ACQUIRING AND INTEGRATING EXTERNAL DATA INTO DATA WAREHOUSES

Are You Familiar With the Most Common Process?

Mattias Strand, Benkt Wangler, Carl-Fredrik Lauren

Department of Computer Science, University of Skövde, Box 408, Sweden

Keywords: Data Warehouse, Data Integration, Data Acquisition

Abstract: Data warehouses (DWs) has become one of the major IT-investments during the last decades and in order to fully exploit the potential of data warehouses, more and more organizations are acquiring and integrating external data into their star-schemas. However, the literature covering external data acquisition and integration is limited. Therefore, in this paper the results of an interview study conducted among banking organizations are presented. The study aimed at identifying different approaches for acquiring and integrating external data into DWs. The results show that there are many different approaches for the acquisition and integration, depending on the purpose and structure of the data being acquired. In addition, the most common external data acquisition and integration process is presented and discussed.

1 INTRODUCTION

The concept of data warehouse (DW) has evolved out of two needs: the business requirement of a company-wide view of information and the need from IT departments to be able to manage company data in a better way. Current database technologies make it possible to effectively store large amounts of internal data in a well-organised way (Connolly and Begg, 2002).

However, to be able to better plan for future success of the corporation, more than just the internal data is needed (Devlin, 1997). Furthermore, Inmon (1999) argues for the importance of integrating internal and external data, as it creates an enhanced foundation for decision-support, i.e. it allows the decision-maker to contrast and verify internal information from an external perspective. On the contrary, Inmon (1999) also claims that the integration of external data with the internal data is the most difficult problem of external data incorporation.

Still, the literature covering this issue is limited and fragmented. Therefore, in this paper, we present the results of an explorative interview study partly aimed at outlining and describing different issues on external data acquisition and integration. In addition, a major purpose of the paper is to outline specific problems that make the integration so problematic. This will be done by concrete examples and explanations as to their relation to the integration efforts.

The rest of the paper is outlined as follows. Firstly, related work is presented in Section 2. In that description we also define external data and give a résumé of existing literature on external data acquisition and integration. Section 3 describes the interview study conducted. In Section 4, analysis and results are given. The paper is concluded (Section 5) with a discussion on the future for external data incorporation and ideas for future work.

2 RELATED WORKS

In this section external data will be defined and introduced. In addition, although the literature coverage of external data acquisition and integration is limited some ideas from related literature will be presented and discussed.

According to Singh (1998), one of the main reasons for a DW is to set free the information that is locked up in the internal operational databases and

Strand M., Wangler B. and Lauren C. (2004).
ACQUIRING AND INTEGRATING EXTERNAL DATA INTO DATA WAREHOUSES - Are You Familiar With the Most Common Process?.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 508-513
DOI: 10.5220/0002601905080513
Copyright © SciTePress

to mix it with information from other external sources. Singh (1998) further advocates that organizations should increase their acquisition of data originating from outside their own systems boundaries. Inmon (1996) is even more outspoken concerning the contribution of external data in a DW. He claims that even though external data do not say anything directly about particular companies, they may still give a lot of valuable information about the environment that the company must work and compete in.

Defining external data is not trivial, since the externality may be considered in different ways. From a DW perspective, several notions are to be found related to the acquiring external data from outside the organization (e.g. Kimball, 1996; Kelly, 1996; Damato, 1999; Oglesby, 1999) However, for this work, the following definition was adopted (Devlin, 1997, p. 135):

"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."

The generality of the definition was considered as important, since the aim of the work was to outline, from a broad, explorative perspective, current approaches for acquiring and integrating external data.

External data has, in a comparison with internal data, some different key characteristics, especially from an acquisition and integration perspective. First and foremost, the acquiring organization has less of the structure of the data. Even worse, the conceptual meaning of the data may also be difficult to interpret. Inmon et al. (2001) also emphasize the transformation problems of external data, as important metadata may be missing, making transformations almost impossible. Still, organizations tend to reduce the impact of these potential problems, by acquiring external data from specialized data suppliers (Strand et al., 2003). Kimball (1996) refers to these as syndicate data supplier.

3 THE INTERVIEW STUDY

Data for the study were primarily collected through interviews. The following steps were used to guide the preparation and accomplishment of the interviews:

1) Selecting the respondents. For this study, it was considered as important to interview organizations that were long-gone in there DW initiatives. Therefore, the financial sector was considered suitable, with a specific focus on banks. The respondents were selected by contacting all companies labelled bank on the yellow-pages. A total amount of 24 companies were contacted. However, since the scope of the study was rather narrow (the companies should have a DW and incorporate external data into it), only 10 banks remained. 8 out of these 10 companies agree on participating, making the final "interview rate" 80%.

2) Constructing the interview questions. The set of interview questions was split into four main groups, besides the usual introductory and concluding questions. The questions were arranged according to the four main activities of the external data integration process, i.e. identification, acquisition, integration, and usage (Strand, 2003). In this paper, results related to acquisition and integration will be accounted for.

3) Initiating the interviews. The interview questions were sent to the respondent in advance so as to let them read through the questions and reflect upon them before the actual interview. A personalized cover letter accompanied the interview questions, the aim of which was to explain the purpose of the study and to guarantee the confidentiality of the collected data, and to explain how the material was to be compiled and validated.

4) Conducting the interviews. Every interview lasted for approximately 60 minutes. After the interviews, the answers were transcribed and sent back to the respondent for reviewing and authorization. In this way, we avoided errors in the material and misinterpretations were corrected.

4 ANALYSIS AND RESULTS

In the following section, the analysis of the interview study is presented, along with the results of the analysis.

4.1 Issues on External Data Acquisition

In this section, issues related to external data acquisition will be described and discussed. Integration approaches are given account for in the next section (Section 4.2).

Firstly, there seems to be two main approaches of the frequency of which the consuming organizations are acquiring the external data from syndicate data suppliers. The most common out of these two was to use a *subscriber service approach*. Subscriber service means that the consumer receives external data on a regular basis, i.e. as any other subscription, according to the established contract

between the two parties. This approach was used by five out of eight corporations (63 %). The second most common approach for acquiring external data is the *on-demand approach*. Corporations using the on-demand approach contact their suppliers when they need external data. This approach was used by three out of eight corporations (37%). The most common motivation for acquiring the data On-demand Approach was solely base on monetary issues. Organizations that had no need for external data on regular basis did not want to pay in vain for the rather expensive data.

One of the respondents claimed that they were only randomly acquiring data from external sources and represents thereby a third, minor approach, i.e. randomly. This approach differs from the on-demand approach since it is not based upon an existing supplier-consumer relationship. Instead a contract is established when needs arise. Moreover, one of the respondents stated that they used both subscription and on-demand for acquiring external data, and that explains the fact that 8 respondents were applying the first two approaches. The reason given for applying both approaches was that the supplier which the organization was subscribing from could not supply with all external data needed. Occasionally, the organization needed other data and therefore they acquired it from another data supplier.

The next issue focused upon was related to how the external data was distributed to the consuming organization. The results of the interview study shows that the external data could be distributed from the suppliers in several different ways. The corporations receive external data from the suppliers by; 1) File Transfer Protocol (FTP) technology, 2) have a DVD-ROM or CD-ROM sent to them, 3) have the external data attached in an E-mail or 4) by accessing a Web-hotel. The FTP technology was the most common approach applied. Seven out of eight (88%) corporations used FTP to receive the external data. The reason for using FTP was that the organizations already had the necessary, underlying technology and that it gives an opportunity to automatically integrate the data, without any manual activities needed. The data was simply pumped into the data transformation tools in the same manner as the internal data. The second most used approach was to receive external data sent on a CD-ROM. This approach was never used stand-alone, but as a complement to FTP transfer.

The other two approaches, i.e. the external data attached in an e-mail or by accessing a Web-hotel were not given too much attention. However, the e-mail approach was considered as convenient when only small amounts of data was needed and when it was needed instantly. In such cases, the external data was attached to the mail and sent to the administrator of the DW. This approach was also used as a complement to FTP. The final approach mentioned in the interviews was the access a Web-hotel. In this case, no other approaches were used. When new external data is available, the corporation receives a message from their supplier and may then access the Web-hotel and download the external data.

The third issue concerned the tools used for extracting, transforming, and loading (ETL) the external data into the data warehouse. In addition, the interviews also brought some interesting viewpoints on whether the ETL operations are automatically performed by these tools or whether there are operations that are performed manually. These tools discussed in this section will for now on be referred to as ETL-tools.

According to the results of the interviews, five out of eight (63%) organizations applied commercial ETL-tools when integrating the external data. The reasons given for this by the organizations are well aligned with the advantages mentioned by Gleason (1997), i.e. the built-in support for metadata generation and the avoidance of the sometimes costly and not always successful, products of in-house development. Especially the generation of metadata was considered as important and this is also emphasized by e.g. Devlin (1997), Inmon et al. (2001), and Marco (2000).

On the contrary, the rest of the respondents were applying in-house developed ETL-tools. Their strongest argument for not investing in commercial tools was the large monetary investment related to such procurement. Another argument given was that by developing an own, in-house solution, the organization preserves the control over what data enters the DW. The respondent claimed that commercial ETL-tools were functioning in a black-box matter, giving that raw external data was entering the tools and resulted in external data mixed in the DW without actually any control of the quality of the data.

The drawback of such solution was that it is difficult to develop such commercial look-alike tools and therefore, for being able to achieve the overall same functionality, the organization had to invest in additional programs. Occasionally, this also resulted in manual activities, where it was difficult to integrate the different components.

4.2 Characterizing External Data Integration Approaches

After the data has been acquired and transformed according to given rules, it will, in some way or another, be integrated into the DW environment. The integration approach chosen depend on the purpose of the external data. Therefore, in this section, different integration approaches will be described and exemplified.

As a result of the interview study and related literature, the following four integration approaches has been identified (Figure 1):

- Star-schema dimension integration
- Dimension attribute integration
- Attribute value integration
- Spread-sheet integration

The first approach is to integrate the external data into a separate dimension in the DW. By integrating external data into separate dimensions the external data is not mixed with the internal data. This may be contributory as external data is sometimes of poor quality. For example, an organization may want to add a business partner's customer data to a star-schema designated towards sales (Figure 1A). However, if the organizations do not want to mix its internal customer data with the business partner's data, that data is then stored in a separate dimension. This approach was applied by two out of eight corporations.

The second approach relates to separate attributes containing the external data, presented and store in dimension which is mostly based on internal data. As an example, one of the respondents mentioned customer ratings. The respondent explained that his organization bought customer ratings from a syndicate data supplier and used that data to segment customers in different types of marketing campaigns. An example of attribute integration is presented in Figure 1B. This approach was used by three out of eight corporations.

The third approach was to consider the external data integrated on a attribute value level, i.e. mixing the data store under one attribute, from both internal and external sources. In this way, the external data is

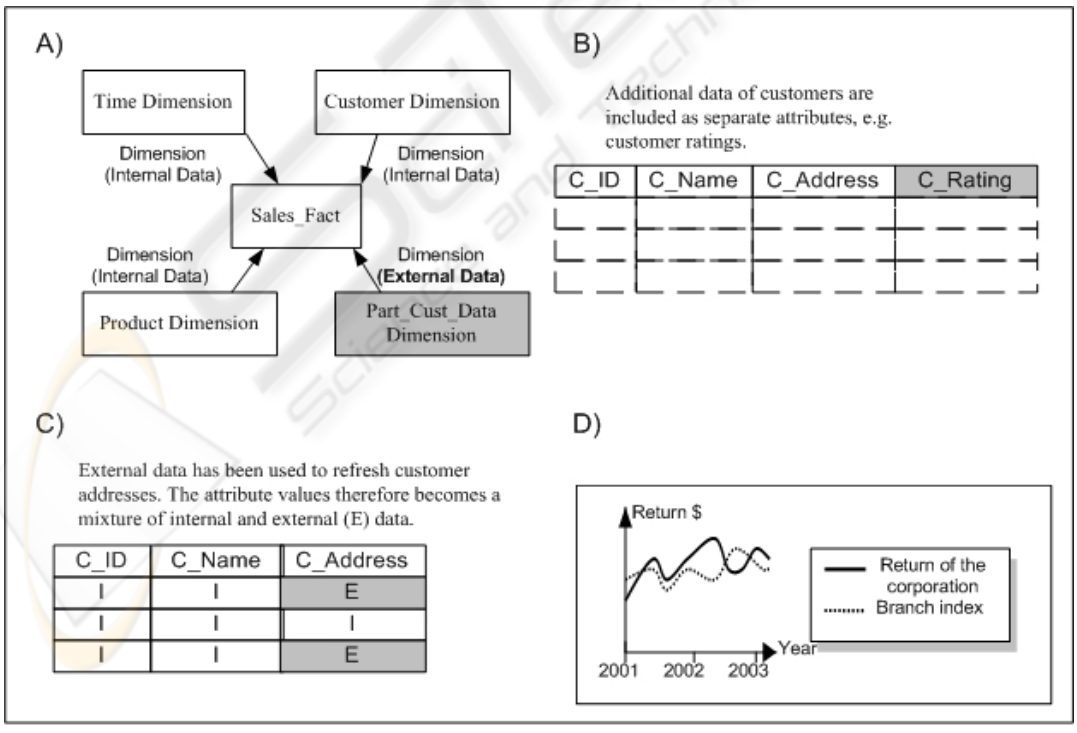

Figure 1A: Example of external data integrated as a star-schema dimension. Figure 1B: Additional data about customers is acquired externally and included as separate attributes, Figure 1C: The C_Address attribute of the customer dimension has been refreshed with external data. Figure 1D: Example of external data integrated via the interface of the GUI

strongly integrated with the external data, making it transparent to the user whether the data is originating from the internal systems or if it is integrated directly from an external source. The most common example given in the interviews was customer address refreshments, in which customer addresses acquired from the internal systems are refreshed with external data. In this sense, the DW becomes the washing machine, the internal data is the dirty laundry, and the external data becomes the soap (Figure 1C). When the customer address data are to be updated, fresh addresses is purchased from a data supplier and the old value is updated and replaced with the new one as the DW is refreshed. An example of this is shown in Figure 1C. This approach was used by four out of eight corporations.

The fourth and last approach was to store the external data in a reference table (Figure 1D). In this sense, the external data is not integrated into the DW. Instead, it is integrated in a spread-sheet manner (concept adopted from Devlin, 1997), where the querying tool uses external data, collected from a reference table outside the DW, to plot the external data into a referential line, allowing for comparison with the internal data result. This approach was only vaguely expressed during the interviews. Still, it was considered as relevant, since Devlin (1997) mentions reference tables as a possible way to handle external data that is somewhat "integration-awkward".

To conclude this section, it was clearly shown that external data may be acquired and integrated in many different ways. Probably, if this study was conducted in a few years time, there would probably be more than one organization that applied more than one integration approach. However, currently one may claim that the most common process of acquiring and integrating external data would be outline as follows; the data is acquired via a subscription service and distributed to the organization by FTP technology. Thereafter, it is scrubbed and cleaned in a commercial ETL-tool and integrated with the internal data as attribute values.

4.3 Common problems

The interviews gave the following problems as most common for hindering organization to integrate external data:

- Data structure problems
- Poor data quality
- Expensive ETL-tools

The most common problem regarding integration of external data into DWs is the difference in the data structure of external data compared to the structure of the internal data already stored in the DWs. This problem is also described in the literature by Devlin (1998) and Inmon et al. (2001). The data must be transformed to fit the data structure of the DW and this is a very time consuming and costly process. This problem could be solved if the data suppliers adjust the external data in the way that the corporations request. However, this requires that the data actually is integrated from external data suppliers. This is the most common approach, but far from the only (Strand et al., 2003).

The problem with poor data quality is a well known problem and mentioned in literature by e.g. Adelman (1997), Inmon et al. (2001), and Strand and Olsson (2003). The information gathered from the interviews presents three main problems regarding data quality. 1) The age/staleness of external data. Data acquired from external sources could be old and when old external data is integrated and then used in a DW, the result is not accurate. Internal data is according to Inmon et al. (2001), time-stamped before it is integrated into a DW. External data must also be time-stamped, but as the source of the external data sometimes is unknown, it may only be time-stamped from an integration perspective and not from a factual age perspective. You may never know for how long the data has been stored at its source, before it was acquired. From the authors' point of view, this problem is usually related to data received from other sources than data suppliers, as data suppliers generally deliver data that is time-stamped. 2) The origin of the external data may be unknown, making it difficult to rely on. Also in this case, it was shown that organizations tend to trust well-known data suppliers and thereby make the origin of the data known. 3) The most obvious problem related to data quality is incorrect data, which may result in important decisions made on incorrect numbers or facts. This is an issue that is difficult to solve, since it depends on how suspicious you are as a decision-maker and on how many sources you base your decisions on. If you solely use the DW data as a baseline for decision-making, you may become very sensitive. However, if you complement the warehouse facts with data and information from other sources, the sensitivity is dramatically decreased.

Finally, advantages and disadvantages of expensive, commercial ETL-tools have already been discussed in detail in Section 4.1 and will not be given more attention.

5 DISCUSSIONS

Based on the experiences and results of the interview study, it is interesting to discuss the

duality of the future of external data integration. In alignment with the results of the interview study presented in Strand and Olsson (2003), indicating that the integration of external data will drastically increase in the future, the result of this study shows that the increased usage of external data may take two different faces. The respondents of this study became divided into two equally big groups, in which one group claimed that the integration of new external data will increase, whereas the other group claimed that it is no necessity that new data will be integrated. Instead, organizations must become much better on fully exploiting the external data that they already integrate. Mostly, economical reasons underlined the second viewpoint, since most organizations claimed that external data (acquired from syndicate data suppliers) is very expensive and therefore, to be able to achieve return on investments, it must be much more exploited. However, other reasons were also presented and security issues and regulating laws were indicated as barriers for increased external data incorporation. Obviously, when acquiring and distributing data from external sources, holes are opened up in the system and these holes must be kept to a minimum.

The respondents aiming for an increased integration of new external data, advocated as their strongest reason that the DW becomes more and more integrated with other systems and this generates new needs for different types of data. Nowadays, it is common that business intelligence initiatives are focused around a technical solution, with a DW in its core. This idea is well supported in literature related to DWs solutions, e.g. Tiwana (2000) and Salmeron (2001).

REFERENCES

Adleman, S., 1997. Data Quality, *Data Warehouse: practical advice from the experts*, (Bischoff, Joyce and Ted Alexander, eds.), New Jersey: Prentice Hall, 122-134.

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

Damato, G. M., 1999. *Strategic information from external sources – a broader picture of business reality for the data warehouse*, acquired from http://www.dwway.com, printed 2003.03.20.

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

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

Inmon, W. H., 1996. *Buliding the data warehouse*, New York: John Wiley and sons, 2nd edition.

Inmon, W. H., 1999. *Integrating internal and external data*, The Bill Inmon.com library LLC, acquired from http://www.billinmon.com/library/articles/intext.asp., printed 2003.02.23.

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

Kelly, S., 1996. *Data Warehousing the route to customization – updated and expanded*, New York: John Wiley & Sons.

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

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

Oglesby, W. E., 1999. *Using external data sources and warehouses to enhance your direct marketing effort*, acquired from http://www.dmreview.com., printed 2003.02.22.

Salmeron, J. L., 2001. EIS data: findings from an evolutionary study. *Journal of systems and software*, Vol.64, Issue 2, 87-172.

Singh, H., 1998. *Data Warehousing: Concepts, Technologies, Implementations, and Management*, New Jersey: Prentice Hall.

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

Strand, M. and Olsson, M. (2003) "The Hamlet dilemma on external data in data warehouses" in *Proceedings of the 5th International Conference on Enterprise Information Systems (ICEIS) - Part 1*, Olivier Camp, Joaquim Filipe, Slimane Hammoudi and Mario Piattini. (Ed.), April 23-26, Angers, France, pp.570-573.

Strand, M. Wangler, B. and Olsson, M., 2003. Incorporating External Data into Data Warehouses: Characterizing and Categorizing Suppliers and Types of External Data". In *AMCIS'03, Americas Conference on Information Systems.* 2460-2468, (CD-ROM).

Tiwana, A., 2000. *The Knowledge Management Toolkit: practical techniques for building a knowledge management system*, New Jersey: Prentice Hall.