have been involved in?

There are companies that offer services where you can send your customer database, and they will add up and update with additional information of addresses. This also washes your own data so you assure the quality of the data, and verify that you have the right information. It is possible to subscribe services so that you have access to this information on-line.

- Demographical data
- Credit report information.
- Customer register, mainly adress register.
- SNI-codes and line of business codes

8. How does the selection of what external data to integrate in a DW occur? What factors is behind the selection?

There must exist internal needs that require the use of external data to be integrated. Users of the data warehouse have business needs and some kind of required analyzes, which requires external data.

9. Why is external data used/integrated in data warehouses, what possibilities and benefits can be related?
It is not necessity, but the use of external data may complement the internal data and could be used to generate new reports. External data may also save costs by keeping better track of customers and their current addresses. By having the right information about the customers you don't spend money on sending information to wrong addresses, or miss to send to customers that expect information. It is mainly in campaign handling against customers that this is useful. Another field of application is to wash internal data through an external source. This is useful when you for instance want make sure that your address register is current.

10. Are there any disadvantages in using/integrating external data?
No disadvantages, but it may exist difficulties with matching external data to internal data. This as the databases is structured different ways, which cause more integration

11. What type of external sources are the most common and interesting for organizations? What areas do they want access to?
As mentioned before it is mainly:

- Customer information
- Address register
- Competitive information
- Credit report information.

12. What are the most common external data sources that provide data?
The banks have a jointly company that collects and putting together register of credit reports.

- UC – Upplysningscentralen.
- SCB – offer different kinds of statistics.
- Branch organizations that offer statistics,
- Consuming statistics

Another external source is to use the Internet to look after competitors, by watching what they offer and to what prices. Nevertheless, it should be remembered that there is a big step to include it into the data warehouse.

Concluding questions
13. What do you consider about the future, regarding the usage/integrating of external data and external sources in data warehouses?
In general, the respondent thinks that the use of external data will increase. This as it is important to have correct and good information about the customers, and as a complement to internal data. Yet, the respondent means that there is so much to collect and analyze from internal system. The situation is often to first of all, structure and compile the internal data, and that many data warehouses are not ready to integrate external data yet.

14. Anything else to add?
No.

Interview – Respondent 10

General/Initial questions

1. In what type of company are you employed?

The company develop and supplies component-based business applications for medium and large enterprises. The Applications, which is based on web and portal technology, offers 60+ enterprise application components used in manufacturing, supply chain management, customer relationship management, financials, engineering, maintenance and human resource. The company has more than 3,200 employees, with sales in 43 countries.

2. How long have you been working in this company?

I have been working here since the beginning of this year, 3 months.

3. Your background and experience in developing of data warehouses?

I have been working since –95 with development of databases and data warehouses. I have technical as well as management experience of data warehouses and I have been involved in about 10 different projects. These projects have involved development of data warehouse in different levels and in diverse business environments

4. How do you define the concept data warehouse?

My definition: *"…a common business adapted platform that contain information that is stored with the purpose and the aim to deliver statistics and/or basic data for analyses and decision-making, over a historical longer period…"*

5. How do you and your company define internal and external data in relation to data warehouses?

Internal data: Data that is collected from the ERP-system that our company develop.
External data: All data that does not originate from the companies own developed systems.

6. Are you working anything with external data, and is any external data integrated into data warehouses? Have there been any requirements to acquire data from external systems?

Yes.

7. What external data have been used/integrated in the DW development projects that you have been involved in?

As we define external data to data that we do not collect from the ERP-system that we develop, this could for instance be data from older systems. External data may also be data from cash registers in shops. Also other statistics and reports. Another example is freight companies, where data about deliver dates could be integrated to analyse delivery precision. This means that these companies open up their systems for us to collect information.

8. How does the selection of what external data to integrate in a DW occur? What factors is behind the selection?

We have a foundation for an own method for this, but it is not very developed. The work usually takes form in small iterations and by workshops, to develop and show prototypes for the customer

9. Why is external data used/integrated in data warehouses, what possibilities and benefits can be related?
It is important when there is a need to integrate external data, to be able to fullfill the customers' requirement.

10. Are there any disadvantages in using/integrating external data?
When working with different data types there are difficulties to match the external data with the internal. Also that a data type could be defined in different ways, which requires additional data wash to be able to integrate it with the internal data. Moreover the respondent believes that the more data sources that is used, the more work is needed in integration.

11. What type of external sources are the most common and interesting for organizations? What areas do they want access to?
- Customer information (mainly in CRM systems).
- Cash registers in stores.
- Data from other businesses in a concern.
- Information about suppliers and deliveries.
- Other enterprise systems.

12. What are the most common external data sources that provide data?
I have not been in much contact with external data sources that for example sells data. An external source that is used is mostly other systems in for instance the same concern and suppliers. See also answer Q 11.

Concluding questions
13. What do you consider about the future, regarding the usage/integrating of external data and external sources into data warehouses?
In general, the respondent believes that companies want to have better control and make use of their data, which motivate to invest more money in data warehouses. With Internet as a distribution channel in relation to that systems becomes more open in business-to-business environments, the respondent believes that the use of external data will increase. This will provide an easier way to collect information from other businesses like partners and suppliers.

14. Anything else to add?
No.

Interview - respondent 11

General/Initial questions

1. In what type of company are you employed?

This respondent is working for the same company as respondent 4, but at a different department in another city.

It is a large concern (13 000 employees) that supplies IT services in Europe. The company provides consulting, systems development and integration, operation and support, product development services for customers, and software services. Their aim is to be a strategic IT partner to its customers.

2. How long have you been working in this company?

I have been working since 1973 (29 years) in the IT-business and the company I used to work for got purchased of the company where I am now employed.

3. Your background and experience in developing of data warehouses?

I have worked in development of economics and personal systems, and in connection with this developed decision-support systems. For 10 years I've been working with data warehouses and this has involved development of small and larger systems. Have participated in 7-8 projects, and this projects has only been in the public sector.

4. How do you define the concept data warehouse?

"*…A repository that make data compilation easier, to use as basis for analyses…*" It is a star-schema that is hierarchical structured, from where dimensional cubes are created.

5. How do you and your company define internal and external data in relation to data warehouses?

Internal data: data from systems within the organisation.
External data: data from systems outside the organisation.

Main questions and follow-up questions

6. Are you working anything with external data, and is any external data integrated into data warehouses? Have there been any requirements to acquire data from external systems?

No, not in the projects I have been working on.

7. Are there any reasons why you do not work with any external data in development of DW?

In the projects I have been involved in, there have not been any needs for external data. The users have not demanded it. I think the reason for this, is that I have only been working in projects in the public sector. I believe that it is more common in data warehouses that are implemented in the private sector.

8. Do you see any advantages or disadvantages, by not work with external data and external sources?

One possible benefit may be that we do not have to wash the data as much.

9. What type of external sources are the most common and interesting for organizations? What areas do they want access to?

Since, I do not have any experience of external data I don't know.

10. What are the most common external data sources that provide data?
Have used any external sources in the projects I have been involved in, but I know that SCB provides statistics. Maybe companies that compiles stock exchange quotations.

Concluding questions
11. What do you consider about the future, regarding the usage/integrating of external data and external sources into data warehouses?
I think the use of external data will increase, above all in the private sector. This as companies more and more wants to compare their own business to others in the same branch. But again, this is dependent on the aim for the data warehouse.

12. Anything else to add?
No.

Interview - respondent 12

General/Initial questions

1. In what type of company are you employed?

I am working on a Consult firm in the IT industry and in the present situation we are 40 employees. Our business concept makes our direction clear:
"…To be a technical knowledge company that deliver IT-consulting services for companies and organisations in the Göteborg region. This comes about by competent consults that puts customer loyalty and quality in the forefront…".

2. How long have you been working in this company?

4 year.

3. Your background and experience in developing of data warehouses?

I have been working with data warehousing for 2 years. I have been involved in 2 projects and in these I had different roles. In the first I was responsible for the reporting side and analyse data. In the current project I am total responsible for the whole ETL-chain (Extract, Transact, Load).

4. How do you define the concept data warehouse?

"A data warehouse is a way to integrate data from several information sources, to make it accessible for the end-users". It is much about the creation of common concepts, and to find the common denominator for all the information sources. This gives new potentials to see new relations and perform analyses that have not been possible earlier.

The work includes the ETL-chain that consists of:

- Build a structured respository model.
- Extract data from existing systems
- Move the data to the data warehouse
- "Wash" the data in respect to placed rules, before it is actual stored in the data warehouse.

5. How do you and your company define internal and external data in relation to data warehouses?

Internal data: data from operational systems within the own corporation. Control of data takes place internal.
External data: data outside the own organisation. Control of data takes place external.

Main questions and follow-up questions

6. Are you working anything with external data, and is any external data integrated into data warehouses? Have there been any requirements to acquire data from external systems?

The answer got to be yes and no. We acquire data in form of ordered files from an external system in another company, yet inside the same concern. This data is external in the sense; we do not have any control at all over it. Still, the information has a direct relation to our business. The information is common for the concern, but the responsibility and the control lies with respectively company. To make it easier, we interpret this a YES continuous.

7. What external data have been used/integrated in the DW development projects that you have been involved in?
We have acquired information about apartments and locals from different real estate systems. See also answer 7.

8. How does the selection of what external data to integrate in a DW occur? What factors is behind the selection?
This is controlled by requirements. What are the needs for information and what surplus value does it add with integration? Does it generate any new possibilities to use the data?

9. Why is external data used/integrated in data warehouses, what possibilities and benefits can be related?
External data may provide tracking of relationships and analyses, and this in turn could mean more effective steering of the business.

10. Are there any disadvantages in using/integrating external data?
The risk lays in that the external information, that we do not have any control over, in some cases could be directly incorrect. A solid validating process is required to guarantee data quality.

11. What type of external sources are the most common and interesting for organizations? What areas do they want access to?
Different marketing investigations and statistical information. In one of our projects this could for instance be age groups and age categories, which are in need of rental apartments.

12. What are the most common external data sources that provide data?
SCB and similar organizations

Concluding questions
13. What do you consider about the future, regarding the usage/integrating of external data and external sources into data warehouses?
It is hard to say. The internal information will naturally have greater priority. It is not until everything internal is prepared and in order, that new possibilities may be drawn by the inclusion of external data. But, I do think that the needs of external data will increase as soon as the internal data is ready. This mainly in the private sector.

14. Anything else to add?
It is hard to give some general answers. How different companies act is dependent on the type of business, their degree of maturity in relation to data warehouses and how far their work with DW development has reached, and also the quality in their internal systems