The design of a Delta Impact Analysis model for Data Warehouses

Master thesis final project

J. Kok, BSc.
Business Information Technology
November 26th, 2007

Graduation Committee:
Dr. R. Müller (UT, MG, IS&CM)
Dr. ir. M. van Keulen (UT, EEMCS, DB)
S. Dijkstra (BI4U)

Management Summary

Motivation
The field of data warehousing has emerged over the last decades. A data warehouse is developed at a moment in time to support business intelligence for an indefinite period of time. During the life of the data warehouse the world around it evolves, including the systems that are a source for the data warehouse. In order for a data warehouse to remain functioning and guarantee the quality of its data, it needs to be adjusted to the evolving world around it. The concept of *Delta Impact Analysis* is used by the company BI4U for the activities of analysing the impact of specific changes. This concept is important because it provides insight into how a data warehouse can be adjusted to the evolving world around it. The motivation to perform this research was the fact that a clear definition on the concept of DIA and what it comprehends was lacking.

Goals
The main goals of the research were to examine the topic of DIA in practice, to gather insights from literature and other research, to design and develop a practical model for DIA, to test the DIA model, and to provide recommendations to better support changes in data warehouse source systems. These goals resulted in the following main problem statement: *How can a Delta Impact Analysis model be designed that supports the process of analyzing the impact of changes in a data warehouse source system situation?*

Approach
The research approach is based on the *design science* framework by Hevner in combination with *action science* theory to validate the resulting DIA model from the research. The design science perspective resulted in an approach that is both rigorous, by performing a literature study, and relevant, by applying the research to practice. The research started by investigating the concept of DIA in practice at BI4U, in order to provide more insight into what it comprehends and what was relevant for the focus of the research. Next a thorough literature study was performed. Finally an artifact was proposed, a model for the process of DIA, which was validated in practice with a field study.

Results
The research offers several contributions. The first contribution is the fact that it was established that several aspects of DIA were unclear at BI4U. No clear guidelines existed with regard to what tasks exactly were part of DIA, and how these should be performed. Also no clear framework was present concerning which types of changes could occur and what kind impact these could have. However, a better vision on the scope of DIA was established, which provided a point of focus for the literature study.

The second contribution is the fact that in the literature study insights from the field of impact analysis were applied to that of data warehousing. Categorizations are offered to classify delta and impact. The delta can be distinguished by its origin: schema, semantic or business rule. A *schema* delta concerns structural changes to a source database, a

semantic delta concerns changes to the meaning or representation of data in the source, and a *business rule* delta concerns changes to the meaning of source data for the data warehouse. The impact can be distinguished as: *information-preserving* or *information-changing* impact on the information supply to the data warehouse. The 'changing' type can be divided by an information supply that has been *reduced, increased* and *redefined*. For the different delta categories it is specified what impact categories can apply, this provides a bridge between defining a delta and determining the impact in impact analysis. To deal with the different types of information-changing impact one has the choice if the core of the data warehouse is *not modified, modified* and/or *modified with regard to historic data*.

The final and most significant contribution of this research is that the results of the first two contributions are combined and applied in an artifact. This artifact is a model for the process of Delta Impact Analysis. The model describes seven main steps that should be performed to execute a solid DIA, from identifying the concrete change and impact till a proposed solution including required resources and costs, communication of this and an evaluation. Evaluation criteria for the process were determined and used for a practical field study to validate the proposed model of this research. The field study provided insight into which aspects can be improved in the future.

Conclusions and discussion

The result of this research, the model for the process of Delta Impact Analysis, provides a solid starting point for tackling source system change situations for BI4U, and also for others in the field of data warehousing.

Several recommendations can be given with regard to practice. First of all the model should be improved by consistently using it to execute Delta Impact Analyses and adjusting it according to the outcome of the evaluation of the process. Also it will be interesting to perform future research on the usability of existing tools and/or the development of a tool to support the steps in the DIA process. With regard to the scientific field several recommendations can be given for future research. For instance more research could be done concerning which resource and cost estimation models and techniques could be used for DIA. Also it will be interesting to research how the DIA model can be expanded to a model that can be applied to tackle any type of change in implemented data warehouse solutions. The topics of *data warehouse evolution* and *meta warehouse* were identified as topics of interest that can support the process of DIA.

Preface

I have dedicated the past 6 years of my life to the development of my knowledge in the field of Business Information Technology. I have spent the past seven months to perform the final project of my MSc education, with this thesis as a result. In the thesis, the results of my research are discussed concerning the design of a Delta Impact Analysis model for Data Warehouses. These results provide a solid starting point for tackling source system change situations for companies active in the field of data warehousing.

I would like to thank Roland Müller and Maurice van Keulen from the University of Twente for their guidance during my final project. During times of uncertainty they were there to keep me on track.

On a personal level I owe my gratitude to my girlfriend, friends and family for supporting me trough the highs and lows that came with this project.

Finally I would like to thank the company BI4U for those resources that were put at my disposal.

Jurriën Kok
Enschede, The Netherlands
November 2007

Table of Contents

List of Figures

List of Tables

1 Introduction

At the start of the computer era most businesses focused on automating their operational activities. These days businesses are more and more interested in using IT to provide them with management information and decision support tools. While much information supporting these new requirements of managers is stored by traditional operational systems, these systems are not designed to be used for this. This can make it difficult to find the right management information, if it can even be provided through the existing user interface. Just adjusting the user interface to provide the required management information is often also infeasible since the data is stored for operational use and it is computationally intensive to provide a useable management overview. The concept of a data warehouse is that of extracting the data from the operational sources and structuring this in a way that it is useable for creating management overviews. The field of data warehousing has been developing over the past decades. [CB05]

Connolly and Begg give a definition of data warehousing as:

"*A subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.*"[CB05]

This means that regarding the data, a data warehouse is focused on the subjects of the organization, instead of the business functions. It integrates multiple systems of the organization to provide one overall view. An operational system is focussed on supporting the business processes and providing the relevant information at the time of executing the business processes, where the data warehouse provides insight into which information was accurate at which time. The data warehouse is updated with intervals from the source systems and is therefore non-volatile.

It is common knowledge that computer software and systems evolve, small and big updates are usually performed within the life cycle of one application, and sometimes (parts of) systems are replaced or new systems are introduced. The evolving of operational systems that are a source for a data warehouse can cause several problems for the functioning of the data warehouse if they are not adapted to. In this research these problems will be discussed in more detail and a solution to deal with these evolving source systems will be proposed by the means of model for the process of *Delta Impact Analysis*.

Delta Impact Analysis, a term used by BI4U, encompasses all activities performed to identify the consequences of a planned or performed change in the source system situation for a data warehouse. The goals of the DIA process are to gain more insight into a change and its impact, to determine the solution directions and to determine and structure the required adjustments for the data warehouse in order to be able to cope with a changing data source. [BI06a] During this research the following definition for Delta Impact Analysis has been determined:

Delta Impact Analysis encompasses all activities performed to identify the consequences and required counter actions of a planned or performed change in the source system situation that influences, or even obstructs, the correct extracting, loading, transformation and/or analysis of the data by the data warehouse.

This chapter gives an introduction to the company BI4U and the motivation to perform the research. Next the objectives of the research are discussed, followed by a definition of the research statement and questions. Finally the scope and the research framework used for this research are discussed.

The rest of this research document will be as visualised in Figure 1. Chapter 2 will provide more insight into the topic of Delta Impact Analysis as currently present within BI4U. In Chapter 3 concepts regarding Delta Impact Analysis from scientific research and experience will be discussed. The knowledge gathered in the first three chapters was the inspiration for Chapter 4: the model for Delta Impact Analysis is proposed. Finally in Chapter 5 the validation of the research is discussed. In Chapter 6 the research will be concluded and discussed and recommendations regarding future research interesting to the field of DIA will be given.

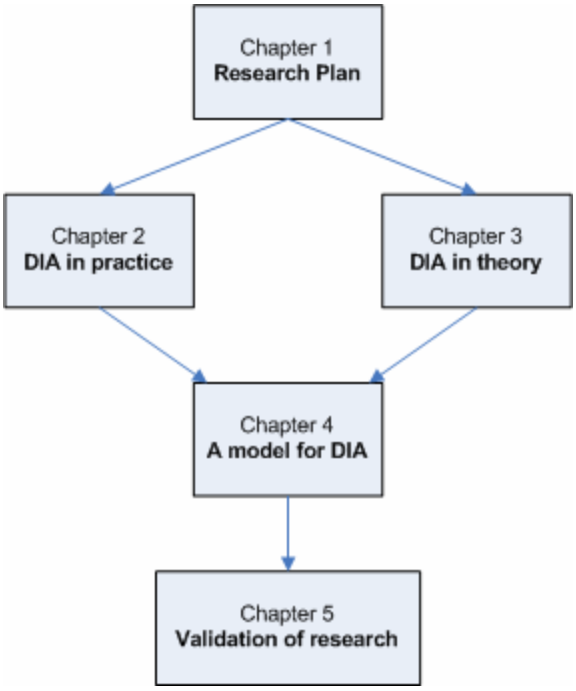


Figure 1. Thesis structure

1.1 Background of BI4U and data warehouses

BI4U was founded in 2002 and currently (2007) has over 50 employees. BI4U used to be active in employing its personnel at other companies, but since two years the company's focus has shifted to offering consultancy services and implementing its own data warehouse solutions. The data warehouse solutions offered by BI4U are based on in

house developed data models and use of third party data warehouse tools. Currently the company is primarily active in the healthcare sector. [BI07b]

Since last year BI4U is developing a so-called *intelligence factory*, previously known as *software factory*, which in fact is a working method, a collection of knowledge, methods, models and modules that support the partly automatic generation of a data warehouse. BI4U uses its intelligence factory for the data warehouse solutions it delivers to its customers.

For example: for its hospital customers BI4U has developed its ARIADNE architecture as part of the intelligence factory. During this project a generic business data warehouse model was defined, which can be used for multiple implementations at different hospital customers. Besides the business data warehouse model, various templates for database structures and packages were defined to be able to link the data warehouse with the various sources of a specific customer. This way these links do not have to be built from scratch for each new implementation.

Delta Impact Analysis concerns the analysis of changes in the source system situation regarding the impact on the packages and databases of a specific data warehouse implementation and/or the generic data warehouse model from the Intelligence Factory. The term “source system situation changes” is used, and not “source system changes”, because it encompasses both changes in specific source systems structures, but also changes in what the data in these source systems mean to the data warehouse. The Intelligence Factory will be discussed in more depth in Chapter 2.

1.2 Motivation

The motivation for this final project was the fact that BI4U suggested researching the topic of DIA. It is the intention of BI4U to include the DIA concept in its Intelligence Factory, in other words to have it be part of the working method of knowledge, methods and models so that it provides a generic method to perform DIA for its data warehouse solutions. This is difficult to do since it is just a concept, no clear instructions and guidelines are being used. Currently DIA is a concept that is embedded in the organization and employees of BI4U, it lacks explicitness, it can be seen as tacit knowledge.

Therefore the intention of the research is to make the concept of DIA explicit. Once this concept is made explicit it, the next step of this research is to relate the concept to literature, research and other companies in the field of data warehousing. Based on this knowledge and comparing the current situation with other insights, a final step of this research is to make a model that can support an effective and qualitative execution of this Delta Impact Analysis process. The design of such a model provides several advantages:

- It provides a structured method of tackling the difficulties that come with the changing nature of source systems.
- The advantage of making tacit knowledge explicit: this model is independent of knowledge residing in employees’ heads and therefore makes it possible to be adopted by others than just the current people involved.

- It provides the opportunity to measure and evaluate the current process of Delta Impact Analysis at BI4U, but also to structurally do so in the future.

In the next section the specific objectives of this research will be discussed more thoroughly.

1.3 Objectives

As explained in the previous paragraph the topic of this research project finds its origin at BI4U. At the same time this research is performed as a final research project for a MSc education of Business Information Technology at Twente University. Therefore the main objectives for this research are to perform research that is of a sufficient academic level and to provide an end result that is valuable to the company BI4U.

More concrete with regard to the research project topic, the objectives for the master thesis project are defined as:

- To make explicit, clarify and document, the DIA approach at BI4U.
- To study literature, research and best practices for insights regarding source systems situation change and DIA.
- Provide recommendations regarding possible improvements to the data warehouse environment at BI4U to better support changes in the source system situation
- The design of a model for the DIA approach that is useable in practice.
- The development of a testable 'prototype' of the model.
- Testing and evaluating the prototype in one or more real life applications(s).

1.4 Problem statement and research questions

To define the problem statement the methodology regarding problem solving by Wieringa and Heerkens is used. [HW06] The authors define *world problems* and *knowledge problems.* A world problem is usually something that is experienced. The problem owner experiences the problem due to the fact that there is a desired situation, a current situation, and there is a gap between these two. A knowledge problem is a problem where the problem owner is in need of information of the world around him. This differs from a world problem with regard to the fact that there is no desire to change a situation.

The problem statement for this research project can be considered as a world problem, and based on the specified objectives it can be defined as follows:

How can a Delta Impact Analysis model be designed that supports the process of analyzing the impact of changes in a data warehouse source system situation?

This world problem can be split in to two sub world problems:

- What is a suitable model for the process of Delta Impact Analysis, based on the current situation at BI4U, literature, research and practice of others in the field of data warehousing?

- Compared to the current ad-hoc situation, does impact analysis using the DIA model improve the result of the Delta Impact Analysis?

This world problem can be solved by answering several knowledge problems. These knowledge problems are formulated in the following research questions:

1) What is currently the essence and the concrete composition of the process of Delta Impact Analysis at BI4U?

2) What can be learned about change impact analysis methods regarding data warehouse / data integration systems by looking at literature, research and practice of others in the data warehouse field?

3) How can a model be developed to describe the process of Delta Impact Analysis that can be tested at BI4U?

4) How can the performance of the developed DIA model be evaluated in terms of efficiency and quality and what are the requirements to this performance?

5) Using the defined metrics to measure performance, what is the performance of the developed DIA model? And, if necessary, how should it be adjusted to be useable in practice?

Using the approach of Verschuren a research model for the research was developed. [Ver98] The model was revised multiple times during the development of the research questions, it is shown in Figure 2. Both the questions and the model were adjusted to fit each other, using the insight gathered during the process. The model can provide more insight regarding the relation of the research questions with the main goal of the research.

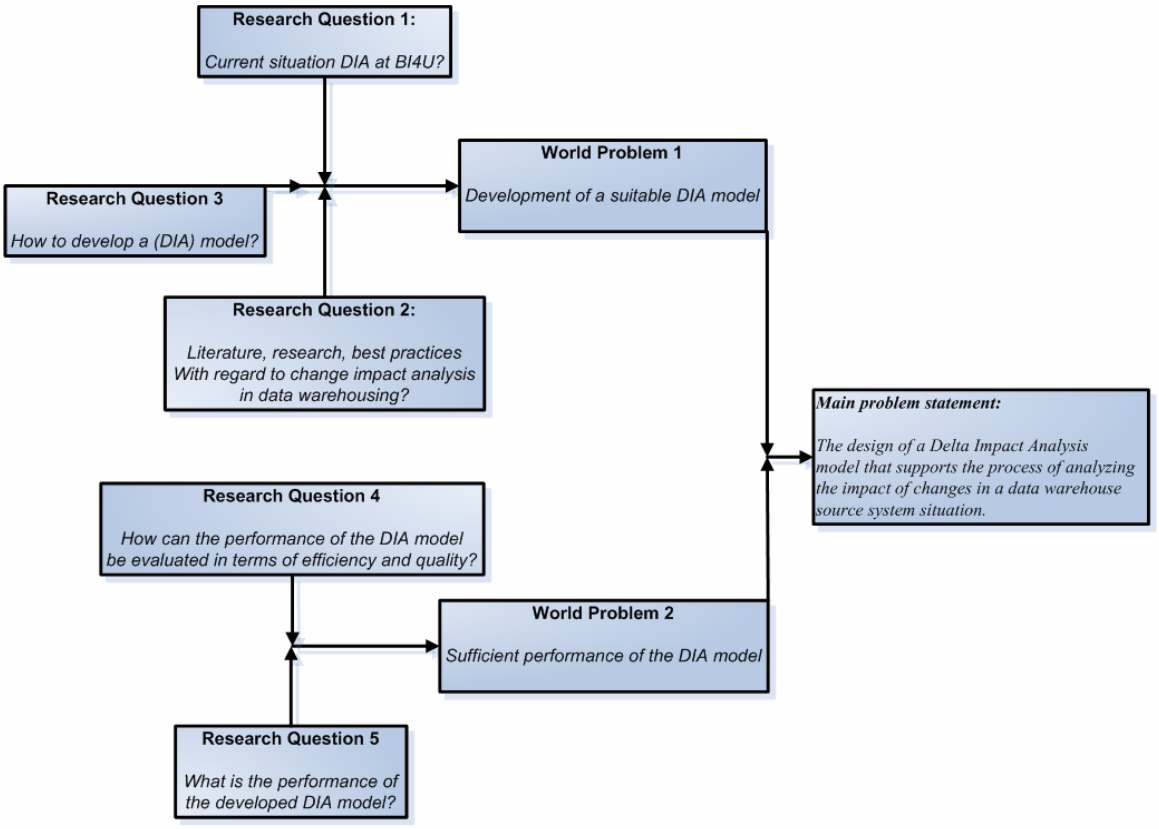


Figure 2. Research model

1.5 Scope

During the research the view on what DIA encompasses changed several times. However the fundamental concept has always been that DIA is triggered by changes regarding the data warehouse source systems. Therefore the scope of the research is not defined by the concepts of *impact* and *impact analysis*, but by the concept of *delta*: the changes that can require an impact analysis. Changes that fall inside the scope of this research are:

- Changes in the schema of data sources.
- Changes in the semantics of data sources.
- Changes in the business rules regarding data sources.

For this research it is assumed that source systems use a relational database to store their data. While in practice this can be different, this aspect is not considered. It can be assumed that this research can easily be applied to the perspective of other types of sources.

Activities that do not involve a Delta Impact Analysis are:

- Changes due to regular data updates in data sources.
- Changes in business rules that do not concern data from source systems.
- Changes due to the solving of mistakes (bugs) in the interpretation of source data.

1.6 Research framework

The Final Project Guide of the Department of Computer Science describes three types of research. [OF04] The first is a *research* project, which involves the formulating of hypotheses and the proving of these with theory. The second is a *design research*, which involves identifying a topic of concern, studying methodology to address this topic, apply this methodology to develop a specification and/or prototype that can be tested. The third is an *empirical research*, which also concerns the formatting of hypotheses, collecting and analyzing of data to prove the hypotheses. The *design research* is applicable to this research since it is the intention to create a specification of a model with the use of methodology.

Hevner et al. discuss their framework concerning a design research project in [HMP04]. The authors provide guidelines that assist researchers in performing a solid research, to evaluate this research and eventually to present this research. They state that IS research involves a mixture of people, organizations, and technology. The authors provide an IS research framework. The framework describes the aspects of Environment (people, organization, technology) and Knowledge Base (foundations, methodologies) which both influence the IS Research aspect (development of theories and building of artifacts, justification and evaluation). The IS Research eventually gives something back to both the Environment, e.g. a solution, and the Knowledge Base, e.g. new insights. Figure 3 provides a graphic interpretation.

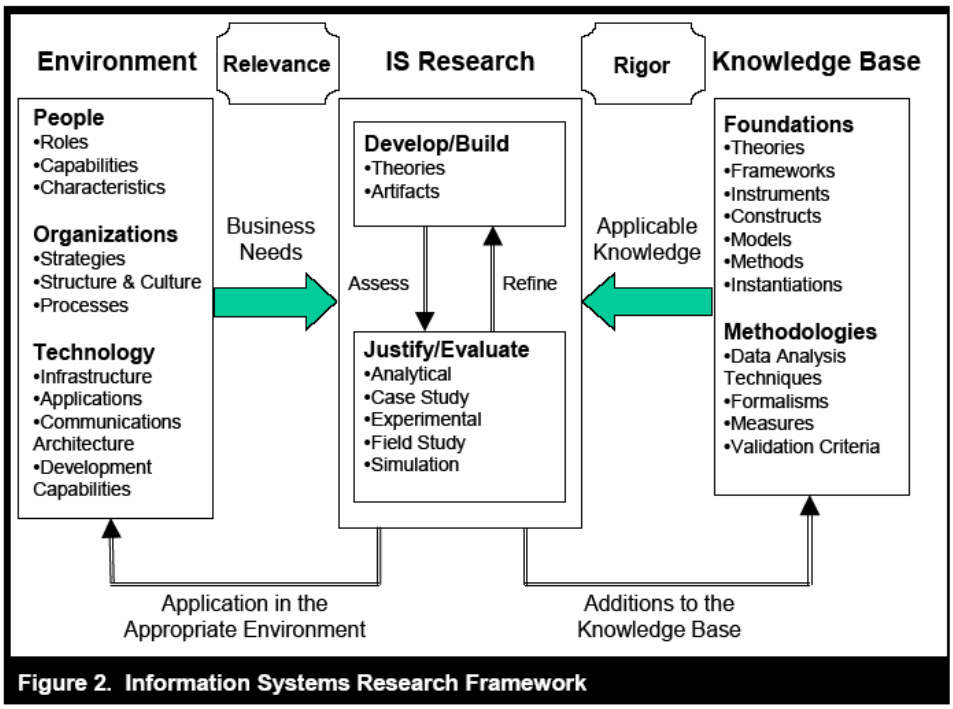


Figure 3. IS Research framework [HMP04]

Besides this framework the authors provide seven guidelines that one can use during a design-science research project. The guidelines should not be followed just to be followed; researchers and reviewers should use their own insight to determine in which cases the guidelines are relevant and when they should be deviated from.

The guidelines:

1. *Design as an artifact*: produce a construct, model, method, or implementation. This should not be independent from people or the organization, also a good perspective on and fit with the organization are crucial.
2. *Problem relevance*: a problem can be defined as the gap between a goal state and a current state of a system. Problem solving can be defined as a search process to reduce or eliminate the gap.
3. *Design Evaluation*: IT Artifacts can be evaluated in terms of functionality, completeness, consistency, accuracy, performance, reliability, usability, fit with the organization, and other quality aspects. A design artifact is complete and effective when it satisfies the requirements and constraints of the problem it was intended to solve.
4. *Research Contributions*: the ultimate assessment for research is to look at what the new and interesting contributions of the research are to the world. This can be one of three: The design artifact, novel foundations, and novel methodologies.
5. *Research Rigor*: it is necessary for all IS research paradigms to be both rigorous and relevant. Designed artifacts are often components of a human-machine problem-solving system. For such artifacts knowledge and behavioral theories and empirical work are necessary to construct and evaluate such artifacts. The main concern is to determine how well an artifact works, and not to theorize about or prove anything about why the artifact works.
6. *Design as a Search Process*: design is essentially a search process to discover an effective solution to a problem, it might not be the one and only or the best solution. Design-science simplifies a problem by decomposing the problem into sub-problems. This may not be realistic enough to have a significant impact on the actual practice, but it can be the start for future research and/or application.
7. *Communication of Research*: design-science research must be presented to both a technology-oriented and a management-oriented audience. This provides the ability for those who want to exploit the research and those who want to do further research, to do so.

The aspects of the framework can be noticed in the research statement and questions. The first research question relates to the Environment and Relevance aspect, the second research question relates to the Knowledge Base and Rigor aspect, the third research question relates to the element Develop/Build of the IS Research aspect, and the fourth research question relates to the element Justify/Evaluate of the IS Research aspect. With respect to the guidelines one can already notice the application of these guidelines in the objectives and research statement paragraphs of this proposal. The link between the research and the framework and its guidelines will be kept in mind during the entire execution of the research project.

2 Delta Impact Analysis at BI4U

The goal of this chapter is to provide an answer to the research question '*What is currently the essence and the concrete composition of the process of Delta Impact Analysis at BI4U?*'.

First the general concepts of data warehousing at BI4U are described. Next, the concepts of *Delta* and *Impact* are discussed separately. Finally the concept of *Impact Analysis* is discussed with respect to what was learnt from the concepts of *Delta* and *Impact*.

2.1 Data warehousing at BI4U

In this section the intelligence factory model of BI4U (Figure 4) will be explained. In [SS05] Sen and Sinha describe several possible architectures for a data warehouse: Data Mart, Enterprise, Hub-and-spoke Data Mart, Enterprise with operational data store, Distributed. The architecture at BI4U is of the type *Enterprise Data Warehouse Architecture.* In this type of architecture there is a central data warehouse on which data analysis is performed which is then stored in RDBMS and or Data Marts through which it can be used for reports. The data warehouse model of BI4U consists of several stages: Data Staging Area (DSA), Interface Staging Area (ISA), Business Data Warehouse (BDW), Data Mart Staging Area (DMSA), Cubes and Reports. In this document the stages will often be referred to as components, since we discuss impacts on the components that facilitate these stages. Data moves through all these components in order to provide a useable outcome to the end user. The different components will be discussed more thoroughly in the next paragraphs and subsections.

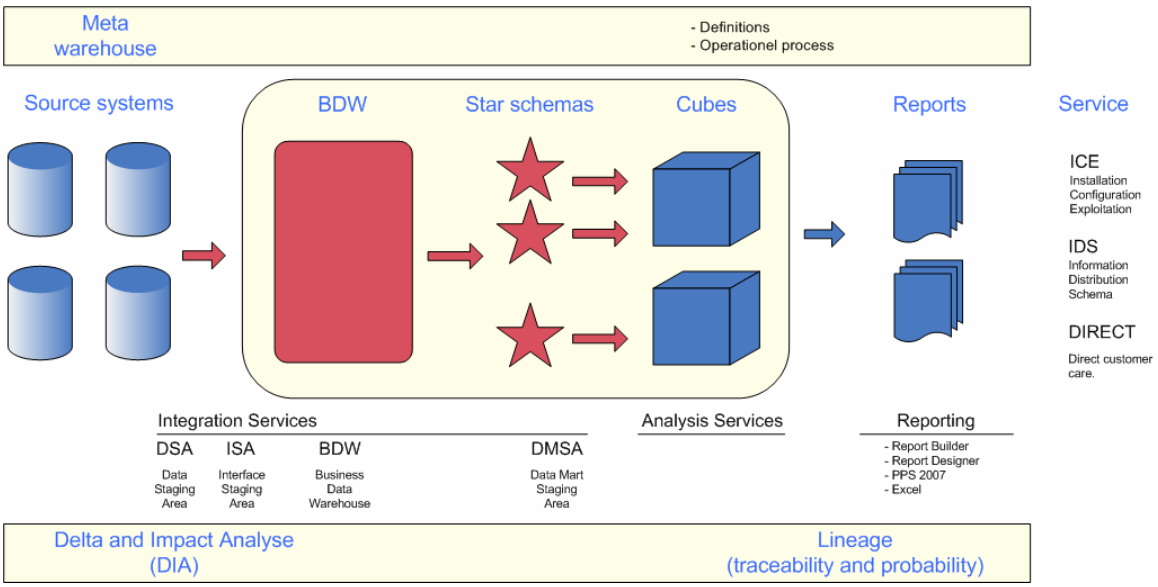


Figure 4. BI4U Intelligence Factory model [BI06b]

The first three steps (Source -> DSA -> ISA -> BDW) can be recognized as the process of Extract, Transform, Load (ETL), which is a common concept in the data warehousing

field. ETL processes involve the cleaning, transforming, combining, duplicating and structuring data in order that it can be used in a data warehouse environment for analysis. [MTK06] In general the term *wrapper* is used in data warehouse literature. The wrapper takes care of connecting a source system to the generic model of the data warehouse. [RKZ00] Relating this to the situation at BI4U this involves the DSA database and the Source-to-DSA package. [BI07a] Between the wrapper and the actual data warehouse something often described as a *middle layer* is present. The middle layer takes care of the filtering and integrating of information from multiple sources. [RKZ00] Relating this to the situation at BI4U this involves the ISA database and the DSA-to-ISA and ISA-to-BDW packages. [BI07a]

The other steps (BDW->DMSA->Cubes->Reports) involve the modification of the data from the data warehouse so that it is optimized for analysis and use for reports. BI4U develops generic business data models that can be used for multiple data warehouse implementations in similar organizations, the ARIADNE BDM (Section 1.1) is an example. This results in a generic structure of the ISA, BDW and DMSA. The Source-to-ISA packages, DSA database and DSA-to-ISA packages are customer specific to map the customer source systems to the generic ISA and BDW. [BI07a]

In the next subsections the various phases in the model will be explained in more detail in order to clarify the elements that can be part of the impacted components in a Delta Impact Analysis. The explanation will be done by describing the phases data will go through and it will be illustrated by using a much simplified example of a source system and a data warehouse. In this example the source system registers employees and information about their work-engagement and absence. The data warehouse in the example is intended to provide insight into the absence of employees.

2.1.1 Source to Data Staging Area

Regarding ETL this phase concerns the extracting part. The goal of the Data Staging Area is to extract data from the source systems with a minimal burden on the source systems. The DSA database is first emptied and then filled with data from the source system by Source-to-DSA packages. The data is extracted without modifying its structure or content. It could be that the entire source database is being copied to the DSA database, but it is also a possibility that irrelevant tables are not extracted. One could describe the database (both content and structure) as: DSA $\subseteq$ Source.

Figure 5 provides an example. In this case the data warehouse is only interested in employees and absence within the company. In this example, data such as the engagement history of employees is not a topic that the data warehouse is interested in.

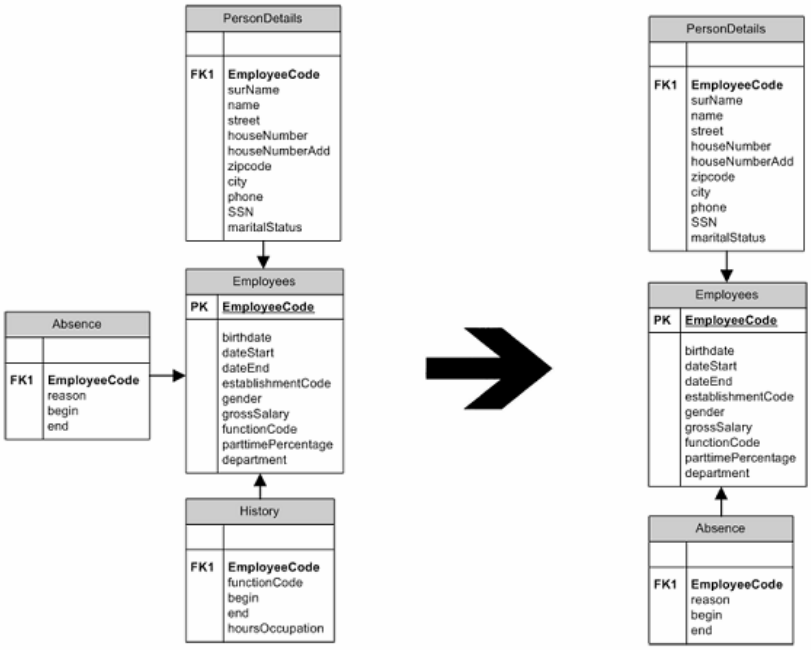


Figure 5. Example Source-to-DSA

2.1.2 Data Staging Area to Interface Staging Area

Regarding ETL this phase concerns the transformation part. The goal of the Interface Staging Area is to restructure the data so that it can be compared with the current data warehouse content and that new data can be inserted in the data warehouse. The ISA database is first emptied and then filled by DSA-to-ISA packages based on the data in the DSA database. The data is transformed to fit the structure of the defined business data model (BDM). The BDM structure is also used for the business data warehouse. As discussed previously, this generic BDM can be used for multiple customer implementations and therefore the structure, naming and type of attributes in the ISA can and probably will differ from the source system and the DSA. It can also result in certain data present in the DSA not being transferred to the ISA, and it can also mean that calculations on data from the DSA are performed and that only the result is being stored in the ISA.

Figure 6 provides an example, based on the example of the previous paragraph. As explained the structure can be different from the source system, which is also the case in this example. For instance the ISA structure has a department and an engagement table, however the data in these tables can be constructed from the data that is present in the source tables. The reason for this structure is that the goal of the data warehouse is to provide insight into absence, and one eventually wants to distinguish absence over different departments. Therefore the department is kept in a separate table and each absence entry can be traced back to a specific engagement in time and therefore to departments. Also it is not interesting to keep all specific data in the data warehouse such

as address and contact information about employees, nor is it required to distinguish between an employee's first and last name. Also naming of attributes slightly differs in some cases, since the data warehouse model is generic and can be used with multiple source systems without modifying it.

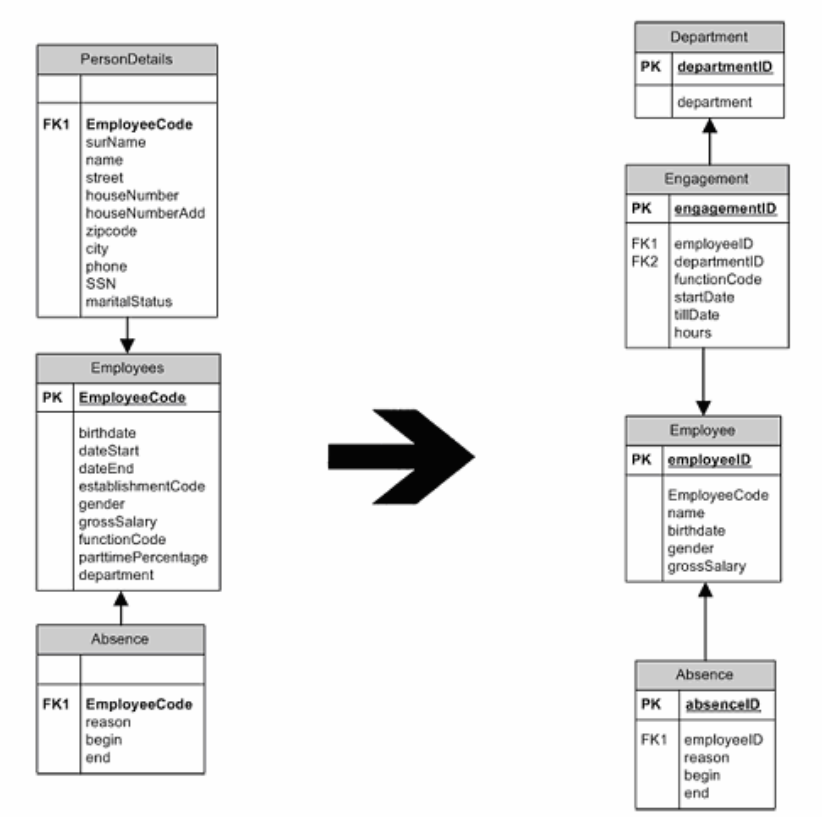


Figure 6. Example DSA-to-ISA

2.1.3 Interface Staging Area to Business Data Warehouse

Regarding ETL this phase concerns the loading part. The goal of the Business Data Warehouse is to store all relevant data that can be used for analysis based on the generic business data model of entities and relations. The BDW database is the core of the data warehouse. It contains all the data that was loaded into the data warehouse at earlier times and the BDW database is filled with new data by the ISA-to-BDW packages based on the data in the ISA database. The structure of the ISA and the BDW database is the same. There are no modifications made to the data. In cases where this is relevant, the data is compared regarding slowly changing dimensions and possibly enriched with time information. Slowly Changing Dimensions as explained in [MTK06] determine if, in the case of change of data compared to what is already in the data warehouse, a record of history is kept or if the old data is overwritten.

Figure 7 provides an example based on the example of the previous paragraph. The data structure of the ISA and BDW is identical.

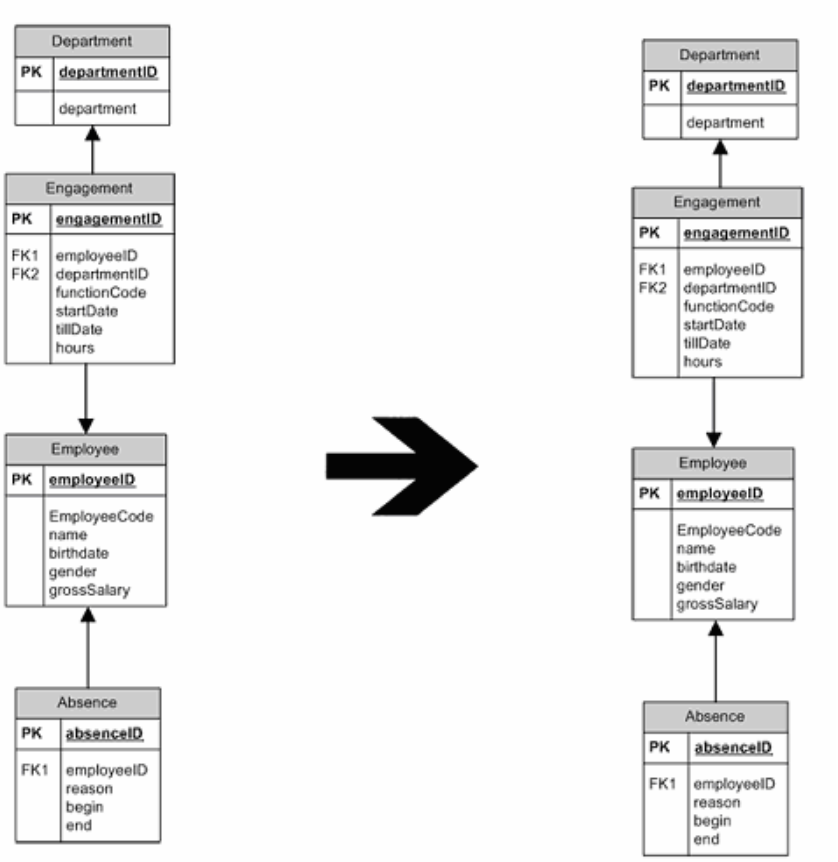


Figure 7. Example ISA-to-BDW

2.1.4 Business Data Warehouse to Data Mart Staging Area

In the BDW to DMSA (Data Mart Staging Area) phase the data is reorganized from a relational model to a dimensional model. The goal of the Data Mart Staging Area is to restructure the data to improve analytical performance. In the relational model of the BDW, analysis would require going trough many relations with join queries, which is intensive for the information needs of a data warehouse. In the DMSA data is related into star schemata with fact and dimension tables in order to structure the data for omptimal analysis performance. The data is structured in such a way that it is easier and faster to relate certain interesting information (facts) to certain dimensions.

Figure 8 provides an example based on the example of the previous paragraph. All data is now stored in such a way that it is related to a specific absence fact entry. Since it is interesting to be able to show reports on e.g. the basis of departments, sex and age classifications these are added as dimensions. During the transformation from BDW to DMSA for each absence entry the age and sex can be determined through the relevant employee and the department can be determined through the engagement of the employee.

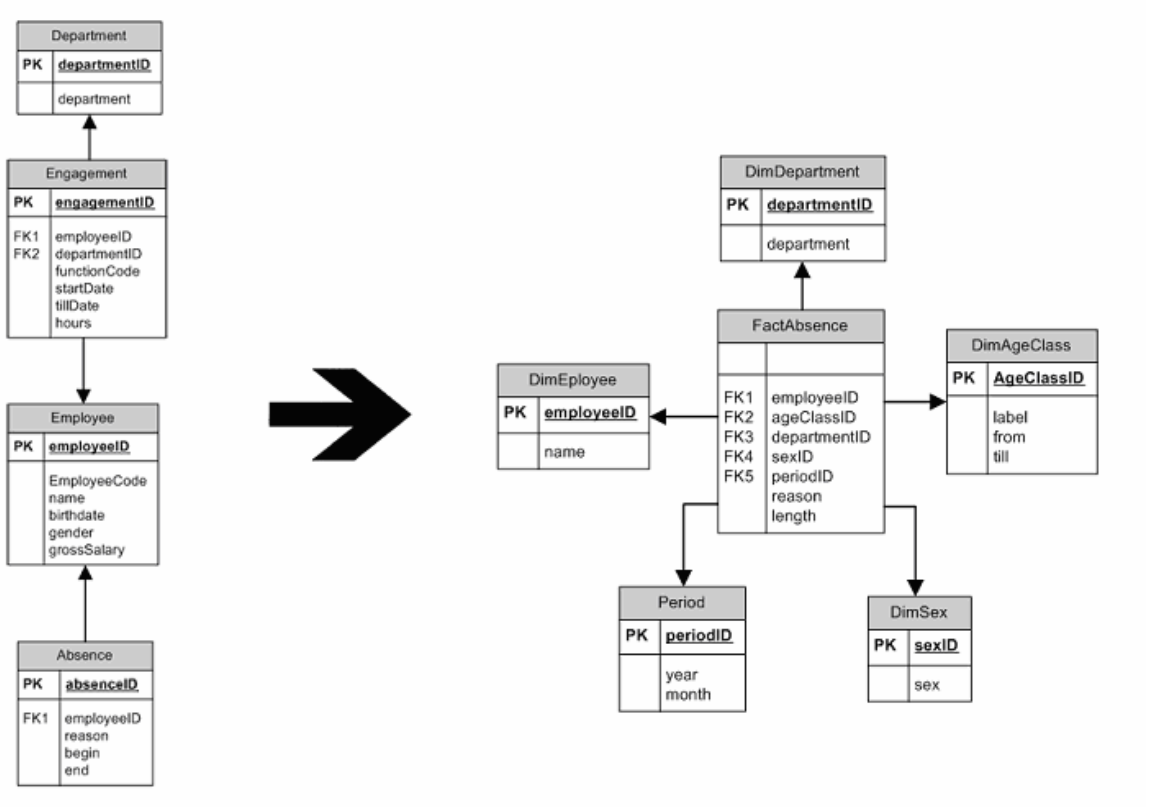


Figure 8. Example BDW-to-DMSA

2.1.5 Data Mart Staging Area to Cubes and reports

In the DMSA to Cube phase the data that has thus far been stored in relational databases is transformed and stored into a multidimensional database. During this transition much of the analysis that will eventually be used in reports is done. The data stored in the cubes can be used to display the desired information in the reports, or in fact the reports will query the cubes for the required information. In the report functionality, all that has to be done is to define which dimensions should be related to each other. Any change in the DMSA will have an impact on one or more cubes and reports.

A multidimensional cube exists of several elements: measures, dimensions and attributes. Measures are the object of the analysis and are derived from fact tables in the DMSA. Dimensions provide the context for measures and these are derived from dimension tables in the DMSA. Attributes define a dimension, and can form a hierarchy of the dimension. A hierarchy of the dimension time can for instance be made out of the attributes day, month, year. The name cube can be confusing as it suggests three dimensions are related to each other for a specific fact of the DMSA. For the purpose of analysis, the number of dimensions is not relevant, this can very well be 2 or 4 or more. The ‘cube’ is just a name for the multidimensional database that supports this analysis. Figure 9 provides a graphic example of the dimensions Period, Department and AgeClass being related to each other in a cube. [CD97]

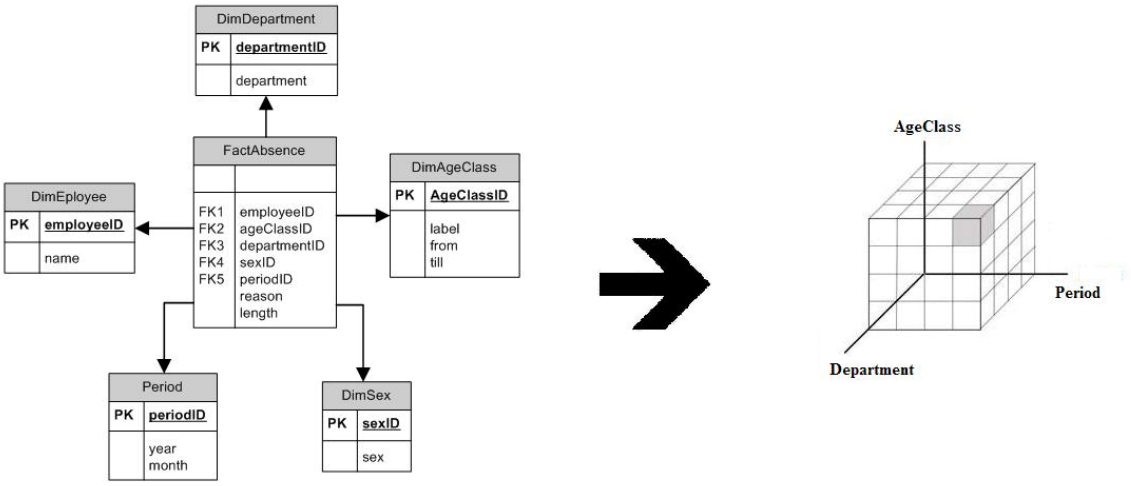


Figure 9. Example DMSA-to-Cubes

Harinarayan mentions three ways in which a cube can be implemented. The first option is to physically materialize the whole data cube. This delivers the best performance for the end-user. All possible calculations are performed and stored beforehand and when a report is requested, the pre-calculated values can be used. However pre-calculating everything is very computational expensive and requires a high amount of storage. The second option is no materialization. This delivers the worst performance to the end-user with regard to response time. However it is cheap with regard to computational expenses and no storage of pre-calculated information is required. The third and last option is to partly materialize the data cube. Only those parts of the cubes that are most important are materialized. This provides a more feasible solution to pre-calculating, since this way it is possible to balance end-user performance with the costs of computational calculations and storage of materialization. [HRU96]

What can be concluded is that there is a tradeoff between on the one hand the costs of computation and storage and on the other hand the response time for the end user when requesting reports and querying the data warehouse. BI4U uses a variant of this last option suggested by Harinarayan, by executing the most used reports every night and caching this data for use if reports are requested the next day.

2.2 Delta

According to the internal DIA document [BI06a] and conversations with Shirly Dijkstra - an information analyst at BI4U- and Marc van der Wielen – a consultant at BI4U-, the Delta in Delta Impact Analysis can have several causes. These causes will be discussed in this section, see Appendix A and Appendix B for summaries of the interviews.

BI4U distinguishes between *technical* Delta Impact Analysis and *functional* Delta Impact Analysis. Both require an analysis of the impact on the data warehouse, but they differ in the type of change. Technical DIA concerns changes in the technical structure of the source systems and Functional DIA concerns functional changes to the data warehouse.

Technical DIA can be required due to changes in the data structure of an existing source system, such as dropping/inserting/renaming of columns and tables and the modifications of primary and foreign keys. This can be caused by for instance minor updates or newly released versions of source systems.

The relation between source systems and the data warehouse is on a database level. This means that changes in the interface, functionality and code of the source system does not impact the data warehouse. The data warehouse is only influenced if the data is stored differently or if the semantics of this stored data changes. BI4U defined a matrix that shows which components of Figure 10 (Section 2.3) can be impacted in one way or another with which type of change. However this only provides high level insight into what impact a change can have on what component, it does not provide the ability to determine which components are actually impacted by a specific delta. The various types of (structural) database changes that can have an impact on the data warehouse as defined by BI4U are [BI06a]:

- Table added, deleted
- Column added, deleted
- Column modified: position, data type, length, obligatory
- Primary key added, deleted, modified
- Foreign key added, deleted, modified

Functional DIA can be required due to the following events:

- Changes in the semantics of data. For instance in the case of a limited set of codes that can be the value for a column. If this limited set changes this will also impact the data warehouse. Or for instance if the meaning of data in a specific column is changed
- Changes in the business rules. BI4U uses the term *business rule* for the rules that define the meaning of information elements in the business data warehouse and reports and how this is calculated from source data. Since the business rules define how certain data elements in the data warehouse are calculated from source data, a change in this can impact other components of the data warehouse and a DIA is required. E.g. in example 2.1, consider situation a where ISA.Employee.name is first constructed as [DSA.PersonDetails.surName + “, “ + DSA.PersonDetails.name] and this is changed to be constructed from just [DSA.PersonDetails.surName]. This is an example with little impact, but one can imagine a larger impact if the change concerns complex calculations and consistency regarding historic data is relevant.

2.3 Impact

In this section it will be discussed what impact a change in a data warehouse source system can have on the data warehouse. In Section 2.1 it was discussed that the data from source systems passes through various databases and transformation processes of the data warehouse before it is provided to the end user by the means of reports. Depending on the type of change any of these databases and processes can be impacted. Figure 10 provides a visual overview of the components on which a delta can have a direct or indirect impact. The big arrows show that a delta can have an impact on any of the data

warehouse companies. The lineated arrows shows the relations between the various databases and transformation processes, which imply a possible direct impact. A technical change in a source system will have a direct impact on the Source-to-DSA package, and changing that component will impact the DSA database, which will impact the DSA-to-ISA package, etc. A business rule change regarding the interpretation of source data will have a direct impact on the DSA-to-ISA package, changing that component will impact the ISA database, etc. Dependent on the characteristics of the delta and the choices made, the impact of a delta can continue all the way trough the reports.

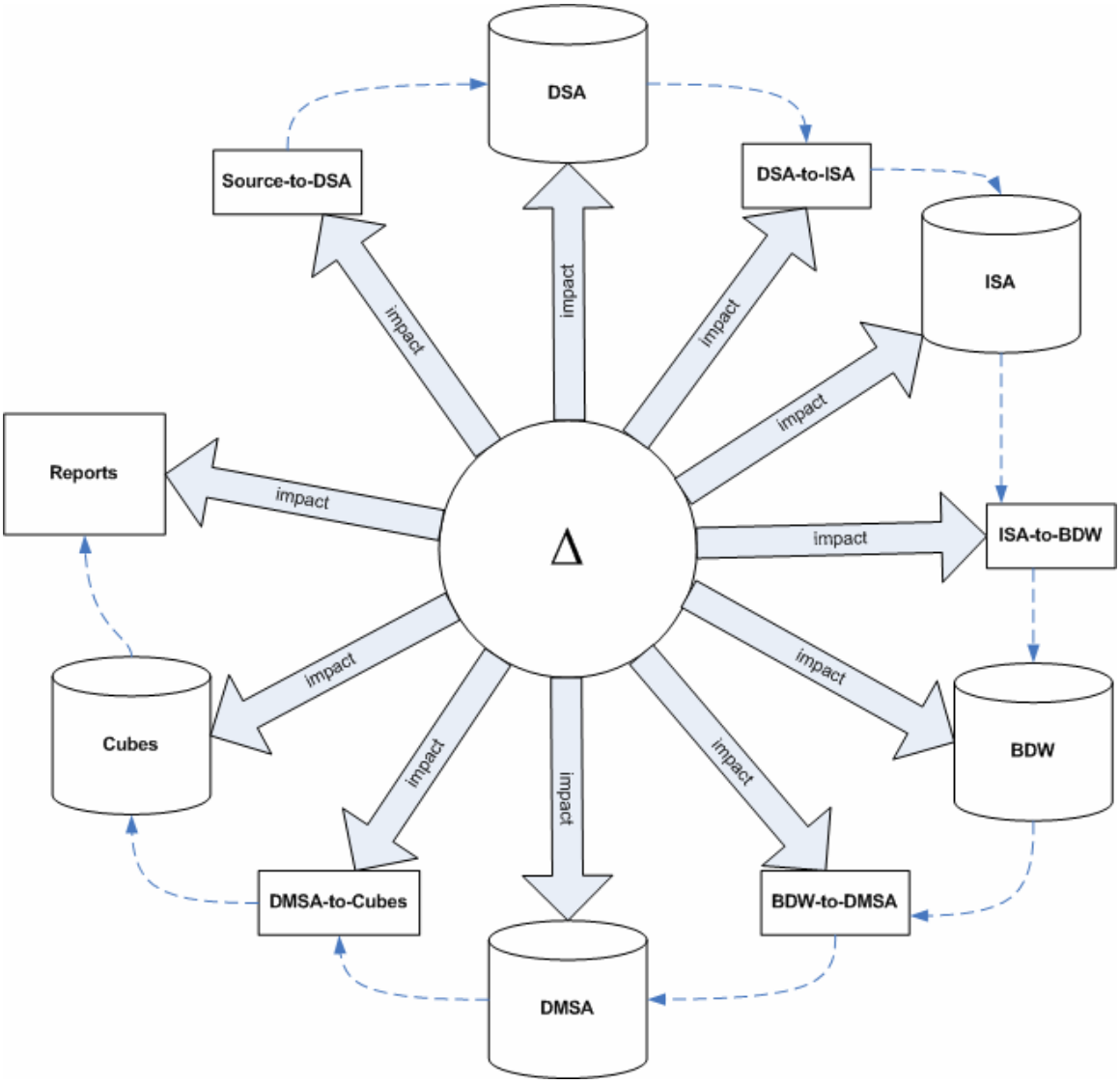


Figure 10. Indirect and direct impact of a delta visualized

The possible impact was discussed in the interview with Shirly Dijkstra (Appendix A) and Marc van der Wielen (Appendix B). Most changes will only impact the Source-to-DSA, DSA and DSA-to-ISA components. As explained these components are distinctive for each data warehouse implementation and try to map the specific source system to the generic ISA, BDW, DMSA and Cube structures. So with any change the goal is to try and map this to the generic part of the data warehouse without modifying the data warehouse. In some cases however it will be impossible to compensate for the changes by adjusting the non-generic components of the data warehouse. For instance if data is no longer provided by the source system or if new data is provided by the source system that can provide functionality that is not yet present in the data warehouse.

To provide some structure in the different types of impact a delta can have, three different types of impact scenario's are defined in this research and explained below: *intact*, *reduced*, *increased*. With *combined information supply* is meant the total set of information that is supplied by all the different data sources of the data warehouse.

Intact information supply
The first scenario is that the combined information supply from sources to the data warehouse remains *intact*. This will require no modifications beyond the DSA-to-ISA component.

Looking back at the examples from 2.1, the following example can be given for intact information supply: the tables PersonDetails and Employees are dropped and the table Emplyees_PersonDetails is inserted. Assuming that in fact this is only a merging of two tables without the removal of any columns, all that needs to be done is adjusting the components till the DSA-to-ISA packages so the same information is supplied to the data warehouse. No other components of the data warehouse will be influenced.

Reduced Information Supply
The second scenario is that the combined information supply from sources to the data warehouse is *reduced*. This can require a change to any component of the data warehouse, assuming information that was being used is no longer accessible.

Looking back at the examples from 2.1, the following example can be given for reduced information supply: if the table PersonDetails is dropped this does not have to influence the components beyond the DSA-to-ISA component, as long as all data required by the data warehouse (for reports) is available from somewhere else. For instance if some columns were moved to the Employees table, or if this information is available through a different source system. Otherwise the ISA, BDW and DMSA components and the components in between will have to be modified in such a way that it will work without name information of employees, since this information is no longer provided by the source systems. The data warehouse can still do absence analysis based on departments, sex and age classifications, but it will not be able to do so on specific employee names.

Increased Information Supply
The third scenario is that the combined information supply from sources to the data warehouse is *increased*. This can require a change to any component of the data warehouse, assuming that it would add functionality to the data warehouse

Looking back at the examples from 2.1, the following example can be given for reduced information supply: if a column Employees.managerID is added with a foreign key constraint to another employee this would not require any modifications beyond the DSA components. However it can be interesting to also be able to do absence analysis based on a specific manager. In that case it would require a change to all components of the data warehouse.

2.4 Impact Analysis

The actual analysis of the delta and its impact encompasses more than just determining a delta and it's impact. Figure 11 provides an activity diagram of the activities involved at BI4U with DIA, this activity diagram and its explanation are based on the internal DIA document [BI06a] and the interview with Shirly Dijkstra (Appendix A).

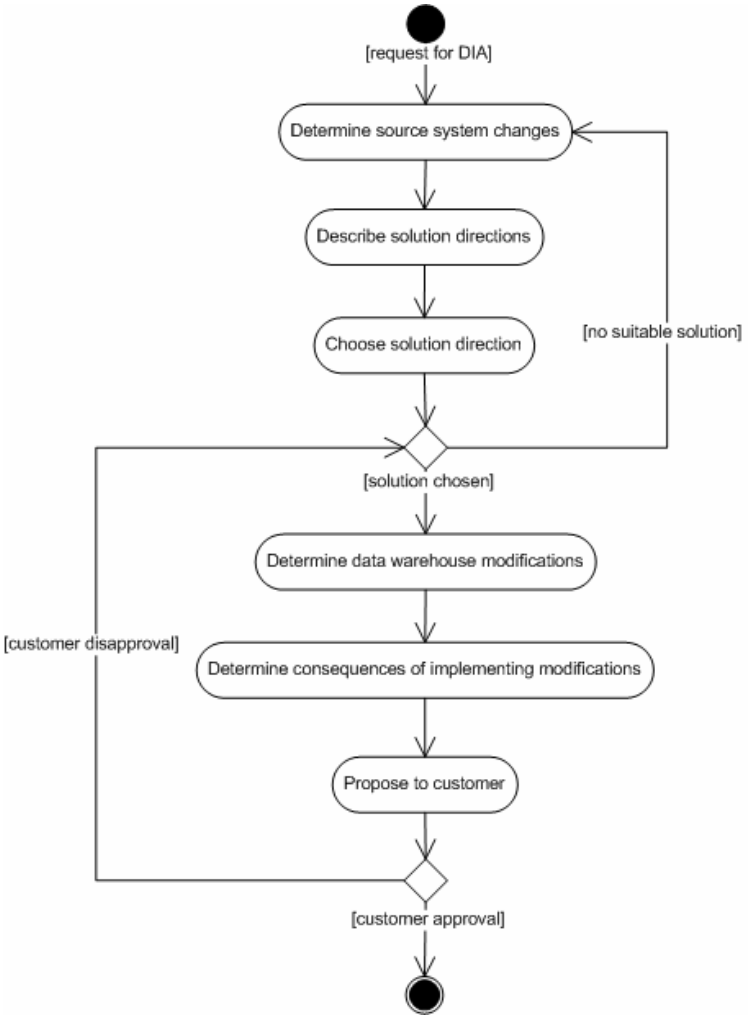


Figure 11. DIA at BI4U - current situation

At first the changes in the source system are described. This includes a description of the context of the change, the requirements of the customer, an overview of the concrete technical changes and the possible solution directions that can compensate for the changes. The changes are usually determined using version change logs or tools such as Red Gate SQL compare. Customer requirements are usually defined by using a structured approach, i.e. using the Zachman framework. In agreement with the customer the most suitable solution direction is chosen, or in case the customer is not satisfied, the previous activities are repeated using the customer feedback. After the solution direction is chosen, the required modifications to the data warehouse components to compensate for the changes are determined. In this step the complexity of each for the required modifications is also determined, based on personal insight. In the next activity the consequences of having to implement these modifications are determined. This concerns a description of the required activities, the planning of time required, the costs for the customer and the service to the customer. Finally the result of the Delta Impact Analysis is proposed to the customer and, if the customer approves, this is the end of the Delta Impact Analysis process and its result can be implemented.

What can be concluded from this information is that the current process of DIA has several shortcomings. First of all, while the steps do guide the employee in the process, most steps lack a clear guideline of how the tasks should be performed. A second shortcoming is the lack of clear criteria for the steps that are designed to function as a measuring point of the progress/completeness of the performed tasks. In the interview it is mentioned that the question “Can I, or someone else, work further with this?” is asked oneself to determine if the result of the activity is sufficient. Besides this, the described BI4U model does not seem to be used as a guideline in practice. The only impact analysis that has been done, if we can even call it IA, has been an, for in-house use, assessment of technical dependencies on certain to be changed columns of a source system.

2.5 Conclusion

In this chapter much has been discussed about how data warehouses work at BI4U. Two possible types of changes, technical and functional, in source systems situation are described. Three types of impact scenario’s are identified: intact information supply, reduced information supply and increased information supply. Also a description of the process on how this impact can be analyzed has been given.

Based on the observed situation of DIA at BI4U several remarks can be made.

In the current model specified by BI4U, the impact on the data warehouse is analyzed after the solution direction is chosen. This is strange since it is likely that one needs to study the implications to the data warehouse first in order to define clear solution directions. It is likely that in practice the implications will be studied in someway before chosing a solution direction, however this is not represented in their current documentation about the process.

With regard to defining the delta and its impact, BI4U only identified the types of structural changes and on which data warehouse components this can have an impact.

What is missing in their documentation are the possible scenario's that can occur that can cause these structural changes. In the desired situation the use of scenario's can improve the process of Delta Impact Analysis because it can provide a less abstract view to a delta, which can simplify the analysis of the impact.

Regarding the analysis it was concluded that a clear guideline of the tasks to be performed are missing and also clear criteria to test the performed tasks are missing. In the desired situation there are more clearly defined guidelines and criteria.

All changes (deltas) are researched manually and then modifications to the data warehouse implementation are performed manually. It might be desirable to have a tool to support the process of DIA with its guidelines, criteria and possible technical scenario's. the use of such a supporting tool can also possibly support (partly) automation of the adjusting of the data warehouse implementation.

3 Delta Impact Analysis in Data Warehouses

The goal of this chapter is to provide an answer to the research question '*What can be learned about change impact analysis methods regarding data warehouse / data integration systems by looking at literature, research and practice of others in the data warehouse field*'. To answer this question one needs to know more about what changes can occur, what their impact can be and how this impact analysis can be performed. The main focus of this chapter is to introduce and discuss theoretical concepts, which can then be related in order to propose a model for Delta Impact Analysis in the next chapter.

In the first section the general topic of data warehousing, changing source systems and the problems that these changes can cause are discussed. The other sections in this chapter can basically be allocated to three concepts: delta, impact and impact analysis. Because the eventual goal of the research is to apply impact analysis to the field of data warehousing, the topic of impact analysis in general is discussed first. Next, the concept of delta will be discussed in three subsections based on the type of delta: schema, semantic and business rule. The reason to split delta this way is that it became clear during the research that these three different types of causes for a delta existed. Finally, the possible impact to a data warehouse that a delta can cause is discussed.

3.1 Data warehouses and changing data sources

Extensive research has been performed on how data warehouses must be maintained. While data warehouses are designed to handle data changes in sources and the processing of this, often too little attention is paid to the changes in the structure of data sources. The structure (tables, columns, data types, foreign key relations, etc.) of a database is described and defined in a so-called database schema. As Chen et al. state in [CZR06], most of the data warehouse research assumes a static data warehouse schema, which is really not a valid assumption in an evolving environment. Sen and Sinha write in [SS05] about the fact how data warehouse solutions often eventually fail because they are too complex and too expensive to change them to fit the evolving needs of the business. These evolving needs include aspects such as end-user improvements, data warehouse schema changes and other factors. One other factor that could be identified of evolving needs are that of changing source systems.

Chen et al. discusses the topic of view maintenance regarding data integration systems using the global-as-view approach. Various mapping techniques have been developed to specify how data of one schema is transformed to the other. One of these techniques is that of a global-as view query. This means that there is one global schema that defines how source systems (schemata) are mapped to gather the required information. Another technique is that of local-as view queries, where for each source system it is defined as a view over the global schema. The authors discuss the problems that can occur with the evolution of the source systems with respect to the global schema. The global schema may become invalid once the source has evolved. During the evolution of the source systems, the mapping or view definition should be maintained to stay consistent and prevent it from becoming invalid. [CZR06]

Rundensteiner et al. discuss several issues that come with the changing nature of data warehouse data sources. [RKZ00] Data warehouse maintenance concerns keeping the data warehouse up-to-date with the data updates, it does not concern the maintenance of the structural changes in source systems. Data warehouse maintenance regarding data updates of information sources is a well-established field, but maintenance under a mixture of data and schema updates remains largely unexplored. They state that this can cause one main issue: maintenance queries submitted to the information source may find an altered schema due to a source schema change that no longer matches the submitted query format. Based on the discussion by Rundensteiner et al., three problems that are caused by the issue can be defined: inaccessibility, invalidity and inconsistency.

Inaccessibility can occur if no adaptation to the delta is made on the wrapper level, because then (parts of) sources will become inaccessible to the data warehouse. Regarding automated change detection algorithms that can detect and adjust the data warehouse processes, Rundensteiner et al. state that these are (computationally) expensive and intrusive and therefore infeasible to implement. They suggest the use of a change history log for each source system. [RKZ00]

Invalidity can occur if the delta is not adapted to on the definition level of the data warehouse. [RKZ00] It might be necessary to modify the definition (schema) of the data warehouse based on the changes in the information source. This can be realized by propagating the changes from the source to the data warehouse schema, but another possibility is that the data warehouse is isolated from the change. Also it is important to keep meta data regarding the origin of information in the data warehouse and also to be able to let the meta data evolve along with the source system evolution.

When preventing *invalidity* from occurring, the third problem of *inconsistency* rises at the data level. If the data warehouse is adapted to fit the changes, it is possible that data gathered in the past is now interpreted wrong (based on the new schema). [RKZ00]

3.2 Impact Analysis

In the previous section it has been discussed what problems can occur by changes in data warehouse source systems. It is important to be able to manage these changes in order to prevent these problems. In this section the topic of performing impact analysis is general will be discussed and this will also be related to data warehousing. First of all the essence of impact analysis will be discussed, after which certain activities of impact analysis are elaborated on.

3.2.1 Essence of Impact Analysis

The main difficulty regarding this research is the fact that little research is performed on the topic of impact analysis in data warehousing. Most research in the field of data warehousing assumes that source systems only evolve on a data content level. If it is even acknowledged that there is a possibility of evolving source systems that can impact the functioning of the data warehouse, the topic of how this impact can be assessed is never discussed thoroughly. Therefore the best option available is to look at impact analysis research in the field of software development. Lindvall noticed the similar lack of

attention for the topic of impact analysis in the field of software development, which is why he researched the topic. [Lin03] Lindvall established that there is sufficient literature out their regarding cost models to assess the costs of implementing change, however these models rely heavily on the outcome of the impact analysis process, on which little research had been done.

Lindvall and Sandahl define impact analysis (IA) as predicting the impact of a new requirement on a system before the requirement is actually implemented. [LS98] Relating this to DIA, it would concern predicting the impact of source changes on the data warehouse before the change occurs. Fasolino and Visaggio discuss the impact of changes in the requirements of software on other components that represent the software, e.g. different levels of models and the actual code. [FV99] Relating the model of Fasolino and Visaggio to DIA on a more abstract level, impact analysis concerns analyzing the impact of a change in a certain source element on elements in other components of the data warehouse, such as source-to-DSA extraction, DSA-to-ISA transformation and ISA-to-BDW loading, etc.

In [AB93] Arnold and Bohner provide a framework to develop an impact analysis method. They distinguish in what an impact analysis can be used for. The first aspect that IA can be used for is as the activity of identifying what has to be modified in order to accommodate a change. The second aspect that IA can be used for is to identify the possible consequences of a specific change. Lindvall discusses the topic of requirement driven impact analysis (RDIA) in software development. [Lin03] He identifies two main activities which both require different output from the IA process: determining change entities and cost estimation. For the first activity it is required to know what exactly changes and for the latter it is only required to know how much changes.

Arthur defined four main goals of impact analysis: determine the scope, develop accurate estimates for required resources, analyze cost and benefits of a change, and communicate the complexity of a change to others. He also states that performing impact analysis is a critical success factor in order to maintain software productively. Not or insufficiently performing impact analysis can result in low estimates, which leads to scheduling and cost problems. It will also lead to an increase in corrective maintenance. [Art88]

Summarizing what has been discussed so far, the main activities of an impact analysis in data warehousing are:

- Determining the changes
- Determining the impacted elements of the data warehouse
- Determining the required modifications to the data warehouse
- Determining the resources and costs required to perform the modifications

Besides performing these activities, the goal of impact analysis is to communicate the complexities involved with a change to others.

To perform impact analysis one needs to take into account several constraints. Lindvall named several constraints for performing (requirements driven) impact analysis in

[Lin03]. First of all it must be performed as accurately as possible, since costs estimations are based on the identified impact. The second constraint is that the costs of performing the impact analysis should not be more than the costs of implementing the change itself. The third constraint is that impact analysis is a planning activity and thus should not involve any alterations of source code. A fourth constraint is the fact that the process of the impact analysis should be sufficiently documented so it can function as a basis for subsequent phases (implementation, etc.). The fifth and final constraint is the fact that the impact analysis process must be evaluated afterwards in order to improve the process and its activities for future use.

Based on a survey by Lindvall it is also concluded that the main source of information with regard to IA for software development is that of knowledgeable developers, the second source is that of the analysis of the actual source code and documentation. [Lin03] Another source of information that Lindvall discusses is that of the original design models. Assuming they are kept up-to-date, models provide a clear abstract view on the system, but they lack information on a detailed level. Therefore it is not that interesting to use for dependency analysis, but it is useful while documenting the results of the impact analysis. [Lin03]

3.2.2 Determining impacted elements: traceability

Most authors in impact analysis research discuss or at least mention the concept of traceability. Traceability is defined by Lindvall and Sandahl as: "*the ability to trace from one abstract model to another*". [LS98] Arnold and Bohner define it as: "*the ability to determine what parts are related to what other parts according to specific relationships*" [AB93]. This concept of traceability is very relevant with regard to DIA, since the essence is about impact on related elements. Therefore this subsection is dedicated to the topic.

The person performing the impact analysis can use the existing packages, database and data warehouse models to look where the changed source elements are used in the data warehouse. The more complex the source changes and/or the use of these elements by the data warehouse, the more complex this work is. This will require a large amount of time and it is also error sensitive when a human is looking for the relations and dependencies on an ad-hoc basis. Fasolino and Visaggio state that manual impact analysis is too human intensive and therefore not possible to succeed. They suggest that automated IA is required in order for IA to be successful. [FV99] This is maybe a bit too black and white, but it would be more valuable if the dependencies and relationships are well documented in order to provide traceability. As Lindvall and Sandahl state in [LS98], if a system has a high level of traceability then it is easier to propagate a change from one model to another model.

The concept of traceability is introduced in two types by Fasolino and Visaggio: internal traceability and external traceability. [FV99] Internal traceability concerns relations between components in the same model, external traceability concerns relations between components of different representation models. In other research the same concepts are sometimes described using other names, Lindvall for instance speaks of vertical and

horizontal traceability in [LS98]. Vertical traceability is the same as internal traceability and horizontal traceability is the same as external traceability.

In order to realise a high level of traceability for a system it is essential to: update all models of the system consistently, to provide vertical traceability and to provide horizontal traceability. [LS98] Relating this to data warehousing, we can consider all the different components and phases in the data warehouse as a different model. So vertical traceability is realized by describing the relationships between elements of the components, e.g. how are columns in the DSA related to other columns in the DSA. Horizontal traceability is realized by describing the relationships between elements in different components of the data warehouse, e.g. how is a certain ISA column related to columns in the DSA. Figure 12 provides a graphic example of the difference between vertical (left) and horizontal (right) traceability. Using vertical traceability information one can see there is a relation between DSA.Employee.EmployeeCode and DSA.PersonDetails.EmployeeCode and DSA.Absence.EmployeeCode. Using horizontal traceability information once can see there is a relation between DSA.PersonDetails.surName and ISA.Employee.name.

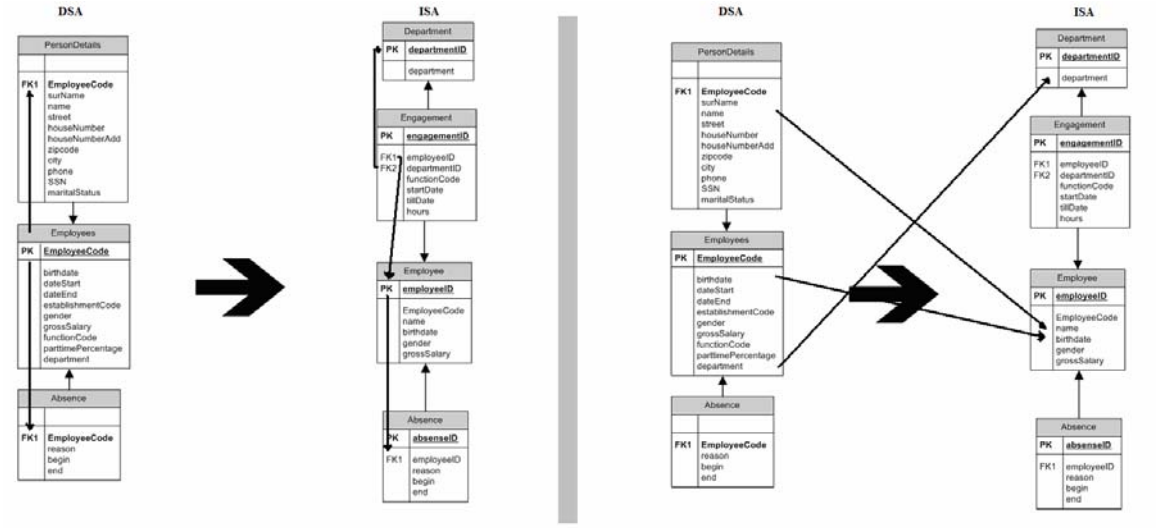


Figure 12. Vertical vs. Horizontal Traceability

Regarding traceability to support impact analysis Lock et al. add to the above that impact analysis without the support of a tool is not feasible in case of most systems. A tool can assist in both the creating, updating and analysis of the traceability information. Where documentation would only provide a long list of traceability relationships, a tool could also visualize the traceability. The topic of traceability visualization is more extensively discussed in [LOC99].

One way to keep an updated model of the data warehouse is that of a meta warehouse in which all relations are specified. Since a meta warehouse is a way to realise traceability and it is not truly a part of the Delta Impact Analysis process, it will not be discussed further in this section. However it does provide a better fundament for DIA and

determining the impacted elements in the DIA process, therefore the topic of meta warehouses is discussed in more depth in Section 6.2.2.

3.2.3 Determining costs and resources

Sneed states that one of the main reasons to perform an impact analysis besides determining the impact on system components is to determine the costs of dealing with this impact. [Sne01] Sneed claims that since change is initiated by the customer, impact analysis is crucial in order to determine the costs of these customizations. In order to estimate the costs of the impact Sneed names four steps:

- Identify the impact domain
- Size the impact domain according to standard metrics
- Adjust the size by the complexity of the impacted system
- Convert the size to required effort using a productivity table that is based on impact analysis and their implementations of the past.

Lethbridge and Laganière defined several principles for cost estimation in software development projects. [LL01] Not all principles will be mentioned here but what can be learned from the principles with regard to cost estimation for DIA will be discussed. First of all it is important to split the cost estimation of a project in to parts, estimate these costs, and eventually add these up. Second it is important to recognise that there is more work than just coding; take into account time required for prototyping, design, inspection, testing, debugging, documenting and deployment at the customer. A third aspect that can be learned is that one should look at past experiences, by using personal judgment and/or algorithmic models one can base (part of) the estimation on past experiences. However, do take into account that the projects will be different on certain aspects such as: people, development processes and skills, customers, demands, technology. A fourth point of attention is to not be naïve, do not be too positive; estimate for three scenario's: best case, worst case and realistic. Afterwards an average can be calculated.

Regarding planning Lethbridge and Laganière state that two aspects need to be distinguished. [LL01] First of all there is the aspect of a time period, which is about the intended begin and end date of the implementation phase. Second there is the development effort, which is about how much time is actively needed to perform certain activities, this should be defined in person-months or person-days. Regarding cost estimation Lethbridge and Laganière suggest that a development effort can be converted to money by multiplying it with a defined weighted average costs per person-month or person-day. [LL01] These weighted average costs are based on salary costs of developer employees, but also all other resources required (by an employee) to perform the work.

3.2.4 Evaluation of Impact Analysis

Lindvall proposes that by comparing that what was predicted to be impacted with what was actually impacted one can evaluate the impact analysis. [Lin97] Arnold and Bohner discuss the possibility of comparing the predicted impact, as they call it the *estimated impact set* (EIS), with the eventual outcome, the *actual impact set* (AIS). [AB93] They state that the goal of any impact analysis model, is to estimate the consequences of certain changes and that the deviation between the predicted impact and the actual impact

is as small as possible, ideally zero. [AB93] Fasolino and Visaggio distinguish two sets that the EIS is composed of, namely the *primary impact set* (PIS) and *secondary impact set* (SIS). The primary impact set is the set of objects that is directly affected by the change and can be easily identified. The PIS is the set that is analyzed in order to identify those objects that are part of the secondary impact set. Together they form the *estimated impact set*. [FV99]

Lindvall and Sandahl also mention that is possible to evaluate the impact analysis after the change is eventually processed. One can compare the predicted outcome of the impact analysis with the actual outcome to determine its effectiveness. Based on this comparison different components of the system can be grouped in the following groups [LS98]:

- Non-impacted components that were predicted to be non-impacted
- Impacted components that were predicted to be impacted
- Non-impacted components that were predicted to be impacted
- Impacted components that were predicted to be non-impacted

The first two groups can be considered as correct estimates. The third group can be related to the concept of false positives (wrongly identified), the fourth group to the concept of false negatives (missed). The authors do not discuss this, but with respect to Delta Impact Analysis it can be difficult to identify components of the third group. This is due to the fact that these are likely to be modified in order to prevent complications after they are analyzed as to-be-impacted, which makes it difficult to determine afterwards if they would not have been impacted without the modification.

Lindvall defined that a basic constraint for good impact analysis is evaluation after performing impact analysis in order to improve the process and its activities for future use. [Lin03] Also evaluation of the actual resources and work effort that were required to implement a change can be documented along with the estimates from the impact analysis. This information can then be used in future impact analyses to base cost estimation of projects that have similarities.

During the planning phase in the beginning of this research project, the initial research question concerned the evaluation of the to-be-developed DIA model on the basis of quality and efficiency. While efficiency can be easily measured by looking at the time required to perform the Delta Impact Analysis, quality is a less concrete concept. Fasolino and Visaggio introduce two concepts with regard to determining the quality of an impact analysis: adequacy and effectiveness. [FV99] The two concepts will be discussed in remains of this section.

Adequacy

Adequacy can be used to determine how well the impact analysis determines the potentially impacted components. The defined indicator by Fasolino and Visaggio is that of Inclusiveness. Unfortunately they limited their definition of this indicator to the value of 1 and 0, where the value is 1 if AIS $\subseteq$ EIS. [FV99] This means that a change with a very large impact, but missing only one of the impacted objects in the IA, will result in the value 0 for inclusiveness. While identifying a large number of incorrect impact

elements will still result in the value 1 for inclusiveness. If the four groups defined by Lindvall and Sandahl, as discussed earlier in this section, are related to adequacy, it can be determined that what one would really like to know to evaluate the adequacy is:

- $AIS \cap EIS$ → the correctly identified impacted objects (correct)
- $(AIS \Delta EIS) \cap EIS$ → the incorrectly identified impacted objects (wrong)
- $(AIS \Delta EIS) \cap AIS$ → the impacted objects that were not identified (missed)

Fasolino and Visaggio did take the size of the impact analysis into account when determining the effectiveness (see next paragraph), but this is also lacking in their determination of the adequacy. [FV99] Taking this into account with regard to adequacy, three indicators are proposed in this research to replace *inclusiveness* in order to determine adequacy:

- Correct-Inclusiveness rate: $|(AIS \cap EIS)| / |EIS|$
- Wrong-Inclusiveness rate: $|((AIS \Delta EIS) \cap EIS)| / |EIS|$
- Missed-Inclusiveness rate: $|((AIS \Delta EIS) \cap AIS)| / |EIS|$

Obviously the correct-inclusiveness should be equal to or approach 1, where as the wrong-inclusiveness and the missed-inclusiveness should be equal to or approach 0. A higher wrong-inclusiveness rate will result in unnecessary labour to investigate the predicted impacted object and therefore less efficiency. A higher missed-inclusiveness rate is more dangerous and can result in components not being adjusted that should have been adjusted, which can result in malfunctioning of the data warehouse.

Effectiveness

Effectiveness can be used to determine if the output of the impact analysis process adds value to the person that has to maintain the system and process the change. With regard to DIA the system is the data warehouse implementation. Fasolino and Vissagio defined three indicators: *ripple-sensitivity*, *sharpness* and *adherence*. [FV99]

Ripple-sensitivity regards how extensive the relative impact of the primary impact set is. This is expressed by using amplification ratio, which relates the secondary impact set to the primary impact set. Sharpness concerns the degree in which the estimated impact does not impact too much of the system It can be expressed using a change rate, which relates the EIS to the size of the entire system. Adherence is used to indicate the degree in which the EIS differs from the AIS. The S-ratio is used to express this. The three formulas that define the discussed indicators are as follows: [FV99]

- Amplification ratio: $|SIS| / |PIS|$
- Change rate: $|EIS| / |System|$
- S-Ratio: $|AIS| / |EIS|$

While the discussed indicators are used by Fasolino and Visaggio to determine the effectiveness of an impact analysis approach, one can doubt the usefulness of the S-ratio. The S-ratio does not take into account the adequacy indicators that were defined in the

previous paragraph. For instance let's assume |AIS| = |EIS|, this gives an optimal S-ratio of 1, but what if half of the EIS was wrong and have of the AIS was missed? The adequacy indicators together provide the same insight as the S-Ratio, and more. The amplification rate and the change rate however are useful to assess effectiveness.

Besides these rates, another way to determine the effectiveness of a performed impact analysis is proposed in this research. One can ask the person that is responsible to deal with the change if the process has added value for him or her. In order to define certain indicators to evaluate the effectiveness of an impact analysis approach one can look back at what the initial goals of the impact analysis were. The person using the outcome of the impact analysis can score the outcome on these aspects. In Section 3.2.1 four goals were stated: determine the scope, develop accurate estimates for required resources, analyze cost and benefits of a change, and communicate the complexity of a change to others. Table 1 is a suggestion for the aspects for which a score can be given with regard to the four goals of IA. Appendix C contains a list of survey questions based on the scoring aspects, these can be used to score the outcome of a Delta Impact Analysis.

Table 1. Scoring aspects DIA effectiveness

Goal	**Scoring aspect**
Scope	Detail of the scope Completeness of the scope
Required resources	Detail of determined required resources Completeness of determined required resources Accuracy of determined required resources
Cost/Benefit analysis	Ability to perform cost/benefit analysis with the provided information
Communicate complexity of a change	Comprehension of the delta and its impact

3.3 Schema Delta

Now the topic of impact analysis has been discussed, the topic of *delta* will be discussed in this section (*schema*) and the following two sections (*semantic* and *business rule*).

The goal of this section is to provide a clear overview of the different changes that can occur in the schema of data warehouse source. On the conceptual level the concepts are usually explained through graphic representation and linguistics, on the logical level concepts are described in a way that they can be interpreted by a computer. [BCN91] Since DIA is mostly about preventing source changes from obstructing or corrupting the extraction and interpretation of data by the data warehouse, discussing it on the logical level is more appropriate.

In this section we will first discuss the different types of source schema change, next we will relate this to the concepts of database refactoring and impact on the total information

supply in order to provide some insight into what the different effects can be of different types of changes in sources.

3.3.1 Types of schema change

Chen et al. define three types of change that require view maintenance. The first is *data updates*, which are non-structural, those that a data warehouse is designed for to process and analyze. The second is *data-preserving*, which does not impact the data logically, i.e. renames and normalizations The third type is *non-data-preserving*, which does not preserve the old data, i.e. deletion of columns. [CZR06] Batini et al. have a similar classification for schema transformations. They do not consider data updates. They classify *information-preserving transformations*, which can be compared to the data-preserving changes of Chen et al. And they classify *information-changing transformations*, which can be compared to non-data preserving changes of Chen et al. They also provide three sub classifications for information-changing transformation: augmenting transformations (information content increases), reducing transformations (information content decreases), non-comparable transformations (other). [BCN91] This type of classification can be related to the classification regarding the information supply that was made in 2.3: intact information supply, increased information supply and reduced information supply.

Since Batini et al. provide a clearer overview of different information-changing transformations, their denomination is more relevant to Delta Impact Analysis than that of Chen et al. Also since information can be considered as the context in which data is interpreted, the term ‘information’ seems more suitable than ‘data’ with regard to the fact if it is preserved or not.

Ambler and Sadalage wrote a book about *Refactoring Databases.*[AS06] The refactoring of a database involves changes to the database that are of a limited extent and that do not impact the semantics. The authors state that it is a primary technique for agile developers. It concerns, as the authors label it, the evolutionary improvement of a database, based on new user requirements or software evolution, but without breaking the functionality. Both the informational semantics -the meaning of the information for the interpreter- and the behavioral semantics –the representation of the data- should be maintained. This is similar to the goal of Delta Impact Analysis for Data Warehouses. One big difference however is the fact that refactorings are discussed by Ambler and Sadalage as modifications and optimizations to the data structure of a system that do not influence the semantics. [AS06] Whereas with DIA it is about what components of the data warehouse are impacted by these modifications and how the data warehouse should be adjusted to cope with this, the semantics of the data warehouse can very well change. The fact remains that the goal of database refactoring is similar to that of DIA, coping with evolution (of the source). Therefore the topic provides a good insight into which types of changes can be executed regarding a database. This list adds several new scenario’s with regard to schema delta compared to those that are documented at BI4U.

Ambler and Sadalage describe many types of refactorings: structural, data quality, referential integrity, architectural and method. [AS06] Structural refactorings concern the

changes in the structure of the schema and are therefore also interesting regarding schema delta. Data quality refactorings concern changes to improve the quality of the information of the database, these changes do not influence the structure or relations of the schema and are therefore not interesting regarding schema delta. Referential integrity refactorings concern the relations between entities in the schema and therefore are interesting with respect to schema delta. Architectural and method refactorings concern changes to the interaction between the program and its database, assuming that the link with the data warehouse is on the data level and not by the use of methods etc. this topic is not interesting with respect to DIA.

In the next sections the various applicable refactorings of Ambler and Sadalage [AS06] will be reorganized based on the information-preserving and information-changing classification by Batini et al. [BCN91]

3.3.2 Information-preserving delta

Relating the insights of view maintenance to database refactoring, we can consider the scenario's listed in Table 2 as information-preserving events. [AS06] [BCN91] Since no data is lost, one can conclude that even though the way the data is stored might be changed, it is still possible to restructure it to the old situation. From this it can be concluded that the scenario's of this kind of change can be processed by adjusting only the ETL processes of a data warehouse.

Table 2. Schema information-preserving events

Merge table	This could be seen as the creating and dropping of columns in separate tables. It would be more valuable to recognize this as the merging of tables. If one would interpret it as just the dropping and creating of columns, the meaning of the creating columns has to be interpreted/defined, new mappings have to be defined, and possibly the link with historic data is lost.
Move column	With regard to DIA this situation is similar to the merging of tables. When it is interpreted as the dropping and creating of tables, valuable information is lost. When it can be identified as a move, this can be preserved.
Rename column	Similar to as explained above. Interpreting a rename instead of just the dropping and creating of a column preserves information.
Rename table	Similar to as explained above. Interpreting a rename instead of just the dropping and creating of a table preserves information.
Replace large object with table	Similar to as explained above. Basically the same information is now stored in a different place the dropping and creating of a table preserves information.
Replace column	Though labeled 'replace column' by Amlber and Sadalage, basically this can be interpreted as the changing of one or more aspects of a column, such as the type, the length, the position or the name. Recognizing this scenario can reduce the required modifications to the data warehouse largely. Since the format of the data that was stored in the past might have changed, it is important to also modify the stored data in the data warehouse so that it can be recognized as being the same.
Replace one-to-many with associative table	A foreign key column is dropped and a new table is created. Interpreting this as a change in how a relation is constructed reduces the required modifications to the data warehouse

Split table	This can be seen as the dropping of columns in one table and the creating of a new table with columns. However recognizing the event of a table split preserves information as with the other scenario's described in this table
Introduce surrogate key	In this case natural information was previously used as a key, which implies redundancy. The redundancy can be removed by introducing a surrogate key. In practice this means the dropping of several (redundant) columns and a relation and the creating of a new column and relation. Identifying this scenario as the removal of several columns that form a natural key and replacing it with a surrogate key can reduce the required changes to the data warehouse.
Replace surrogate key with natural key	In practice this means the dropping of one column and its relation(s) and the creating of new columns and a relation. Identifying this scenario as the removal of a surrogate key and replacing it with several columns that form a natural key can reduce the required changes to the data warehouse.
Drop view	In case the ETL processes of the data warehouse are extracting data from a specific view that is being dropped, it is useful to recognize this is the dropping of a view. If it is identified as a dropped view, one can determine from which tables/columns the view was constructed and the ETL processes can be adjusted to get the data from there
Rename view	This could be interpreted as the removing of a view and creating of a new view. Interpreting this as a renaming of the view would limit the amount of adjustments that have to be made to the data warehouse to fit the new situation

3.3.3 Information-changing delta

Ambler and Sadalage consider all refactorings as changes that do not influence the semantics. In some cases this is realized by removing references in application code to removed aspects of the schema, thus by reducing the functionality of the application using the database. [AS06] Relating this to data warehousing and the types of change, this can be considered as information-changing refactorings. Table 3 provides an overview of the structural events that are information-changing

Table 3. Schema information-changing events

Merge column	This could be seen as the dropping of two or more columns and the creating of one new column. However it would be more valuable if a merging of columns can be identified. This way old mappings can be adjusted, instead of deleted and the creation of a new mapping. Also it can be decided if historic data is adjusted to fit the new situation, or to maintain a historic link in another way
Drop column	A column is dropped and not replaced by another one. If the data warehouse uses this column for source data the data warehouse will need to be adjusted to no longer try and get data from this column and also what it should do with the historic data
Drop table	A table is dropped and not replaced by another one. If the data warehouse uses this column for source data the data warehouse will need to be adjusted to no longer try and get data from this table and also what it should do with the historic data
Split column	Interpreting this as the dropping of a column and creating of new columns makes it impossible to inherit information about the meaning of the original column. Recognizing the splitting of a column can be valuable regarding the link with historic data, how this link should be maintained can vary.
Introduce column	A new column is created. This could provide the data warehouse with an opportunity to gather more useable data. The priority to process such a change is not high, since the data warehouse will have no problems to function without being modified to the change.

Introduce table	A new table is created. This could provide the data warehouse with an opportunity to gather more useable data. The priority to process such a change is not high, since the data warehouse will have no problems to function without being modified to the change.

The ‘merge column’ event is considered as an information-changing event where information is *reduced*. Even though the combined value of two separate columns is the same as the value of the resulting column of a merge, information regarding their separation is lost: the information supply is reduced. Take for example the merging of a PersonDetails.street and PersonDetails.housenumber column into a PersonDetails.address column. In some cases a streetname might contain numbers (“Boulevard 1945”), so after the merge it is no longer possible to just split the address back into a streetname and a housenumber. The ‘split column’ is the opposite of ‘merge column’ and using the same logic as with a ‘merge column’ one can say that it is an information-changing restructuring where the information supply is *increased*.

Processing changes of the ‘drop’ events in Table 3 are similar to information-preserving events: modifying the ETL processes of the data warehouse will be sufficient to keep the data warehouse functioning correctly. However, even though the data warehouse will function correctly, certain dimensions of the data warehouse, data mart, cube and report components will no longer be filled with new information: the information supply is reduced. Since the main goal of data warehousing is to provide specific reports, it is very interesting, if not crucial, to also determine the impact of which information will be missing from these components in the future. Thus the information-changing event can have an impact on all components as visualized in Figure 10 (Section 2.3).

The ‘introduce’ events in Table 3 require no complex modifications to keep the data warehouse functional. Possibly a small modification in the SRC-to-DSA packages is required, to ignore the introduced elements. However the data warehouse can become more ‘intelligent’ if other components, such as the data warehouse model, the data marts, cubes and reports of the data warehouse are adjusted to interpret newly added tables and columns. The information supply is increased by these events.

3.4 Semantic Delta

In the previous section the topic of schema delta has been discussed. But what if the schema of the database remains intact, but the use of the schema changes; what if there is a *semantic delta*? In some cases a semantic delta concerns the representation of the data and it only requires a technical change to the transformation packages, but in other cases it concerns the meaning of the data and it is also required to change the actual business rule and its implementation in the transformation packages.

Ambler and Sadalage distinguish between informational semantics and behavioral semantics. [AS06] Informational semantics concerns the true meaning of data stored in a column to a user. Where behavioral semantics concerns the way in which this true information is represented. No specific literature concerning the types of semantic changes in data sources of data warehouses was found, therefore the distinction made by Ambler and Sadalage is used to describe several possible types of semantic changes. This

distinction is useful if we want to relate semantic delta to the impact on the information supply. Analogous to schema delta we distinguish between an information-preserving and an information-changing semantic delta.

3.4.1 Information-preserving delta

Changes in behavioral semantics, the representation of the data, are changes that do not influence information supply to the data warehouse if they are correctly identified and managed. The following examples are semantic changes that preserve the information supply:

- A date is first stored in a string as DD-MM-YYYY and this is changed and stored as MM-DD-YYYY. While the information stored is the same, it is now represented in another way.
- Assume a column that is of the type enumeration with the following possible values: A, B, C. These values typically each represent something, e.g. respectively situation 1, 2 and 3. If the used enumeration values are changed to D, E, F , but the meaning that they represent does not. All ‘A’ values become ‘D’ and still represent ‘scenario 1’, all ‘B’ values become ‘E’ and still represent ‘scenario 2’, etc.

3.4.2 Information-changing delta

Changes in informational semantics, the meaning of the data, are changes that influence the information supply. Similar to schema delta a semantic delta can *reduce* or *increase* the information supply. Another possibility however, is that it just changes the element in meaning, in this case the information supply is *redefined*.

To provide a clear picture, a distinction is made between what the change represents for the source system using the data: application or contents. Is it a change in the *application* of the data, so is it going to represent something else than before? Or is it a change in the *contents* of data, so does it still represent the same thing, but is it stored in a different way? The following examples are semantic changes that change the information supply.:

- **Application:**
  - A money value is first stored in EUR and this is changed to be stored in USD. The schema remains the same, also the way in which values are stored are the same (e.g. floating point), but the meaning of the information that is stored, is changed.
  - A money value first represents an amount before tax and this is now changed to represent a value after tax
  - A column ‘name’ that is used to store the surname but this is changed to now store the first name.
- **Contents:**
  - Assume a column that is of the type enumeration with the following possible values: A, B, C. These values each represent something, e.g. respectively situation 1, 2 and 3. If the used enumeration values are expanded to: A, B, C, D. The data warehouse does not know that now a possible value is ‘D’ that represents a new scenario 4.

- An orderdate was first stored in a column of the type string as an unix timestamp (seconds past since 1/1/1970) and is now stored as a string in the format “DD-MM-YYYY”.

When rationalizing, those examples representing a change in the *application* of the data are of the information-changing *redefining* type. Those examples representing a change in the *content* of the data, concern either an information-changing delta of the *increasing* or *reducing* type. The addition of ‘D’ to the enumeration can be considered as an *increasing* information-changing delta. The storing of a string date in day-month-year instead of a timestamp in seconds can be considered as a *reducing* information-changing delta.

3.5 Business Rule Delta

Embury et al. define a business rule as that what defines or constrains a certain aspect of an organization. [EWD06] Marco and Jennings define it as “*The logic applied to calculate or otherwise derive a business-related value*” [MJ04]. Marco distinguishes business rules based on what they are used to define: processes or data. Business rules for a process describe the definition and the quality rules. [Mar07] An example would be that a column ‘order_total’ is defined as a numeric value representing the total price of an order after tax in Euros and is calculated as ‘((item_quantity * item_price) + shipping – discount) * tax_rate)’. Business rules regarding data describe the definition and the domain. [Mar07] An example would be that item_price is a numeric value representing a product price before tax in Euros.

In data warehousing the business rules are basically that what define the meaning of the information in the data warehouse and that what is shown in the reports. On a higher level a business rule can define certain business concepts. On a lower level business rules are expressed in the data warehouse by the packages that transform the data between the various data warehouse components. In the DSA-to-ISA packages business rules are expressed to define how elements in the BDW are composed of source data. In the BDW-to-DMSA packages business rules are expressed to define the different facts and dimensions from the data in the BDW. These representations can be very complex and involve many tables, columns, tuples, calculations and constraints.

Unlike with schema and semantic delta a business rule delta always is an information changing delta, since it always concerns the meaning of data in the data warehouse at all of the components. Depending on the fact if a business rule is changed, deleted or added the information supply can be considered to be *redefined, reduced* or *increased.*

Figure 13 visualizes how the DSA-to-ISA transformation packages implement sets of business rules. The set of business rules of BR3 for instance could (pseudo description) look like this:

- X = Join DSA.Employees and DSA.PersonDetails on EmployeeCode,
- ISA.Employee.EmployeeId = hash(X.EmployeeCode, X.name)
- ISA.EmployeeCode = X.EmployeeCode
- ISA.Employee.name = [X.surName] + “, “ + [X.name]

- ISA.Employee.birthdate = Str_To_Unixtime(X.birthdate, “DDMMYYY”)
- ISA.gender = if(X.gender=1){m}elseif(X.gender=0){f}
- ISA.grossSalary = X.GrossSalary

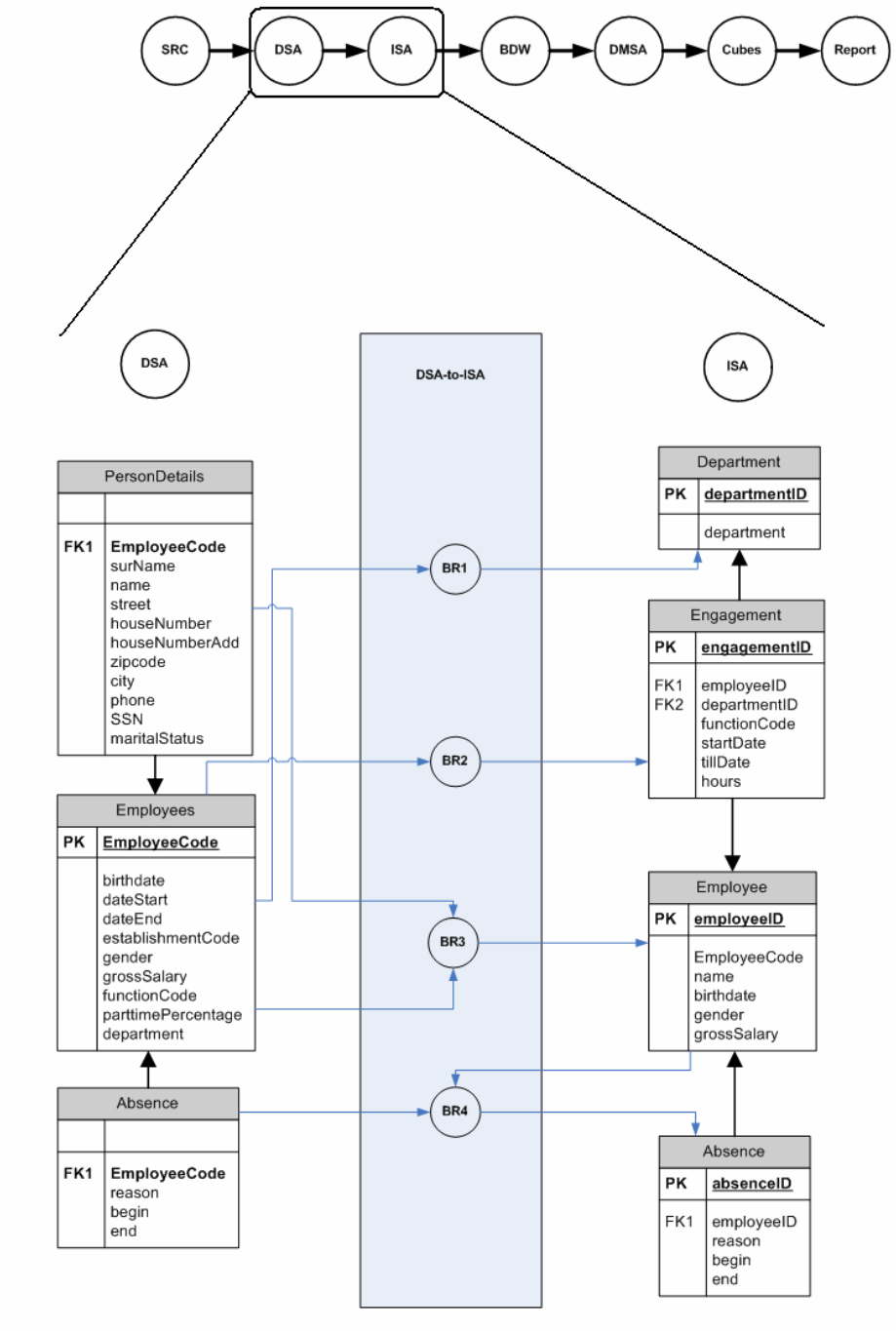


Figure 13 Business Rules DSA-to-ISA

The business rules are defined during the initial information analysis that is performed before a data warehouse implementation. During the implementation process, these

business rules can possibly be adjusted when new insights arise. After the implementation it is possible that new developments, such as customer demands or new insights, lead to the desire to redefine certain business rules. But also both a *schema delta* and a *semantic delta* can lead to the redefinition of a business rule if they are of the information-changing classification.

Imagine a situation where [ISA.Employee.name] was first defined by the business rule [X.surName] + ", " + [X.name] and this is changed to just [X.surName]. This change in a business rule can be the result of a *schema delta* classified as reduced information supply, in other words, the source column PersonDetails.name is removed by the vendor of the source system. The change of the business rule can also be due to the fact that the end user no longer considers a name to have to include the first name. While this is a simple example of a changing business rule and different possible causes for the change, in practice it can be far more complex.

An example of a business rule change in the BDW->DMSA stage can be the business rule that defines what the end-user considers as absence in the reports. For instance let's assume that at first absence was considered to represent any absence entry in the source system, but that this is changed to no longer include absence entries with the reason 'maternity-leave'. In addition to this the end-user could still request to introduce a new fact in the DMSA and an accompanying cube and report that specifically provide analysis regarding maternity leave. Due to the scope of the research, the role of business rules in Delta Impact Analysis is limited to those rules that concern the transformations of source data to the data warehouse. Therefore changes in business rules implemented in packages other than the DSA-to-ISA stage are not considered as a delta for DIA, however it is likely that changes in these business rules can be tackled in a similar way.

3.6 Impact

In the previous sections of this chapter a distinction was made between schema, semantic and business rule delta. For each delta it was discussed what the possible impact on the information supply from the data source to the data warehouse can be: *preserving* or *changing*. With regard to the actual impact on the data warehouse the same distinction between information-preserving and information-changing will be maintained.

The delta has an impact because certain elements are dependent on the elements that are changed. Therefore first the topic of relations that create these dependencies will be discussed. After this the possible impact scenarios caused by both information-preserving delta and information-changing delta will be discussed.

3.6.1 Relation between source schema and data warehouse model

In order to determine the possible impacts of a source change it is interesting to discuss the relation between the source system and the data warehouse. The ETL processes are the processes in between the original source data and the eventual data in the data warehouse. The transformations on the data during these processes is what defines the relation between the one end and the other. Therefore in order to determine the actual

relationship it is relevant to discuss the possible transformations that can occur to the data.

Source systems are mapped to the data warehouse in so called mapping expressions. These mapping expressions define which data elements of a data warehouse sources are extracted and what transformation are performed on this data. A mapping can exist in several forms [PER03]:

- An attribute of a table is copied as is
- Several attributes of one or more tables are extracted and a calculation is performed
- One or several attributes from multiple tuples are extracted and a calculation is performed
- External variable, such as a version or timestamp

These mappings are implemented in the ETL packages. Relations between elements in the different data warehouse components can be deducted from these packages. However, the packages are created to transform the data and not to describe the relations, therefore they are not structured in a way to support easy identification of relationships between elements. Regarding DIA it is useful if these relations are also specified in documentation by providing a definition of all mappings. Looking at literature there is no specific method that is suggested to use to document this.

For this research work of Marotta and Ruggia was considered. Marotta and Ruggia discusses 14 different types of transformations that can be performed on source data to retrieve the desired data for the use in the data warehouse. [MR02] [Mar00] They reference to previous work of Marotta in which these transformations were more extensively discussed. Classifying transformations according to their work can be insightful in clarifying what certain data undergoes on its way from source to its destination in a data warehouse. For instance it could help in determining the complexity of certain relationships, by scaling the transformation types on a scale of complexity. However determining the classification of a transformation can be abstract and classifying every transformation in a data warehouse from source to report is so labor intensive, that one can doubt if this work will ever be rewarding. Therefore it will not be discussed any further.

Fan and Poulovassilis discuss how relations between different schema's can be modeled by AutoMed. AutoMed, as the authors explain, is an integration and transformation system for heterogeneous data sources that can deal with multiple types of data models. The system relies on a limited set of primitives that can be used to describe a schema and also how one schema is transformed into another. The possibility to describe how one schema can be incrementally transformed into another, and another, is useable to express schema evolution. [FP03] [FP04]

In essence AutoMed makes it possible to describe how any schema can be transformed to another schema by the means of primitive transformations. AutoMed uses a low level hypergraph-based data model (HDM) that consists of nodes, edges and constraints. These

nodes, edges and constraints can be used to define a construct in a model. Fan and Poulovassilis define *Attribute (Att)* and *Relation (Rel)* constructs for a model that represents a relational schema and *Fact, Dimension (Dim), Attribute (Att)* and *Hierarchy* constructs for a model that represents a multidimensional schema. They also define a set of primitive transformations that can be applied to the constructs to describe the transformation from one schema to another: *Add, Delete, Extend, Contract* and *Rename*. *Add* and *Delete* transformations introduce or remove a construct of the schema based on other constructs that are already present in the schema. *Extend* and *Contract* transformations introduce or remove a construct that is partly or entirely-not based on other constructs that are present in the schema. *Rename* transformations involve the renaming of a schema construct. The primitive transformations applied to the constructs result in transformations such as *addRel, addAttr, delRel, delAtt,* etc. Using these transformations it is possible to define a schema and how it is transformed into another schema. [FP03]

As Fan and Poulovassilis explain the change from schema S to S_{new} can be described in a transformation pathway 'S -> $S_{new'}$. This change can be propagated by prefixing the reversed pathway (S_{new} -> $S_{)}$ to the transformation sequence. Putting this in the perspective of data warehouse transformations, a change in a schema SRC to SRC_{new} can be propagated to the current data warehouse situation by reversing this pathway and prefixing this to the sequence of transformation pathways as 'SRC_{new} -> SRC -> DSA -> ISA -> BDW -> DMSA -> Cubes'. This defined pathway now reflects how data is transferred from the updated source and transformed trough the data warehouse. When the independent transformations of the 'SRC -> $SRC_{new'}$ transformation pathway are considered, there are three possible scenario's on how to propagate these changes dependent on the type of transformation: [FP03]

- It concerns an *add*, *delete* or *rename* transformation. Related to this research, these transformations preserve the information supply, since add and delete transformations introduce or remove constructs that are based on other constructs in the schema. No other changes are required further in the transformation pathway sequence.
- It concerns a *contract* transformation. This means that a construct is removed that is partly or entirely not based on another construct in the schema. The reversed pathway 'SRC_{new} -> SRC' will be an *extend* transformation on the new construct, however this requires that it is possible to gather the information from somewhere, which is unlikely if it is removed from a source. The transformation pathway sequence needs to be modified downstream. All constructs based on the construct of the transformation need to be examined and modified or removed . Related to this research this transformation changes the information supply, it reduces it.
- If it concerns an *extend* transformation. The reversed pathway 'SRC_{new} -> SRC' will be a *contract* transformation on the new construct, which will make the transformation pathway sequence consistent and will not require any more modifications downstream the transformation pathway sequence. However, it might be interesting to make the data warehouse more intelligent by removing the newly added contract transformation at the beginning of the sequence of

transformation pathways and propagating this downstream the sequence to make the it consistent again. Related to this research this transformation changes the information supply, it increases it.

As an illustration, assume that to the source system of the example in 2.1 a new column ‘ethnicity’ is introduced to the PersonDetails column. This would imply that in ‘SRC -> SRC_{new}’ an *extend* transformation is performed. This can be propagated to the data warehouse by adding ‘SRC_{new} -> SRC’ , with a *contract* transformation, to the transformation pathway sequence. The transformation pathway sequence is now consistent, but the increase in information supply is not utilized. In order to utilize this, the *contract* transformation should be omitted, and the DSA->ISA pathway needs to be expanded with an *add* transformation for the ethnicity attribute in the Employee relation and so on for the other pathways in the sequence.

3.6.2 Information-preserving Impact

In case of a delta that is information-preserving there is only one problem that needs to be tackled, specifically that what was earlier defined as *inaccessibility*.

Since both *delta* and *impact* are split in two ways (*preserving, changing*), it makes sense to also distinguish the components of the data warehouse as present at BI4U in to two groups, with the business data warehouse model separating the two. With respect to the first group of components, this are the Source-to-DSA packages, the DSA database and the DSA-to-ISA packages. In common data warehouse terminology these are the Extract and Transform processes, the first two aspects of ETL. The second group is that of the ISA-to-BDW packages, the BDW database, the BDW-to-DMSA packages, the DMSA database, the cubes, and the reports. We will refer to this second group as the core of the data warehouse.

The practical reason to split of the components of these two groups is the fact that we can relate this to the distinction in information-preserving and information-changing delta. An information-preserving delta will only have impact on the first group of components, where as an information-changing delta can impact components of both groups.

So basically schema and semantic delta that are information-preserving will only require modifications to the first group of data warehouse components to prevent inaccessibility. The modifications will have to take care of the fact that the information loaded in to the ISA after the delta, is equal to the information loaded in to the ISA before the delta.

3.6.3 Information-changing Impact

In any scenario of a delta that is information-changing the problem of *inaccessibility* needs to be tackled, and therefore such an event will have an impact on components of the first group of the data warehouse. What the impact and required modifications are for the second group of components, the core of the data warehouse, depends on the type of event and the choices that are made. Dependent on the choice the problems of *invalidity* and *inconsistency* also need to be tackled.

A *schema* or *semantic* delta that is information-changing in a reducing or increasing manner, can impact components of the second groups of the data warehouse, depending on the fact if the changes are propagated to the data warehouse or not. If changes are not propagated to the data warehouse, only components from the first group are impacted. A reason not to propagate information increasing changes, can be that there is no desire to utilize the increase. A reason not to propagate information reducing changes, can be that the information that is reduced was not utilized anyway E.g. there is no need to propagate a reducing change of a column from unixtime to a DDMMYYYY string, if the data warehouse only extracts the month and year from the information.

A *semantic* or *business rule* delta that is information redefining, changes the meaning of data in the data warehouse and therefore implies impact on all data warehouse components. Just as with a schema delta, if the business rule delta reduces or increases the information supply one can choose to propagate these changes to the data warehouse or not, this will determine if the impact concerns only the first or also the second group of components.

In case of *reduced* information supply, historic data in the data warehouse will not be equal to that of new data that is inserted in the data warehouse. Based on what was discussed by Fan and Poulovassilis in the previous section, several options can be distinguished to create a fit between the old and the new situation:

- Not modifying the core data warehouse. In some cases this implies that certain information is left empty in new entries in the data warehouse. In some cases it might be that the reduced information was not utilised anyway, and there are no implications.
- Modify the core data warehouse and historic entries of the database to fit the new situation. In case of reduced information supply, this implies removal of the elements from the core data warehouse. This means that information of historic data is reduced in order to fit the new situation.

In case of *increased* information supply historic information in the data warehouse will not be equal to that of new data that is inserted in the data warehouse, but this is not necessarily a problem because contrary to reduced information supply creating a fit will not result in a loss of information. The options in case of increased information supply are:

- Not modifying the core data warehouse, which means that nothing is done with the added information supply and that the intelligence of the data warehouse will remain the same.
- Modify the core data warehouse to support the added information supply. Historic entries will have an empty value for the added parts of the model.
- Modify the core data warehouse to support the added information supply and calculate historic values for the elements that are added to the data warehouse model.

In case of *redefined* information supply there is no choice about what to do. Since no distinction can be made between different definitions at different times the core data warehouse and all historic values of the data warehouse need to be calculated according to the new definition.

Ideally it would not be a requirement to recalculate historic values. Calculating these values can possibly be very complicated or even impossible, or maybe it is undesirable because historic information should be according to an old definition. In this case it would be important to know which content of the data warehouse is based on the old definition and which data is based on the new definition. Basically the user wants to be able to know where information in the reports came from and how aggregations and calculations were made based on what business rule at what time, Fan and Poulovassilis call this concept 'lineage'. [FP04] This would require some sort of support for data warehouse versioning. This support for versioning also provides a cleaner solution when it is desired to keep historic information in case of reducing changes: new data for reduced elements will not have a zero value, but they will not exist in the new version. The same applies to information increasing changes where no historic values are calculated: the values will not be zero for these elements, they just do not exist in the old version of the data warehouse. Whether evolution and versioning is supported or not does not matter much for the process of DIA. However it does provide support for DIA and reduces the possible required modifications that are determined in the DIA process, therefore the topic of data warehouse versioning is discussed in more depth in Section 6.2.1.

Whether the calculating of historic values is a necessity, due to lack of evolution support, or a choice, does not really matter to this research. What does matter is providing some insight into the work required when taking historic information into consideration. Golfarelli specifies several actions that can be required to calculate the information for historic entries when a data warehouse schema is modified. The possible action depends on the type of element that is added to the DMSA (star) schema: measure, property of a hierarchy, or a new dimension: [GLR04]

- *Measure*: In case the added element is a derived measure, i.e. one measure is split in two or more measures, the values for existing data can be calculated from the measure it is being derived from. In case the added element is a non-derived measure, the values must be estimated, or calculated with logic from other measures, or possibly the values can not be estimated or calculated.
- *Property*: In case the added element is a property of a hierarchy all elements in this dimension need to be updated with their specific value for the added property.
- *Dimension*: In case a dimension is added, the historic fact entries that this dimension applies to can be updated with their value for the added dimension. This value can be calculated according to a specified business rule. Another possibility is that these values can not be determined for historic data.

3.7 Conclusion

First the topic of data warehousing was discussed and the three main problems with source system changes -inaccessibility, invalidity, inconsistency- were introduced. This clarified the need for a solid method to perform impact analysis for data warehouses.

With respect to impact analysis, existing research and models were discussed, the main activities of impact analysis were established, the topic of traceability was introduced, which can function as a building stone for impact analysis, and finally evaluation in impact analysis was discussed and evaluation measures were proposed. The knowledge gathered about the impact analysis process in general, can be used to provide the basic outline for the concepts that should come back in the Delta Impact Analysis model that is described in the next chapter. The evaluation measures can also be used in the eventual validation of the DIA model proposed in this research.

The most significant contributions of this chapter are the different categorizations that are provided for delta and impact. The first categorization is that a delta can now be distinguished by its origin: schema, semantic or business rule. A *schema* delta concerns technical/structural changes to a source database, a *semantic* delta concerns changes to the meaning or representation of data in the source, and a *business rule* delta concerns changes to the meaning of source data for the data warehouse. The second categorization is that a delta can now be distinguished by its impact: *information-preserving* or *information-changing*. Since this categorization is both relevant to the concept of delta and the concept of impact, this provides a bridge between defining a delta and determining the impact in impact analysis.

The two types of impact that were identified are *information-preserving* and *information-changing* information supply, where the 'changing' type can be divided by an information supply that has been *reduced*, *increased* and *redefined*. To deal with the different types of information-changing impact one has the choice if the data warehouse is *not modified*, *modified* or *modified including history*.

The concepts discussed in this chapter can be used in the DIA model to better classify the different types of delta, resulting impact and the choices that can be made in the process of impact analysis.

4 A Delta Impact Analysis model for Data Warehouses

The goal of this chapter is to provide an answer to the question '*How can a model be developed to describe the process of Delta Impact Analysis that can be tested at BI4U?'*. This is done by taking all the knowledge gathered so far about DIA at BI4U, Impact Analysis models, delta and impact scenario's in data warehousing, and to combine this into one model that can be used to perform a solid Delta Impact Analysis for a data warehouse. The combining and relating of this knowledge from these different knowledge sources and research fields into one model is considered to be the main contribution of this research.

In Section 3.2 the main activities of an impact analysis process have been established. Regarding software engineering Lethbridge and Laganière state that to solve a customer's problem it is essential that one understands the business environment, the problems and the available solutions of the problems of the customer. Once this is done one can meet with the customer to decide how to tackle the problems. [LL01] This relates to what was discussed in Section 2.4, the DIA process at BI4U of determining the solution directions and proposing these to the customer. Combining this knowledge, the following steps can be identified for the process of Delta Impact Analysis:

1. Determining the change set
2. Determining the impact set
3. Describe solution space
4. Determining the required modifications to the data warehouse
5. Determining the resources and costs required to perform the modifications
6. Communicate the results of the DIA
7. Evaluating the DIA process

Figure 14 below visualises the proposed DIA model and the seven steps.

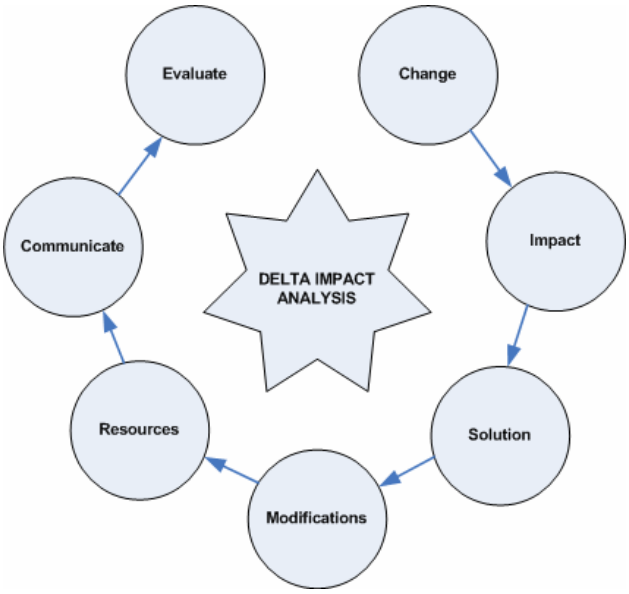


Figure 14. Model for Delta Impact Analysis

With respect to the terminology used in Chapter 3, the first two steps would have been respectively called '*Determining the primary impact set'* and '*Determining the secondary impact set*', but to make the model more easy to understand, it was chosen to use different terminology.

Lethbridge and Laganière discuss the topic of quality assurance in software engineering. Quality assurance can be divided into validation and verification. [LL01] Validation takes place before implementation and concerns checking that the defined requirements will solve the customer's problems. Verification takes place after implementation and concerns making sure that all requirements have been taken into account in the implementation. So where validation is a part of impact analysis, verification is a contribution of the DIA process to the implementation process.

Figure 15 is an activity diagram of the suggested DIA process. The diagram is based on what Delta Impact Analysis means for BI4U (Chapter 2) and the established goals and the gathered insights of and what can be learned from research about DIA (Chapter 3). One can clearly identify the seven main steps as activities in the diagram. It also shows which activities have output that is added to the DIA document; the deliverable of the process. The validation as suggested by Lethbridge and Laganière is implemented by the validation activities that are in between the steps representing the seven main activities. Validation can be done by the person performing the impact analysis or by a colleague. Validation should be done by questioning if the information is complete and if it adds value to solving the problem.

It is the intention that this process is executed by a developer with generic technical knowledge of data warehouse solutions, who has access to several information sources regarding the specific data warehouse solution that were discussed in section 3.2.1:

- Knowledgeable developers.
- The actual source code and documentation of the data warehouse solution.
- The original design models of the data warehouse solution.

In each section of this chapter one of the steps will be discussed in more depth as to how this step can be performed and what the activities actually are.

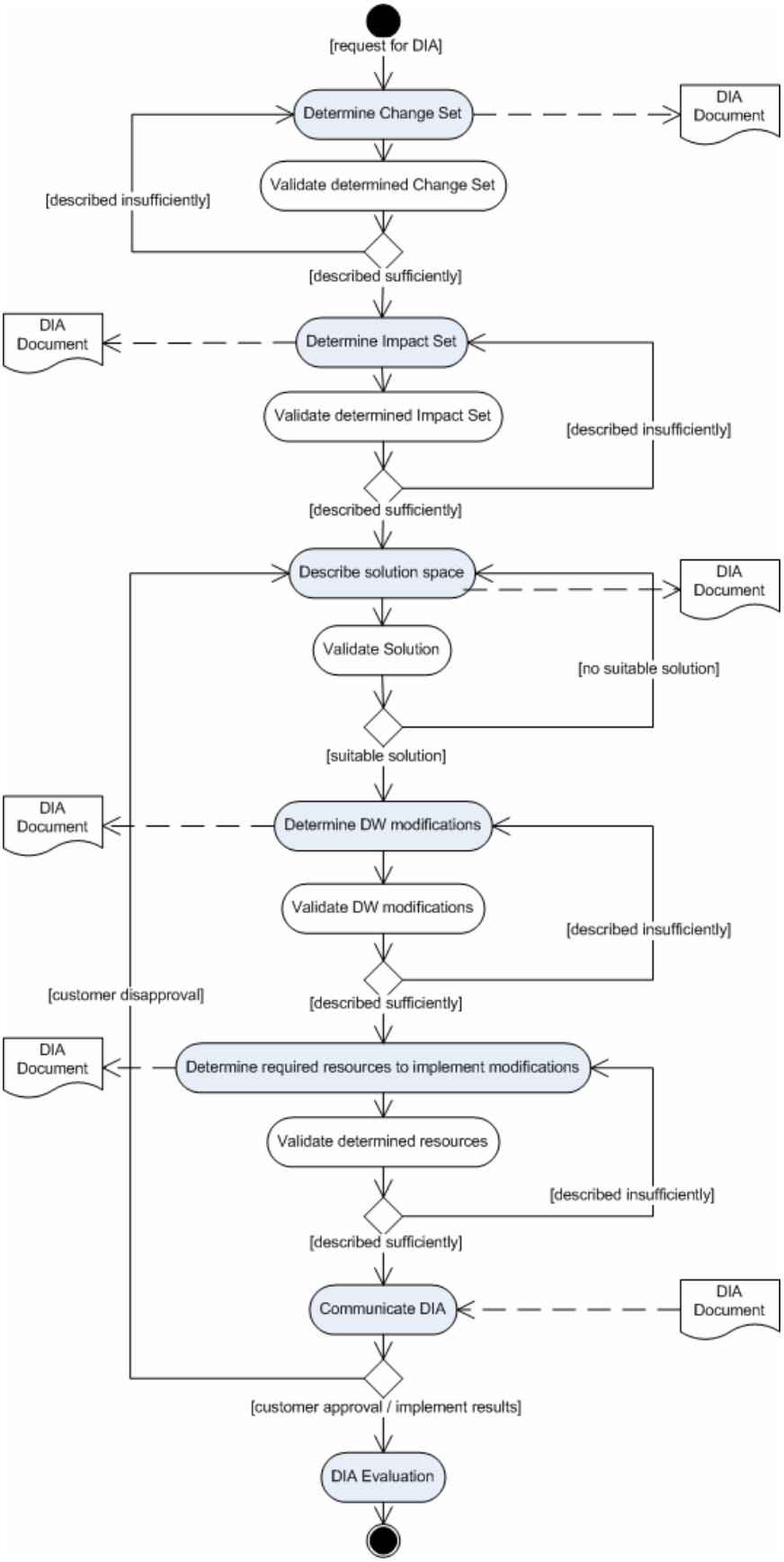

[request for DIA]
Determine Change Set
DIA Document
Validate determined Change Set
[described insufficiently]
[described sufficiently]
DIA Document
Determine Impact Set
Validate determined Impact Set
[described insufficiently]
[described sufficiently]
Describe solution space
DIA Document
Validate Solution
[no suitable solution]
[suitable solution]
DIA Document
Determine DW modifications
Validate DW modifications
[described insufficiently]
[customer disapproval]
[described sufficiently]
DIA Document
Determine required resources to implement modifications
Validate determined resources
[described insufficiently]
[described sufficiently]
Communicate DIA
DIA Document
[customer approval / implement results]
DIA Evaluation

Figure 15. Process of Delta Impact Analysis

4.1 Determine Change Set

Figure 16 is an activity diagram that describes the process of how to determine the *change set* (CS). The very first step is classifying the type of delta of which the impact is to be analyzed. The techniques to determine the change set will have to be chosen based on the fact whether it concerns a schema, semantic or business rule delta. Therefore these three types of delta will be discussed in the next sub-sections.

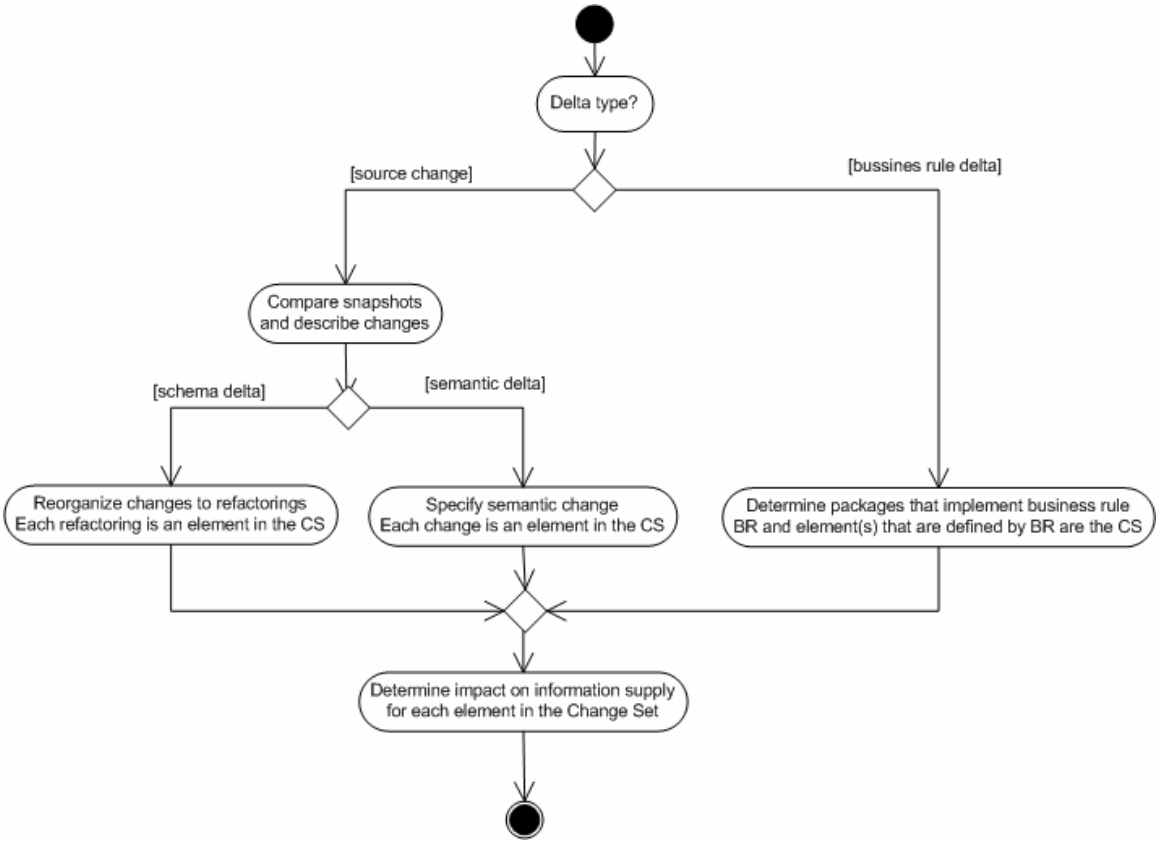


Figure 16. Determine Change Set (CS)

The possible impact on the information supply for the different types of delta as discussed in Chapter 3 are summarized in Table 4.

Table 4. Possible impact on information supply

	Preserved	**Reduced**	**Increased**	**Redefined**
Schema	x	x	x	
Semantic	x	x	x	x
Business Rule		x	x	x

4.1.1 Schema Delta

In order to create a complete overview of the changes a database compare tool as suggested in Section 2.4 (*Red Gate SQL Compare*[1]) can be used to compare the two different versions of the source database. As discussed in Section 3.3 all structural changes found can then be mapped to a specific refactoring. Based on the refactoring it can be determined what the implications are to the information supply of the source system to the data warehouse, Table 5 provides an overview.

In Section 3.3 it was discussed how refactorings that change the information supply require more effort to determine the impact and making the modifications to keep the data warehouse working and consistent. Since most refactorings are combinations built out of the primitive introduce/drop column/table refactorings it requires logical thinking of the person performing the DIA to determine that certain primitive refactorings together form a refactoring that keeps in the information supply intact. Possibly the database compare tool can assist in determining relations between certain changes.

Table 5. Information supply impact of refactorings

Information supply	Event
Preserved information supply	Merge table, Split table, Move column, Rename column, Rename table, Replace large object with table, Replace one-to-many with associative table Replace column, Introduce surrogate key, Replace surrogate key with natural key, Drop view, Rename view
Reduced information supply	Merge column, Drop column, Drop table
Increased information supply	Split column, Introduce column, Introduce table

Elaborating on the first example from 2.3 the database compare tool would identify the following structural changes:

- Drop table ‘Employees’
- Drop table ‘PersonDetails’
- Insert table ‘Employees_PersonDetails’

Instead of recognizing two information reducing and one information increasing refactoring, some analysis can result in the recognition that this in fact is one ‘merge table’ refactoring. Based on this classification one can then also establish that this refactoring results in a preserved information supply of the source system to the data warehouse. This fact is useable in the next step of determining the impacted elements.

The result of this step will be a list of refactorings. For each refactoring it is defined which structural changes are part of this refactoring, and what the impact on the information supply is for this refactoring.

4.1.2 Semantic Delta

Unlike *schema delta*, a *semantic delta* can not be identified by comparing database schemata. It will be necessary to compare two versions of a database on the actual

[1] http://www.red-gate.com/products/SQL_Compare/index.htm

content level. For instance a tool like *Red Gate SQL Data Compare*[2] could be used for this. Figure 17 and Figure 18 provide an example of a source system 'SRC' with Table 'X' that contains a column 'Z' that represents the last login date of a user. To identify a semantic delta, ideally one would compare two database snapshots of the source system, the current and the new version, without any executed operational data updates performed in the time between the two snapshots (Figure 17). If operational data updates have been executed in the time between the two snapshot were taken, then this comparison has to be done intelligently, since changes can be due to both data updates (Figure 18a) and semantic change (Figure 18b). This could be done by for instance alerting for the fact that for all tuples in a table, the data of a specific column is different between the two database snapshots. While this could still be the result of data updates, a semantic change is also a logical cause for such an event.

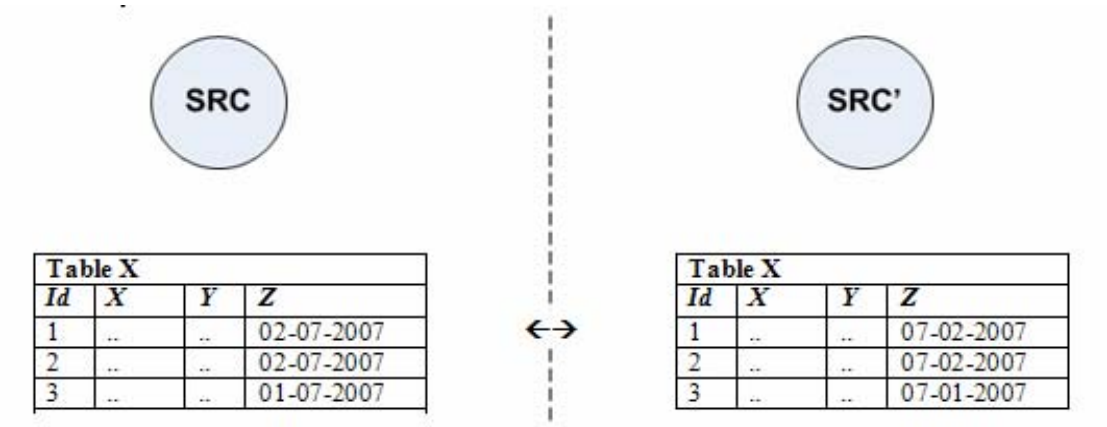


Figure 17. Database snapshot, no DU possible

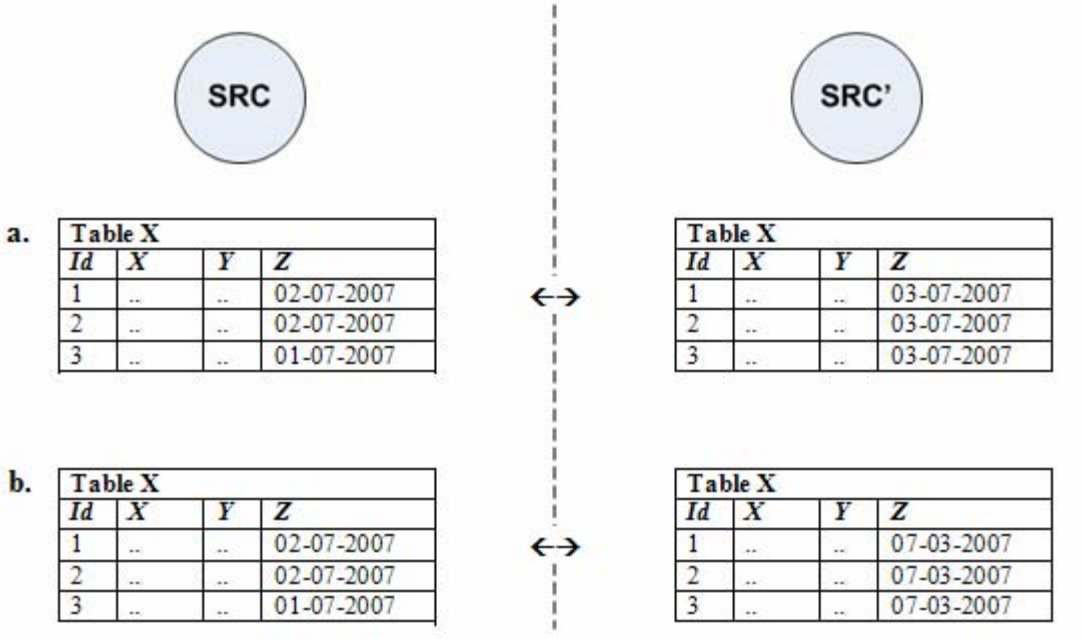


Figure 18. Database snapshot, DU possible

[2] http://www.red-gate.com/products/SQL_Data_Compare/index.htm

The elements that are semantically changed form the change set. The person performing the impact analysis will have to assess what type of change each element of in the change set is, and if the information supply to the data warehouse is preserved or changed. In case the information supply is changed it will have to be assessed whether it is reduced, increased or redefined. In order to assess this the person can investigate what kind of semantic change it concerns by comparing snapshots and logic reasoning, with the suggested examples as a reference point. For the examples mentioned in Section 3.4 it will be discussed how one could identify these.

Information-preserving

- *Example "DD-MM-YYYY" => "MM-DD-YYYY":*
  See Figure 17
- *Example Enum (A, B, C) => into Enum (D, E, F):*
  For all existing entries with the value A, this changed to D.
  For all existing entries with the value B, this changed to E.
  For all existing entries with the value C, this changed to F.

Information-changing: redefining (application of data)

- *Money value changed from EUR => USD:*
  For all existing entries the amount has changed. Every change is according to the same ratio. E.g. If an entry with the value 1000 changed to 800, then an entry with the value 100 should have changed to 80.
- *A money value is changed from representing an amount before tax to an amount after tax:*
  The values of all entries can be changed, but not necessarily in a comparable way, this depends on what the tax rules and rates are. It is not that difficult to identify that the semantics of the data changed, but it can be very difficult to identify what the change actually is.
- *A 'name' column represents first name instead of surname:*
  For all existing entries the value has changed to something different. An exception that can complicate the recognition of this fact is if the first name of a person is equal to his surname.

Information-changing: reducing or increasing (contents of data)

- *Enum (A, B, C) expanded to Enum (A, B, C, D):*
  Looking at the different possible values for all entries, it no longer only is A, B or C. Certain entries now have the value D.
- *Orderdate stored in Unix timestamp => "DD-MM-YYYY":*
  For all existing entries the value has changed from something formatted as a string of numbers to a string representing a date.

4.1.3 Business Rule Delta

As discussed in Section 3.5, a business rule can be very complex and therefore it can be implemented by complex transformations gathering data from many elements.

For each of the business rule changes, it has to be specified if the information supply is redefined, reduced (deleted business rule) or increased (added business rule). In the case of a deletion or a redefinition of a business rule it has to be identified by which transformation package this business rule is implemented. The elements that are defined by the business rule form the change set for this delta. In case a business rule is added the elements that are input for the business rule form the change set.

4.2 Determine Impact Set

What one wants to do in this activity step of the impact analysis, is to determine the set of elements of the data warehouse that are estimated to be impacted (the impact set; IS), by each of the elements that are in the change set (CS). In the perspective of DIA this concerns the process of mapping the impact on the data warehouse of each of the schema refactorings, identified semantic changes and identified business rule changes that were specified in the previous step. Possible types of impacted elements are database tables and columns and business rules implemented by the packages.

Figure 19 is an activity diagram that describes the process of determining the impact set. Which data warehouse components are included in tracing the dependencies, is dependent on the impact on the information supply of the delta. Using traceability information in combination with the type of delta and its impact on the information supply it will be possible to define the impacted elements on a component level. The result of this step is an overview of all elements that are part of the change set, and for each of these what their relationships are with the elements of the other data warehouse components (impact set). This overview can be provided either visually or textually, or both.

In case it concerns a schema or semantic change that preserves the information supply, it is only necessary to determine the dependencies till the DSA-to-ISA components. In case it concerns a business rule delta, it is not applicable to look at dependencies at SRC and DSA components, since the business rules are implemented in the DSA-to-ISA component.

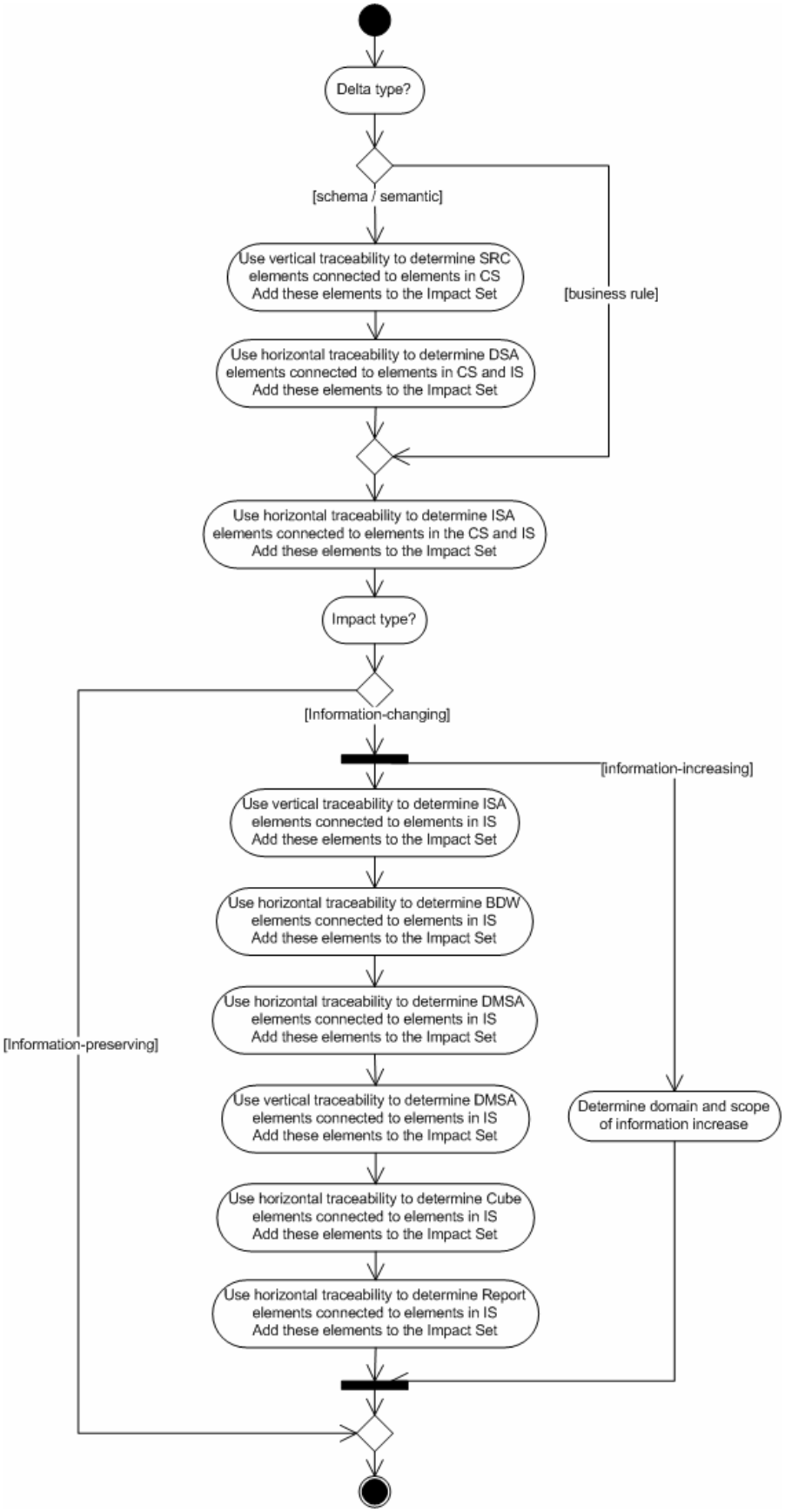

Delta type?
[schema / semantic]
[business rule]
Use vertical traceability to determine SRC elements connected to elements in CS Add these elements to the Impact Set
Use horizontal traceability to determine DSA elements connected to elements in CS and IS Add these elements to the Impact Set
Use horizontal traceability to determine ISA elements connected to elements in the CS and IS Add these elements to the Impact Set
Impact type?
[Information-changing]
[information-increasing]
[Information-preserving]
Use vertical traceability to determine ISA elements connected to elements in IS Add these elements to the Impact Set
Use horizontal traceability to determine BDW elements connected to elements in IS Add these elements to the Impact Set
Use horizontal traceability to determine DMSA elements connected to elements in IS Add these elements to the Impact Set
Use vertical traceability to determine DMSA elements connected to elements in IS Add these elements to the Impact Set
Use horizontal traceability to determine Cube elements connected to elements in IS Add these elements to the Impact Set
Use horizontal traceability to determine Report elements connected to elements in IS Add these elements to the Impact Set
Determine domain and scope of information increase

Figure 19. Determine Impact Set (IS)

As discussed in Section 3.2.2 vertical and horizontal traceability can be achieved by documenting all relationships between attributes within the same component and between the attributes of the different components. For a data warehouse this comes down to documenting the elements of the different components (mostly database structures) and the relations between those components that are represented in the transformation packages. This could be done manually for each impact analysis, but it would be far more efficient to create documentation of all relations during the data warehouse design and to in maintain an up-to-date version of this during the lifetime of the data warehouse. A meta warehouse would be a suitable solution for this, the topic of a meta warehouse will be discussed in more depth in Section 6.2.2.

Red Gate offers the tool *SQL Dependency Tracker*[3], which claims to be able to determine and visualize dependencies between database objects, also cross multiple databases, in Microsoft SQL environments. It will be interesting to look into these kinds of tools to support this step of the Delta Impact Analysis process.

Elaborating on the example from section 2.1, one wants to know the impact of a merge column refactoring of source columns PersonDetails.name and PersonDetails.surName into the new column PersonDetails.fullname. The fact that this refactoring reduces the information supply means that potentially all components of the data warehouse can be impacted. Traceability information will result in the knowledge visualized in Appendix D. It is then the choice of the person performing the Delta Impact Analysis to determine if this change is resolved in the DSA-to-ISA stage or if it is propagated to the other components of the data warehouse. If it is, the traceability provides the insight into which elements of these components are impacted.

[3] http://www.red-gate.com/products/SQL_Dependency_Tracker/index.htm

4.3 Describe solution space

In the case of Delta Impact Analysis this step is also valuable because it will provide a point for customer involvement. When impact analysis is performed in a scope where there is no customer involved, the solution space can be described in order to discuss this with those people in the organization that have the decision making function for the project, they could be seen as an internal customer

The solution space is defined by performing the next step of 'determining the required modifications' on an orienting level: giving some interpretation to the determined data warehouse impact and establishing if there are more ways to tackle a specific delta and its impact. The impact-activity tables (Table 6 and Table 7) can function as an inspiration for the different possible solutions. For example in the case of increased information supply one solution direction can be how to deal with the impact without using the increase in information supply, and another solution direction can be how to deal with the impact by utilizing the increase in information supply by modifying the ISA->Reports components. Based on the identified impacted elements a rough estimation on the difference in size of the different solutions can be given by thinking of the required modifications superficially.

Figure 20 visualizes the process of describing the solution space. The result of this step in the DIA process is an overview of the possible solutions for each of deltas that were defined in the first step of the process, concluded by a choice of the solution.

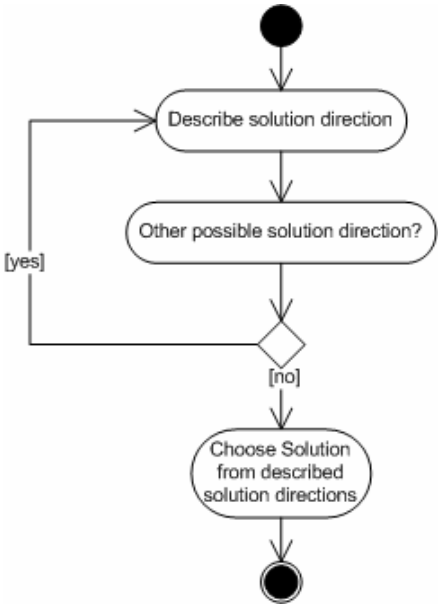


Figure 20. Describing solution space

4.4 Determine required modifications

In the previous steps all changes have been categorized with regard to the type of delta and classified by the type of impact on the information supply. Besides this, it should be clear which delta, impacts which elements of which components based on traceability information. The chosen solution direction also provides insight about what type of modifications are required to be performed.

Figure 21 provides a diagram to determine the required types of modifications to manage a specific change element in the change set. This means for each element of the change set, all the steps of the diagram are gone trough to determine the modifications required to implement the solution to the impact for that specific delta. Depending on the situation, modifications out of one or more of the three categories need to be performed.

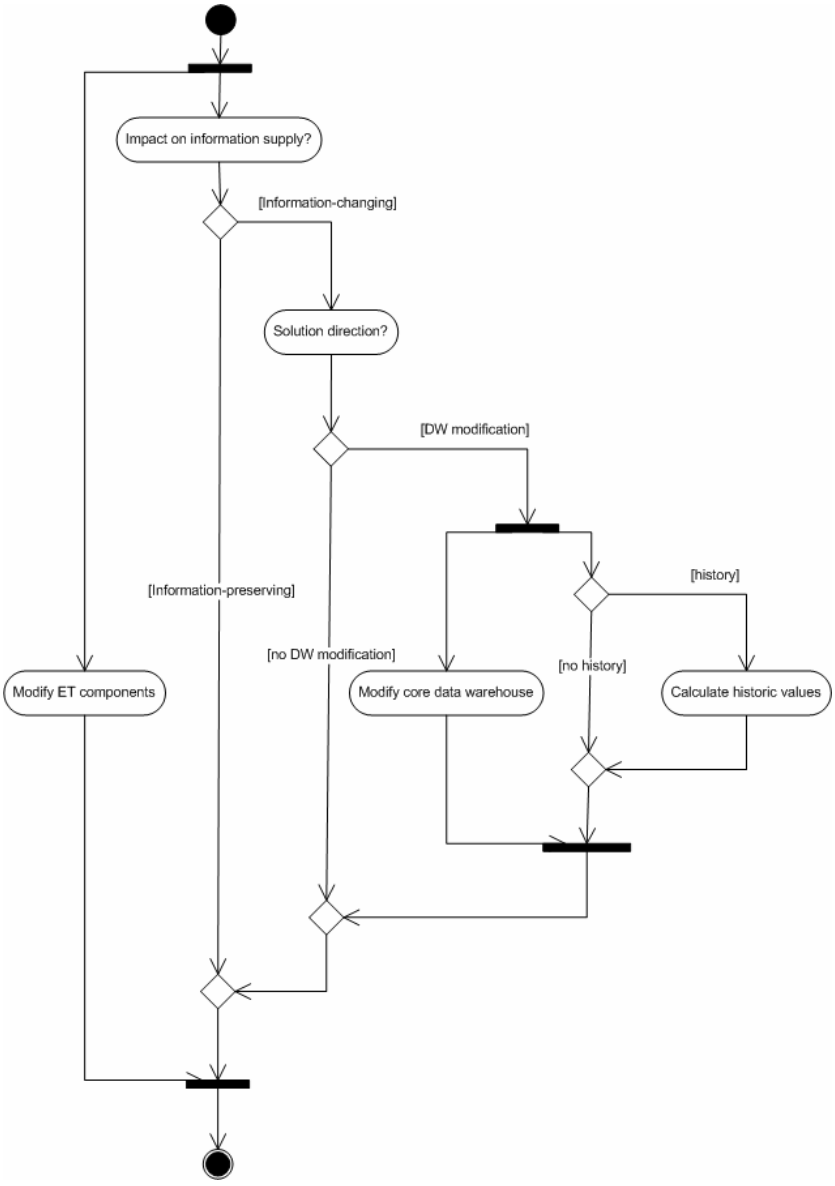


Figure 21. Determine required modifications

Table 6 provides an overview of required and possible modifications for each group of data warehouse components, depending on the impact on the information supply of the delta. In Section 3.6 the topic of data warehouse evolution and versioning was also discussed, Table 7 provides an overview relevant to when data warehouse evolution and versioning is supported. When a column is filled with 'must' this implies that the activities must be performed, when a column is filled with 'can' it is a choice to perform the activity or not, when a column is '-' this means that there are no applicable activities of that type in that situation.

Table 6. Impact-activity, without evolution/versioning support

	Preserved	**Reduced**	**Increased**	**Redefined**
SRC -> DSA-to-ISA (ET)				
Modify	must	must	must	must
ISA -> Reports (Core)				
Modify	-	can	can	must
Calculate historic values	-	-	can	***must***

Table 7. Impact-activity, with evolution/versioning support

	Preserved	**Reduced**	**Increased**	**Redefined**
SRC -> DSA-to-ISA (ET)				
Modify	must	must	must	must
ISA -> Reports (Core)				
Modify	-	can	can	must
Calculate historic values	-	-	can	***can***

The actual required modifications for each delta depend on the impact on the information supply and the chosen solution direction for the delta. The three categories of modifications listed in the table explained:

1. *Modifying the ET components*: modifications to the elements of the Source-to-DSA, DSA-to-ISA packages and DSA database.
2. *Modifying the core data warehouse:* removing or introducing the new element in the ISA, BDW and DMSA databases, the ISA-to-BDW, BDW-to-DMSA, DMSA-to-Cubes packages, the cubes and the reports.
3. *Calculating historic values*: activities that involve calculating historic values according to the new situation. and modifying existing data warehouse values.

The result of this step of the DIA process is an overview of the required modifications to the ET and core data warehouse components for each delta that was defined in the first step of the process. Modifications regarding ET and core data warehouse components need to be specified specifically with regard to which elements need to be modified and in which way. Activities regarding the calculating of historic values can be more specifically specified according to the different possible activities listed in Table 8. The activities in Table 8 are based on the possible actions specified by Golfarelli et. al, as discussed in Section 3.6.3. [GLR04]

Table 8. Calculation actions

Analysis function	Action
Measure	Estimate values for the new measure
Derived measure	Compute the values for the new measure
Property	Consistently add values for the new property
Dimension	Disaggregate measure values according to a business rule

4.5 Determine resources and costs required

In Section 3.2.3 four required steps were mentioned in order to be able to determine the costs of a certain delta. The impact domain is already identified by the DIA process steps of determining the change set and the impact set. The remaining steps can be modelled in activities as shown in Figure 22.

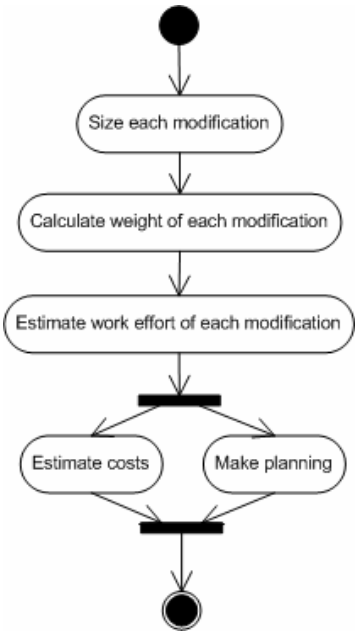


Figure 22. Determine resources and costs required

Arthur discusses sizing of impact analysis in software evolution. One possibility is that of statistic analysis, where one looks at the lines of code and decision logic (if, else, and, or not, etc.) to determine the complexity of an impacted component. [Art88] Sizing of the impact domain regarding data warehouses could be approached in a similar way. For instance the impacted elements in each component of the data warehouse could be counted.

In order to adjust the sized impact domain to the complexity of the impacted components it must be defined how we can determine the complexity. A distinction was already made between the different modification activities that can be performed. This same distinction can be used to relate these different types of activities to a level of complexity.

- Modify elements in the Source -> DSA-to-ISA components [1]
- Modify elements in core Data Warehouse components [2]
- Calculate history values for the Data Warehouse [3]

For instance the weight of complexity of an impacted element of category 1 could be 4, because it does not require any complex activities, only some modifications of the ET databases and packages. Where the weight of an impacted element of category 2 can be 7, since more complex activities are going to have to be performed. This would mean that if 5 elements of category 1 are impacted this has a weighted impact of 5 * 4 = 20, where 5 impacted elements of category 2 would have a weighted impact of 5 * 7 = 35. Determining the actual weight of these categories will have to be done based on more professional insight and experience.

The next step is to translate the defined weighted impact to estimated work effort. The goal is to estimate how much work effort it requires to handle the determined impact of a certain weight. This estimation should also take into account the 'optimist factor', the margin to catch underprediction, which Lindvall discusses in [Lin03]. This can be done based on information of past Delta Impact Analysis that were executed. E.g. if in the past the execution of a DIA with a weighted average complexity of 60 took 90 minutes to execute, the translation factor can be defined as 1.5. Obviously this can be done on a more specific level, where more historic information is used and possibly a translation factor is defined for each of the complexity categories.

Table 9 provides an example of a table that can function as a reference in order to calculate the required effort for a modification of a specific complexity. The X values will have to be determined based on expert intuition. The Y values can be calculated before each impact analysis based on information about past impact analyses and the workload that was actually experienced with the implementation. It is very well possible that in practice more complexity categories will be recognized, which can be an addition to those listed in Table 9.

Table 9. Modification complexity categories

Category	Modification Type	Weight	Effort factor
1	Modify Source -> DSA-to-ISA components	X_1	Y_1
2	Modify core Data Warehouse components	X_2	Y_2
3	Calculate history values	X_3	Y_3

The final step is to calculate the costs and make a planning for the modifications based on the estimated work effort. As discussed in Section 3.2.3 a planning should take into account both time with respect to work effort as to a period in which the effort is performed.

To provide a basis for planning and cost estimation the *Esterling Time Study Model* as described by Arifoglu is presented. [Ari93] This model was chosen due to its applicability in smaller projects and the fact that empirically determined values for parameters are provided. The following formulas can be deducted:

- Calendar-time to person-time ratio = 7 / (5 * n * w)
- w = 0,125 * [7,53 + (21*(n-1)/60)]
- n = number of people working on the project

Assuming 1 person implements the results of the DIA the *calendar-time* to *person-time* factor would be 1,487. So if 50 hours are estimated to be required, and there are 25 person hours available a week, for planning purposes one should use a timespan of (50/25) * 5 *1,487 = 14,87 calendar days. This ratio can function as a starting point and can be adjusted according to experience.

The result of this step in the DIA process is an estimation of costs and a planning for the required modifications for each chosen solution.

4.6 Communicate the results of DIA

In the introduction of this research it was suggested that a Delta Impact Analysis is performed on customer request. Even though this might not always be the case, the step of developing a final document to communicate the results, is considered a step in the DIA process. The combination of all documentation from the previous steps, will result in a solid document consisting of the identified delta, its possible impact, the solution space, the required modifications of the chosen solution and costs and resources required to implement the change and required modifications. With some final editing this document can be presented to the customer as a proposal to the customer to tackle the delta.

The activity of creating a document of the DIA process is performed alongside the previously performed activities. Lindvall identified the correct documentation of the impact analysis process as a constraint for impact analysis. [Lin03] The result of each step in the DIA process should be documented, so it can be used for the next step and eventually during the implementation stage of the change.

Based on the activities performed in the Delta Impact Analysis, a suitable way to structure a DIA document would be as suggested in Figure 23 on the next page.

1. Change Set
    1.1 Delta A
    1.2 Delta B
      Etc.
2. Delta A
    2.1 Estimated Impact
        2.1.1 DSA
        2.2.2 BDW
        2.2.3 DMSA
        2.2.4 Cubes / Reports
    2.2 Solution Space
        2.2.1 Solution 1
        2.2.2 Solution 2
          Etc.
        2.2.x Chosen Solution
    2.3 Required modifications
        2.3.1 DSA
        2.3.2 ISA
        2.3.3 BDW
            Etc.
    2.4 Resources required
3. Delta B
   Etc.
4. Costs
5. Planning
6. Evaluation

Figure 23. Document structure for documenting DIA

4.7 Evaluation of the DIA process

The possibility of evaluation was discussed in Section 3.2.4. Evaluation is not something that can be done before implementation. To evaluate the performed Delta Impact Analysis, its results first have to be implemented and the actual outcome can than be compared to that what was estimated during the Delta Impact Analysis process. The implementation itself however is not part of the process of Delta Impact Analysis.

The evaluation measures proposed in Section 3.2.4 can be used for this final evaluation step of the DIA process. The evaluation can be used to determine the adequacy and effectiveness of a performed Delta Impact Analysis in general. The evaluation can also be used to adjust the Delta Impact Analysis model for future use, i.e. to improve the performance of estimating the impact set and the required resources.

4.7.1 Adequacy

The adequacy of an impact analysis can be assessed by the following indicators:

- Correct-Inclusiveness rate: $|(AIS \cap EIS)| \ / \ |EIS|$
- Wrong-Inclusiveness rate: $|((AIS \ \Delta \ EIS) \cap EIS)| \ / \ |EIS|$
- Missed-Inclusiveness rate: $|((AIS \ \Delta \ EIS) \cap AIS)| \ / \ |EIS|$

In order to calculate a value for these indicators the following must be done:

- Determine Estimated Impact Set (EIS), the elements estimated to be impacted, which is the sum of the Change Set (CS) and Impact Set (IS).
- Determine Actual Impact Set (AIS), those elements that were actually impacted during implementation, must be determined
- Determine Wrong Impact Set (WIS), those elements that are in the EIS that are not in the AIS, must be determined
- Determine Missed Impact Set (MIS), those elements that are in the AIS that are not in the EIS, must be determined
- Elements in the AIS, EIS, WIS, MIS should be counted.

With respect to evaluation to facilitate learning, it is important to provide feedback on the process, in order to improve the current process of DIA. Therefore the persons executing the DIA and the person implementing the results should be asked to answer the following questions as thorough as possible after implementing the DIA:

- Were the steps provided in the DIA model sufficient to perform the DIA? If not, in which way differed the actual performed steps from the model?
- What are the differences between the estimated impact and the actual impact?
- What are the differences between the estimated modifications and the actual modifications?
- What are the differences between the estimated resources & costs and the actual resources & costs?
- What are the differences between the planning and the actual time-to-implement?
- In which way can/should the DIA model be adjusted to improve this for the execution of future DIA's?
- Are there any other remarks with regard to the performed DIA?

4.7.2 Effectiveness

The effectiveness of an impact analysis can be assed by the following indicators:

- Amplification ratio: $|IS| / |CS|$
- Change rate: $|EIS| / |System|$
- Time spent to perform DIA

In order to calculate a value for these indicators the following must also be done:

- Total number of elements in the system should be counted
- Elements in the change set and impact set should be counted

Besides using calculations, another way to evaluate the effectiveness of a performed impact analysis is by asking the opinion of the person implementing the change as suggested in Section 3.2.4 by using the survey questions in Appendix C. The person implementing the change can very well be the same person that performed the Delta Impact Analysis. With respect to evaluation (bias and objectivity) it is preferred that the change is implemented by a different person, however the implementation might be performed better by the same person, especially if the Delta Impact Analysis was not documented correctly.

5 Validation of the DIA model

In this chapter the topic of validation of the proposed model for Delta Impact Analysis will be discussed. First it will be discussed how the model is validated in a field study. Second, validation is realised by reflecting on the design science research framework of Hevner et al. [HMP04]

5.1 Field study

Lee suggests in [Lee07] that the design science approach combined with *action research* can result in a greater rigor and a greater relevance than either approaches by themselves. Baskerville discusses the topic of action research regarding information systems in [Bas99]. The main idea behind action research is that social processes are best studied by introducing changes into the social processes and observing the effects of the changes. In action research the researcher both observes and participates in the research. The validation of this research can be done using the action research approach: introducing a change and observing the effects. The validation of the DIA model can be done by establishing that impact analysis by the organization is performed equal or better after introducing the change: the use of the proposed DIA model. A way to establish this is to perform a field study in which impact analysis has to be performed twice on one specific case, or once on two very similar cases. The first impact analysis should be executed before the change (before introducing the model), and the second impact analysis after the change (after introducing and using the model). The results of the performed impact analyses can then be compared.

Baskerville names three effects that are inevitable when performing action research: an interpretivist viewpoint, an idiographic viewpoint and a qualitative nature. [Bas99] An interpretivist viewpoint concerns the fact that the researcher can become part of the research, the researcher's value and knowledge become part of the observations and thereby invade the observations made. This should be taken into account when observing, interpreting and presenting the results of the research. An idiographic viewpoint concerns the fact that a social setting is strongly dependent on the people involved in the setting and thereby it is close to impossible to replicate the exact situation. Therefore it is very important to clearly identify and document the context of the social setting and the people involved. The result of an interpretivist viewpoint and an idiographic viewpoint is the fact that the research will be of a qualitative nature. It is important to establish this fact and to not present the research as quantitative.

Baskerville distinguishes five phases of action research. [Bas99] The first phase concerns that of *diagnosing* what the main problems intended to be solved are. This will be done in paragraph 5.1.1 by proposing the hypotheses that are to be tested. The second phase is that of *action planning*, which concerns the planning of the to-be-performed activities to solve the identified problems. This will be discussed in paragraph 5.1.2 with the field study design. The third phase is that of *action taking*, where the planned activities are actually executed. This will be described in 5.1.4. The fourth phase is that of *evaluating*, where the results of the executed actions are compared to their solving effect on the identified problems. This will be discussed in paragraph 5.1.5. The fifth phase is that of

learning, which will be reflected upon in paragraph 0. The limitations of the field study will be discussed in paragraph 5.1.3.

5.1.1 Hypotheses

In Section 3.2.4 multiple methods were discussed that can be used to measure the performance of an impact analysis with regard to adequacy and effectiveness. These methods will be used to define the hypotheses that will be tested by the field study. However the amplification rate and change rate to determine effectiveness are not relevant with regard to comparing the effectiveness of two impact analysis approaches. Because using different impact analysis approaches for a specific case should result in the same amplification rate and change rate, unless wrong/missed impact is assessed by an impact analysis approach, and that aspect is already evaluated by the adequacy indicators. Therefore these indicators are not translated into hypotheses.

The main hypothesis is that the performance of the executed impact analyses based on the model equals or surpasses the executed impact analyses without knowledge of the model. For convenience, we will refer to impact analysis based on the model as DIA_M , and to impact analysis without knowledge of the model as DIA_I (with I for Intuitive). The first three hypotheses are based on the concept of adequacy, where the next four are based on the possible scoring aspects regarding the effectiveness concept. The final hypothesis is the fact that structure provided by the model will result in less time required to perform a DIA.

The following hypothesizes will be tested:

- H1: *The Correct-Inclusiveness rate* $DIA_M >= DIA_I$
- H2: *The Wrong-Inclusiveness rate* $DIA_M <= DIA_I$
- H3: *The Missed-Inclusiveness rate* $DIA_M <= DIA_I$
- H4: *Average score on aspect 'Scope'* $DIA_M >= DIA_I$
- H5: *Average score on aspect 'Required resources'* $DIA_M >= DIA_I$
- H6: *Average score on aspect 'Cost/Benefit analysis'* $DIA_M >= DIA_I$
- H7: *Average score on aspect 'Communicate complexity of a change'* $DIA_M >= DIA_I$
- H8: *Time to perform* $DIA_M <= DIA_I$

5.1.2 Design

Since the resources of the company were limited it was only possible to use one subject for the field study, therefore the setup of the field study is as follows. Two similar and comparable delta cases are defined, and one subject will perform delta impact analyses on both cases. The first case is analyzed before having knowledge of the model. After completion of the first DIA, the model that is proposed in this research is explained to the subject. Next, the second case is analyzed using the proposed model. In order to best represent the real world situation, the subject performing the Delta Impact Analysis should be a person that would usually be appointed to perform an impact analysis when a source system is changed.

The scoring of the performed impact analysis on the concept of adequacy can be done after the implementation of the change, using the outcome of the impact analysis to

determine the AIS. In case the results are not implemented, an alternative to this is that the person that would use the outcome of the impact analysis to implement the modifications, carefully analyzes the situation and assesses what the AIS is according to him or her.

The scoring of the performed impact analysis on the concept of effectiveness is also done by the person that would use the outcome of the impact analysis to implement the modifications. This scoring can be done before the actual implementation of the impact analysis results.

For each delta case the scores of the DIA_M can be compared to DIA_I to test which hypothesizes hold.

Several kinds of people can be distinguished in the field study:

- Researcher R: The person performing this research.
- Expert E: The expert person available for consultation. This is also the person assessing the adequacy and effectiveness of the impact analyses.
- Subject S: The person executing the impact analyses.

The following requirements need to be satisfied by the field study

- S will be provided with the first delta case and asked to perform an impact analysis in the way he or she is used to (DIA_I).
- After having performed the first DIA, S will be given access to the research and will be given an explanation of the model that is described in this research.
- S will be provided with the second delta case and asked to perform an impact analysis using the model proposed in this research (DIA_M).
- S will be able to consult R during the execution of DIA_M for questions regarding the model.
- S will be able to consult E during the execution of both DIA_I and DIA_M as a source for expert knowledge. Since this is common when executing a DIA.
- Directly after finishing the impact analysis both S and E will be surveyed separately to assess the quality of DIA_I and DIA_M.

5.1.3 Limitations

Regarding the design of the field study three remarks can be made. First of all in the ideal situation multiple subjects would perform Delta Impact Analysis on the same two (or more) cases. Which cases would be used for of DIA_I and DIA_M could then also be alternated between the various subjects. This would provide results in which personal skills and characteristics of the specific cases can be ruled out; these results would better represent the real world. The second remark that can be made, is that it was not possible to hide the fact if the to-be-evaluated output was based on DIA_I or DIA_M from expert E. This was due to the fact that the expert was involved during the earlier stages of the research and she would easily have been able identify the used approach to perform the DIA. Therefore also no effort was invested in trying to hide this aspect. To prevent bias, ideally the results of the executed DIA's should be judged by experts who would not be

able to assess if the resulting work is based on the resulting model of the research or based on the intuitive approach. The third remark that can be made is the fact that the feedback by the subject might be less objective due to the fact that he would logically value his work as high as possible. However the difference in feedback between DIA_I and DIA_M on his own work does indicate that he feels that the work he has done using the model is of a better quality than without the model.

Regarding the actual execution one main remark can be made. The provided cases did not have the best fit with this research. It did concern delta cases, however these delta did not regard source system changes, but changes in the reporting possibilities of the data warehouse. Therefore it was not possible to assess the usefulness of categorizations for delta and impact that are proposed in this research. However, the cases did still provide the opportunity to assess the usefulness of providing a model with steps that describe the process of Delta Impact Analysis.

5.1.4 Execution

The field study was monitored by the researcher. Besides this a screen recorder tool was used to record the actions of the subject on the computer, and a camera was set to record the complete duration of the field study. The screen and camera recordings were used afterwards by the researcher to guarantee a correct documentation of the field study.

The requirements as described in the previous paragraph were strictly followed. The subject was given the explanation (Appendix E) that two DIA's were to be performed by him: the first without any further explanation by the researcher about the research, the second after an explanation about the research was given. Also he was informed that afterwards it would be evaluated what the differences between the two performed DIA's were. What was different from the original research design was the fact that Expert E was not actually present during the field study, however this person was consulted once by the subject by phone. Also the field study took some more time than planned, where it was expected that everything could be done in half a day, it eventually took a whole day to perform the two DIA's and evaluate the results with the subject.

In the remains of this section four phases of the field study will be discussed separately: execution of Case 1, intervention, execution of Case 2, evaluation.

Case 1

The first case concerned a request for change (RFC) concerning the data warehouse model. Where currently it is possible to group reports on a specific performance indicator, the request for change is that this grouping ability is expanded to support a wider definition of the performance indicator.

The observed work approach of the subject can be summarized as:

- Assess the case.
- Consider the possible solution.
- Describe the required changes to realise the solution.
- Describe a test-case for after implementation.

To consider the possible solution and describe the impact, the subject consulted printouts of the data model of the data warehouse and once consulted the implemented data warehouse regarding the data cube structure. While describing the impact of the solution the subject doubted if his direction was correct, consulted a colleague by phone and then adjusted his solution and modified the described impact accordingly.

Since the subject was hesitant, the researcher emphasised several times to the subject that he should perform his work as he would under normal circumstances. This was done by for instance asking questions such as “What would you usually do now?” and “Is this what you would normally propose as the deliverable of the DIA?”. This last question possibly resulted in the subject adding a summary to beginning of the document.

The total time to perform the DIA took about 1 hour and 50 minutes.

Intervention (introducing the change)
The intervention occurred as intended in the requirements that were set. After the subject finished the DIA for the first case, certain information about the research was given to him:

- A Dutch translation of the management summary of this thesis to clarify the context (obviously without conclusions from this field study). See Appendix H.
- A document with a step-plan on how to perform a DIA (in Dutch). This step-plan described the seven steps that are proposed in this research, however it described them in a way that the subject could perform these steps without having to read or consult the actual research. See Appendix F.
- A template document for documenting the executed DIA that fits with the provided step-plan. See Appendix G.

The step-plan and template document were explained by going trough them and clarifying when unclear for the subject. About 1 hour was spent on the intervention.

Case 2
The second case concerned a request for change (RFC) initiated by the end-user concerning the reporting capabilities. Currently the end-user can distinguish certain costs by 4 different codes, which are based on 2 possible characteristics (each with 2 possible values) of the costs. In the future they want to be able to also distinguish costs on the values of just one of these characteristics.

The observed work approach of the subject can be summarized as:

- Assess the case.
- Describe the change.
- Describe the impact.
- Consider the different solutions, their advantages/benefits, choosing a solution.
- Describing the required modifications.
- Describing the required resources and costs.
- Checking if the document was suitable as a deliverable.

This approach reflects the steps presented in this research. During the research the researcher was involved more than during the first case in order to realise the fact that the model as proposed in this research was represented in the actions to perform the DIA:

- During the description of the change the researcher and subject discussed how to best perform this step. Since the provided case did not concern a delta regarding the source system situation the categorizations regarding schema, semantic and business rule delta were not applicable. The subject was asked to best describe the scope of the change, as is advised in the step-plan for information increasing delta.
- During the describing of the impact, the subject was instructed not to describe a specific solution yet, but to focus on the possible impact on elements related to the change.
- Regarding the describing of the solution space, the subject was emphasised not to just think of a solution and immediately elaborate on that for the remains of the DIA. Just as proposed in this research he was asked to consider if there were multiple solutions available and to weigh these against each other and then choose a solution.
- The subject found it difficult to assess the actual size of the impact with regard to work required. He was instructed to do this to the best of his ability based on common reasoning.

During the DIA the subject consulted printouts of the data model of the data warehouse and consulted the implemented data warehouse multiple times regarding the data cube, DMSA and BDW structure and related transformation packages.

The total time to perform the DIA was about 3 hours and 50 minutes. It should be noted however that this was party caused due to a mistake in the assessment of the impact, which required a repetition of part of the work.

Evaluation

After the subject finished the DIA for the second case, it was time to evaluate. During the evaluation the subject was given a document with evaluation questions. The expert was emailed the resulting documents of the two DIA's and a document with the evaluation questions. The evaluation consisted of a Dutch version of the questions as described in Section 4.7 and Appendix C. These questions allowed to reflect upon certain aspects by scoring on a 5-point scale (strongly disagree, disagree, neither agree nor disagree, agree, strongly agree) and also to gain feedback by asking open questions.

The practice learned that the results of a DIA are not that exact as assumed for the evaluation approach. The outcome of the DIA can be considered as a clarification of the scope of the change, impact and how it can be dealt with. Calculating ratio's such as the inclusiveness rates was not really possible, because clear assessments regarding the quantity of changing elements are not made. This might partly be due to the fact that it did not concern source system changes, where clear reasoning from changing elements would have been better possible.

5.1.5 Evaluation

The field study will be evaluated by discussing the remarks of the involved parties regarding the field study and by discussing the scores of the performed DIA's.

Researcher remarks

During the field study several things caught the attention of the researcher. First of all the lack of structure while performing the first DIA_I case. The subject started working in the RFC document, and thinking right away about what should be modified. The second thing that caught attention was the fact that the approach for DIA_I was solution-oriented from the start, little time was spent on clarifying the scope of the change and assessing the available options. The step-plan used for the DIA_M case resulted in a shift in the focus from solution-orientation to a focus on the process to understand the change and its possible impact. The third thing that caught attention was the fact that clear argumentation and clarification was lacking in the resulting document of DIA_I. Another interesting fact is that DIA_I was limited to the description of elements that should be changed; the required resources and costs were not considered. The last remark that can be made is that the time to perform DIA_M took about twice as much time than to perform DIA_I. This partly was caused due to a wrong assessment by the subject which required a repetition of part of the work, however it was also partly caused by the activities that were performed extra.

Subject remarks

Regarding the comparability of the results of the two cases it can be noted that the subject randomly mentioned the fact that the second case was similar to the first case. This supports the requirement of the field study setup to have two similar delta cases.

The subject was explicitly asked if he was able to perform the DIA and if it was required to deviate from the step-plan. The subject indicated that he did not deviate from the step-plan, he did however remark that it was difficult to assess the resources and costs and make a planning.

When asked to relate the performed DIA_M with DIA_I, the subject gave several remarks on executing a DIA with the use of the model. First of all he indicated that the model provides more structure. Which also, according to the subject, contributed to the recognition of a mistake in the assessment of the impact. Second, it stimulates thinking about the possible scenario's. Third, the performed DIA is more complete, including cost benefit analysis. And finally, the subject presumes that the resulting DIA document provides a clearer picture to the different possible audiences (builder, superior, customer).

Regarding the question how the DIA model can be adjusted to be improved for future use, the subject suggested that in the first step it should become more clear that a delta can have its origin in a source, but also in the requirements of the customer. Also he indicated that it would be useful to have a visual perspective on the DIA process. By process he did not mean the seven steps of the DIA model, but the data warehouse process of related components.

As a final comment the subject stated that in his opinion the use of the step-plan adds value to performing a DIA. However, it should be used and optimized to better fit practice in the future.

Expert remarks

From the survey questions several remarks by the expert are insightful. The expert was content with how for DIAM the scope was described clearly and the fact that it was transparent what the considerations were and which choices were made for what reasons. It shows that he methodology provided by the model is useable by people with less experience, resulting in better decision making. Also the DIAM approach resulted in far more insight in the required resources and costs for the expert. The expert thinks that the fact that a structure was provided benefited the accuracy of the estimation. However the expert also suggested that the model can be improved with respect to this aspect by developing it further to provide more guidance in how to estimate the required resources and costs. This is also important because this information is important in the final decision making.

As a final note the expert stated that it is clear that (relatively) much time is required to perform DIAM compared to DIAI, however in return for this we gain transparency.

Scoring of the DIA's

Table 10 and Table 11 show the outcome of the evaluation questions of the subject and the expert. The scores indicate the added value of the proposed DIA model. The values in the table are based on the 5-point scale that was used:

1 = strongly disagree
2 = disagree
3 = neither agree nor disagree
4 = agree
5 = strongly agree

Table 10. Evaluation by Subject

Measure	**DIAI**	**DIAM**	**Difference**
Feedback score Scope	4	4	0
Feedback score Resources & costs	1	4	+3
Feedback score Cost/benefit analysis	2	3	+1
Feedback score Communicate complexity of delta	3	4	+1
Average total feedback score	*2,5*	*3,75*	*+1,25*

Table 11. Evaluation by Expert

Measure	**DIAI**	**DIAM**	**Difference**
Feedback score Scope	3	4	+1
Feedback score Resources & costs	1	3,33	+2,33
Feedback score Cost/benefit analysis	2	5	+3
Feedback score Communicate complexity of delta	4	4	0
Average total feedback score	*2,5*	*4,08*	*+1,58*

5.1.6 Reflection

Baskerville distinguishes three types of learning experiences. [Bas99]. Each type will be discussed and reflected upon based on the insights gathered by the field study.

Double-loop learning

The first learning experience is that of *double-loop learning*, which concerns that what needs to be changed about the organization to reflect the new knowledge gained by the research.

Based on the scores and feedback by the subject it can be concluded that the use of a step-plan for the process of DIA adds value as opposed to the current ad-hoc style. Therefore the organization should be adjusted by actually implementing the model for the process of DIA. However the consideration of the trade-off between the added value and the extra time required should be made carefully; is the extra time required worth the added value of the results?

During the field study it became clear that emphasizing only a limited amount information to the subject had a large influence on the process. Possibly it is interesting to provide a set of guidelines for executing a DIA. These guidelines can be introduced in the organization standalone (as a 'light' version of the model) or together with the step-plan to perform DIA. An initial suggestion for a set of guidelines could be:

- First focus on the problem, not the solution: what is the change and its impact.
- Consider multiple solutions and weigh their advantages and disadvantages.
- Document and back up all considerations and decisions.
- Learn by trial and error: execute all steps of the process to the best of your capabilities and learn from this for the future.

Future iteration

The second learning experience is that what can be learned from this experience for future action research iterations. How can the model be improved for use in practice in the future?

The fact that the step-plan was still useful for cases that did not concern source system changes shows that the model can be expanded by supporting more scenario's than just source system changes. Adding for instance the scenario of a change in reporting requirement and how to better determine the impact for this scenario will make the model more complete and applicable to practice. Providing visuals with regard to how to assess the impacted components in the different scenario's will benefit the comprehension of the person performing the DIA.

Since it became clear that the subject found it difficult to assess the required resources and estimate the costs, and the expert also indicated this aspect should be developed further, it will be interesting to perform further research concerning resource and cost estimation models and how people can learn to better assess these aspects. The DIA model can then be expanded with this knowledge.

The model should be adjusted more to what is applicable in practice. The evaluation should be reduced to evaluation surveys that offer the possibility to score and reflect upon certain aspects of the DIA process and results.

Scientific community
The third learning experience is with regard to the importance to the scientific community if the research was a success or a failure, and if so, what was the cause of failure. This is knowledge that can be learned from for future research by others.

The fact that for this field study only one person was used to perform a DIA on two cases implies that one should not put too much value in the outcome of this field study. It can be used as a learning experience for both the model and the field study setup. The described field study setup can be adjusted with what was learned and applied in a more scientifically correct setup, by removing or reducing the limitations as discussed in Section 5.1.3. Including the fact that the actual content of the cases should have a better fit with the research topic: source system changes. This way the proposed categorizations for delta and impact can also be validated and reflected upon.

Nothing can be concluded with regard to hypotheses H1, H2 and H3, since it was not possible to calculate inclusiveness ratio's. The scores show that hypotheses H4, H5, H6 and H7 hold. Looking at the time spent on the performed DIA's it can be concluded that hypothesis H8 does not hold. However, this could be explained by the fact that the DIA is performed more thoroughly with the use of the model, which results in a higher quality of the outcome of the process.

5.2 Information Systems Research

In the original design of this research, the IS Research model of Hevner et al. is used to define the research strategy. [HMP04] Therefore a suitable validation of this research project will be to discuss how the performed research conforms to the guidelines set by Hevner et al. In each of the subsections a specific guideline will be reflected upon.

5.2.1 Problem relevance

Hevner emphasizes problem relevance by stating that this can be accomplished if (part of) the problem is solved by the research results. The initial research model described the problem as a gap between the current situation and the desired situation.

The research model was used as a red line throughout the research, which can also be noticed by the structure of this thesis. The problem gap was basically the fact that that there was no clear model for the process of Delta Impact Analysis. By presenting a valid model for the process in this thesis, it can be concluded that the gap has been reduced, which implies problem relevance. Even if the model could not have been validated, the research would still have been relevant since the gap would have been reduced by the knowledge gathered about what is not a suitable model for Delta Impact Analysis.

5.2.2 Research Rigor

Rigor in the research was realized by performing a thorough literature study taking current research as a basis, while realizing relevance by taking the practice at BI4U as a frame of reference. This resulted in solid contributions to the field of data warehousing based on both scientific research and practice. The ultimate acknowledgement of both rigor and relevant research was that the resulting Delta Impact Analysis model for data warehouses was evaluated in practice by a field study using criteria that are based on scientific research.

5.2.3 Design as a Search Process

One main problem statement was defined for this research. To solve the problem the research was divided into 5 sub-research questions. The relation between these questions and the main research question was specified in the research model. Answering each research question involved a process of searching for insights and literature. At certain times, by trying to answer one of the sub-research questions insight was gathered regarding another sub-research question, which resulted in adjusting the focus. This was especially the case with the 2nd and 3rd research question. This method of research did not result in 'the solution', nor are all possible solutions taken into account. The result is a proposed solution, the presented model, that reduces the main problem in some way. The search process for a better solution can continue as future research.

5.2.4 Design as an Artifact

The product of this research is the model for the process of Delta Impact Analysis as specified in Chapter 4. This model is based on both the practical situation at BI4U and insights gathered from the research field. Due to this, it is based on scientific theory, but still contains a fit with the organization. This model can be considered the artifact of this research.

5.2.5 Design Evaluation

The artifact of the research was evaluated by the performed field study. Hevner et al. state that a design artifact is effective when it satisfies the requirements and constraints of the problem it was intended to solve. Therefore several hypotheses were defined and tested if these held. The results of the field study show that the DIA model assists the person executing the DIA in the process and results in a better outcome of the process. From the field study it could also be concluded that more research and improvements to the artifact are desirable to provide a more solid solution that is useable in practice in the future.

5.2.6 Research Contributions

The research resulted in several contributions. First of all a perspective was gained on the practical situation concerning Delta Impact Analysis at BI4U. Besides providing a frame of reference for the research, this also added value for the company BI4U by gaining perspective on the current process of Delta Impact Analysis. A second contribution of the research is that different research and literature was studied and related to each other. This resulted in several categorizations for possible types of delta and their impact. The third contribution of the research is that the knowledge gained from the practice at BI4U and the literature on data warehouses, delta and impact is related to impact analysis

theory. This resulted in an impact analysis model for changes in the scope of data warehouse source systems, this is the final and most significant contribution of the research.

5.2.7 Communication of Research

This thesis was written while keeping in mind that the audience can be both technology-oriented and management-oriented. This is also taken into account for the final presentation of the research. The final thesis has been read by multiple people, including the supervisor at BI4U and the supervisors of the University of Twente, of which the first is part of the ‘Management & Governance’ faculty, and the other is part of the ‘Electrical Engineering, Mathematics and Computer Science’ faculty. Feedback from readers was processed to improve the readability of the thesis. While the ultimate assessment can only be done after this work has been read by multiple management- and technology-oriented people, based on current feedback it can be concluded that the research is accessible for both types of people.

5.2.8 Conclusion

The research project was reflected upon concerning all seven guidelines by Hevner et al. The reflection shows that all guidelines can be applied to the performed research. It can be concluded that the performed research was influenced by both practice and theory, which resulted in both rigor and relevance, the development of an artifact, and an evaluation of the developed artifact.

6 Conclusions & Recommendations

As a conclusion to this thesis the results of the research will be summarized and related to the initial research setup. After this it will be discussed what these results mean to BI4U and the research community.

6.1 Results

In the initial research plan several objectives were defined that resulted in the problem statement "*How can a Delta Impact Analysis model be designed that supports the process of analyzing the impact of changes in a data warehouse source system situation?*". To solve this problem, five research questions were defined in the initial research plan. The structure of this section is determined by reflecting the results of the research with regard to these research questions.

What are the essence and the current situation regarding the concrete composition of the process of Delta Impact Analysis as used within BI4U?
The knowledge gathered with regard to this objective is represented in the content of Chapter 2, which can be considered the documentation and clarification of data warehousing and DIA at BI4U. The contributions of this research were a definition for Delta Impact Analysis, a description of the process at BI4U and a consciousness development about what could be improved and researched about the topic. No clear guidelines existed at BI4U with regard to what tasks exactly were part of DIA, and how these should be performed. Also no clear framework was present concerning which types of changes could occur and which kind impact these could have.

What can be learned about change impact analysis methods regarding data warehouse / data integration systems by looking at literature, research and practice of others in the data warehouse field?
This resulted in the content of Chapter 3, which contains contributions of the research by providing a clear overview and classification types for the possible delta that can occur, the impact that can be caused and what the activities and goals of impact analysis are. The delta can be distinguished by its origin: schema, semantic or business rule. A *schema* delta concerns technical/structural changes to a source database, a *semantic* delta concerns changes to the meaning or representation of data in the source, and a *business rule* delta concerns changes to the meaning of source data for the data warehouse. The impact can be distinguished as: *information-preserving* or *information-changing*. The 'changing' type can be divided by an information supply to the data warehouse that has been *reduced*, *increased* and *redefined*. For the different delta categories it is specified what impact categories can apply, this provides a bridge between defining a delta and determining the impact in impact analysis. To deal with the different types of information-changing impact one has the choice if the data warehouse is *not modified*, *modified* and/or *modified with regard to historic data*.

How can a model be developed to describe the process of Delta Impact Analysis that can be tested at BI4U?

The previous research questions can be considered the means to achieve the goal of this research question, which is also how it is presented in the initial research model. Answering this question resulted in the most significant contribution of this research is by combining and applying the results of the previous research questions in an artifact. This artifact is a model for the process of Delta Impact Analysis. The model describes seven main steps that should be performed to execute a solid DIA:

1. Determining the change set (primary impact set)
2. Determining the impact set (secondary impact set)
3. Describe solution space
4. Determining the required modifications to the data warehouse
5. Determining the resources and costs required to perform the modifications
6. Communicating the results of DIA
7. Evaluating the DIA process

The usability of the model in practice was determined by the final two objectives.

How can the performance of the developed DIA model be evaluated in terms of efficiency and quality and what are the requirements to this performance?
The research proposes several measures to evaluate DIA in terms of performance. Several metrics regarding the adequacy and effectiveness of a performed impact analysis were found in impact analysis literature. This resulted in the use of the metrics of *correct-*, *wrong* and *missed-inclusiveness* to evaluate the adequacy. The metrics for the effectiveness were inspired by the established goals of performing an impact analysis (clarify scope, estimate resources and costs, ability to weigh cost and benefits, communicate complexity of a change) and translated into a survey to assess these metrics. The performance can be evaluated by performing a field study: performing DIA with and without use of the model and scoring these on the defined metrics.

Using the defined metrics to measure performance, what is the performance of the developed DIA model? And, if necessary, how should it be adjusted to be useable in practice?
While the field study shows that the performance of DIA with the use of the DIA model is considered better than without use of the model, not too much value can be given to this fact. To be able to truly conclude this, a more thorough field study will need to be executed.

What is more important is what was can be learnt from the field study. First of all it is suggested that a set of guidelines for executing a DIA is created. These guidelines can be introduced in the organization as a standalone solution or together with the step-plan to perform DIA. It is however advised that the organization should introduce the complete step-plan for DIA. The second thing that can be learnt is that expanding the model by supporting more scenario's than only source system changes will make the model more complete and applicable to practice. Providing visuals to assess the impact on data warehouse components is also considered a valuable contribution to perform DIA. Furthermore it will be interesting to perform future research concerning resource and cost estimation models and how people can learn to better assess these aspects. Finally the

evaluation should be reduced to evaluation surveys that offer the possibility to score and reflect upon certain aspects of the DIA process and results.

What was not defined as an objective of the research, but what is a relevant result of the research, is the validation of the research in Section 5.2 by reflecting on the research framework of Hevner et al. that was discussed in the initial research plan. This means that this research is not only validated according to practice, but also according to a scientific research framework. From this reflection it can be concluded that the research was performed according to the guidelines of the framework; influenced by both practice and theory, which resulted in both rigor and relevance, the development of an artifact, and an evaluation of the developed artifact.

6.2 Discussion and future research

The result of this research, the model for the process of Delta Impact Analysis, provides a solid starting point for tackling source system change situations for BI4U, and also for others in the field of data warehousing. It was expected that there would have been several opportunities to experience practical cases where Delta Impact Analysis was performed. It should be noted that this was not the case during the development of the model and that most practical knowledge was gathered by discussing the concept with a DIA expert at BI4U, and researching existing documents of the company. However the fact that the model was eventually validated in practice by a field study demonstrates its value for companies active in the field of data warehousing. However it is likely that the model should be adjusted more to provide a better fit with practice. Therefore the model should be consistently used for Delta Impact Analysis and adjusted according to the outcome of the evaluation of the process.

It was established that visualization of traceability information is important. Several tools were mentioned in the research that can support the process of DIA by providing more insights in the technical impact of a source change: Red Gate SQL Compare, SQL Data Compare and Dependency Tracker. However these tools were not further examined for this research. It will be interesting to perform future research on the usability of existing tools and/or the development of a tool to support steps of the DIA process. The topic of a tool for the visualisation of DIA and Lineage is currently being researched by another intern at BI4U.

Little research has been done in the field of Delta Impact Analysis. The proposed model is a mixture of several findings in the literature from different fields. In certain aspects the model might lack a certain detail for specific tasks. For instance more research could be done concerning which cost estimation models and techniques could be used for DIA. Also the scope of the research was limited to changes in data warehouse source systems with a relational database, changes in the data warehouse can also occur with other types of sources and at other levels than the source. It will be interesting to research how the DIA model can be expanded to eventually become a complete model to deal with any type of change in implemented data warehouse solutions.

Besides just researching the current situation and literature relevant to developing a model for the Delta Impact Analysis process, the objective to look at improvements was defined. This objective resulted in a sensitive attitude with regard to opportunities to improve the process of DIA. During the research it became clear that this objective led to a situation where the results of the main research became intertwined with extra information that was not essential to developing a model for the process of DIA. This resulted in some sort of an information overload. Therefore it was decided to conclude this thesis by discussing suggestions for future research for the topics that are not essential to the core of the research, but are to this defined objective. The topics are *Data warehouse evolution*, which makes it possible to distinguish different versions of the data warehouse structure over time, and *Meta warehouse*, which provides support to document all relations between the different elements and components in the data warehouse.

6.2.1 Data warehouse evolution

Golfarelli et al. discuss the topic of data warehouse evolution. They distinguish two types of evolution: on the *data* (extensional) level and on the *schema* (intentional) level. [GLR04] [GLR06] Regarding the *data* level often it is incorrectly assumed that dimensions are static and they are defined without a relation to time. One can take the example of countries that are used as a dimension in a data warehouse. The world around us changes and over time The Federal Republic of Germany (FRG) and the German Democratic Republic (GDR) became one Germany. The country Czechoslovakia was split into Czech and Slovakia. [EKK02] Evolution on the data level is not further discussed, since this is outside the scope of Delta Impact Analysis. Evolution on the *schema* level can occur due to new insights, business needs etc. [GLR04] With regard to Delta Impact Analysis, a delta in a source system schema, semantics or a business rule can result in evolution of the data warehouse on the schema level.

According to Golfarelli et al. [GLR06], the solutions provided by research regarding data warehouse evolution often do not support versioning and those that do, do not discuss how querying across multiple versions can be done. Therefore the authors provide an approach to evolution on the schema level by supporting schema versioning and cross version querying in data warehouses. The authors discuss the evolution of schemata defined using their DFM formalism. They propose support for evolution by discussing three aspects: schema graphs, augmented schemata, and schema histories.

A schema graph is a graph of the schema that shows the functional dependencies between the different elements of the schema. The authors define four modification operations to manipulate a schema. [GLR06] They introduce $Add_A()$, $Del_A()$, $Add_F()$ and $Del_F()$, where A stands for attribute and F stands for functional dependency. These operations can be used to add and delete attributes and functional dependencies to/from a schema.

A version as defined by the authors is a schema that represents the business requirements at a given time. [GLR06] Using a sequence of these operations combined can be used to distinguish different versions of a schema, e.g. $S_{new} = New(S, Add_A(x), Add_F(y \rightarrow x))$, where S is the current schema, and x a new attribute that is added to the schema. The sequence of operations that lead to a new version of the schema is called a transaction.

An augmented schema is a variation of a specific schema version that makes an older schema compatible to a newer schema. [GLR06] Usually an augmented schema is created at the time of a change to make the old schema compatible with the current schema. The designer can choose which differences between the schema versions are propagated to the augmented schema. The authors state that creating an augmented schema is only relevant for add operations. Propagating Del operations makes no sense in their point of view, since they assume querying functionality that takes care that old schema versions can still be queried without a problem.

Consider a schema S that has already been changed over time, specifically at t1 and t2, respectively producing versions S1 and S2. The augmented schema of S1 can differ from schema S2, if not all changes have been propagated to the augmented schema of S1 at the time S2 was introduced. At current time, t3, a schema transaction is performed, and assuming that the designer chooses to propagate all changes to previous versions of the schema, the augmented schemata of S1 and S2 can be defined by adding all attributes and functional dependencies that have been introduced in 'S2 → S3' to the already existing augmented schema(ta). Figure 24 provides an abstract example of this, augmenting the changes from 'S2 → S3' would imply adding al attributes and functional dependencies concerned with the introduction of H to the old schemata.

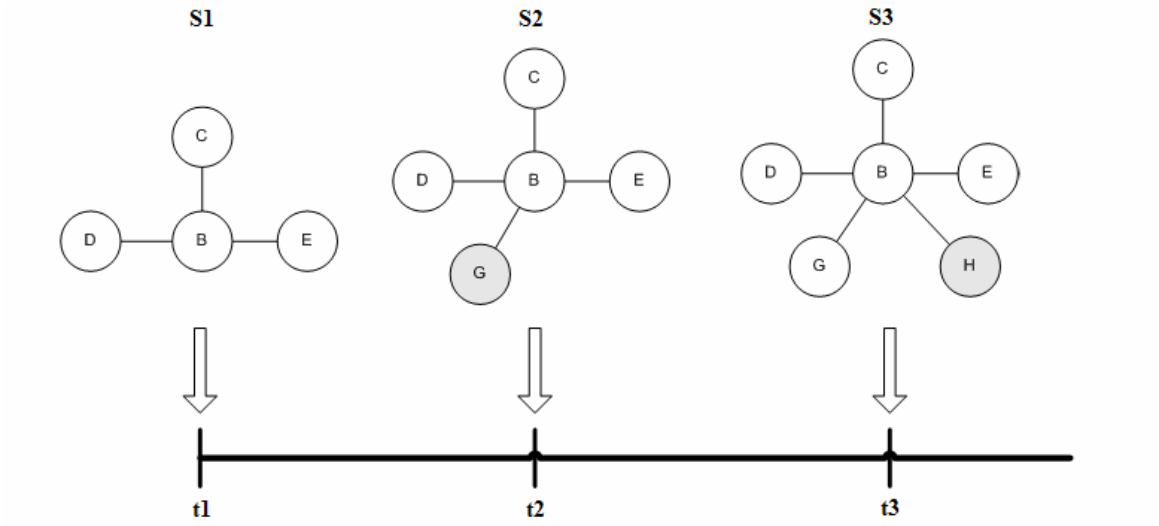


Figure 24. Schema versions at t1, t2 and t3

Assuming the changes of S2 (G) were not augmented to S1 and the changes of S3 (H) were augmented to both S1 and S2, Figure 25 provides a graphic overview of the augmented schemata at t3.

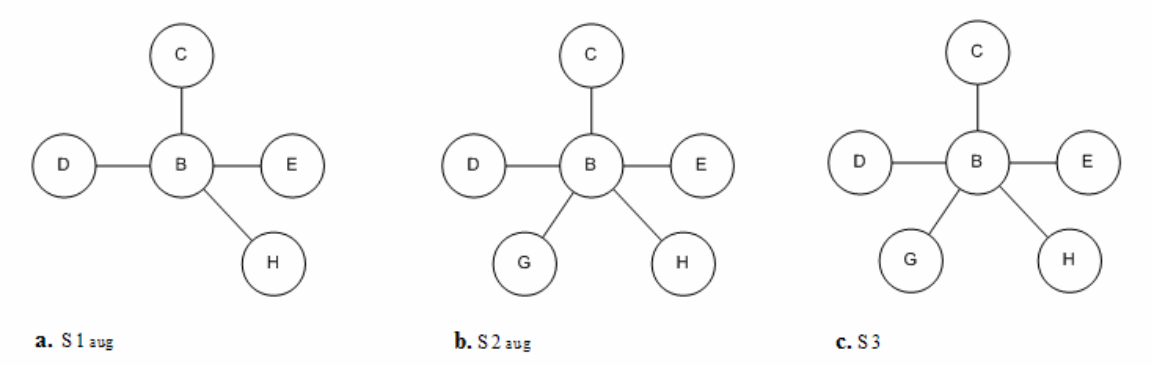


Figure 25. Schemata at t3

Querying the data warehouse concerning data applicable to time interval T, where in interval T multiple schema versions apply, is possible on the intersection of the schemata. Thus an information request regarding data concerning both t1, t2 and t3, is possible on the elements that are present in '$S1_{aug} \otimes S2_{aug} \otimes S3$', an information request regarding data that concerns both t2 and t3, but not t1, is possible on the elements that are present in '$S2_{aug} \otimes S3$'. Figure 26 provides graphic insight as to what is query able with these augmented schemata.

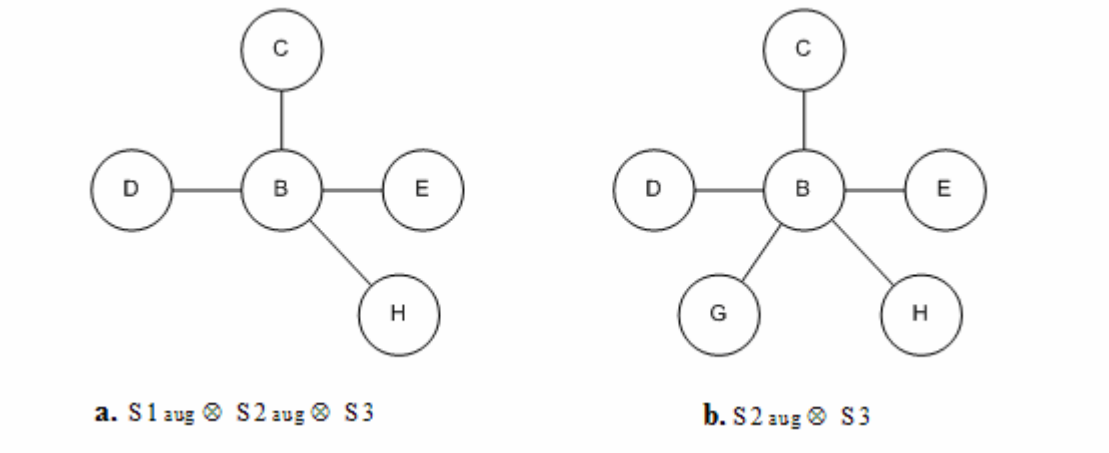


Figure 26. Querying multiple schema versions at t3

To support querying functionality over multiple versions Golfarelli et al. suggest two possibilities for the querying interface: implicit and explicit. [GLR04] Where in the implicit case given a time interval T, the widest intersection of the schema is calculated and presented to the user as query able. In the explicit case the user chooses what to query, thus the required intersection of schemata, and the time interval T is calculated over which the user can query this information.

The approach of Golfarelli et al. bares some resemblance to what Fan and Poulovassilis discuss (3.6.1, 3.6.2). [FP03] [FP04] With the provided versioning support it would also be possible to perform lineage as mentioned by Fan and Poulovassilis, i.e. where does the data in reports come from, and based on which business rules at which time. Fan and Poulovassilis discuss the transformation from one schema to another on a more abstract

level. Their approach can be used to not only define transformations from one schema version to another version (schema evolution), but also to define the transformation from a schema of one system/component to another (e.g. DSA -> ISA). While this provides a complete solution, this does make it more complex to actually implement and maintain such a solution.

Most of this research project concerned how to deal with changes (evolution) of sources, and the concept of data warehouse evolution has been briefly mentioned at some point. Data warehouse evolution supports the distinction between different versions of the data warehouse at different moments in time. As clarified in Section 3.6.3 and Section 4.4 (Table 6 and Table 7), this can reduce the impact of an information-changing delta, and also provides more insights of introduced delta of the past. Even though it is not a necessity for performing Delta Impact Analysis, it is recommended to be integrated in the data warehouse solution because reducing the complexity of possible impacts will also reduce the complexity of performing a DIA. Therefore it is relevant to further research the topic, the complexity that it comes with it, and how it can be implemented in practice.

6.2.2 Meta warehouse

Vaduva and Dittrich claim that in order to withstand the increase in data warehouse processes and management complexity, the consistent management of relevant metadata is required. In their article they provide an overview of metadata management; what kind of information about the data needs to be generated and kept for data warehouses and how should this be stored and managed. Besides this they also discuss the weaknesses that come with partial solutions for metadata management. The complexity of a data warehouse grows with the amount of data sources, the level of heterogeneity, the diversity in loading the data warehouse and how many applications are using the data warehouse. Metadata management can help to oversee this growing complexity. [VD01]

Types of meta data

Vaduva and Dittrich make a distinction between the type of application for the meta data: *business* metadata, for end-user use, and *technical* metadata, for database administrators and technical modules of the data warehouse. [VD01] Stöhr et al. call this semantic metadata and technical metadata and explain that this distinction is mainly made to support the different types of people working with it (technical and non-technical). [SMR99] Another distinction is made between between, *descriptive metadata*, information related to the data structures of the sources and the data warehouse, and *transformational* meta data, information regarding the transformation of data. [VD01] With regard to Delta Impact Analysis descriptive metadata is relevant, since this concerns the structures of data sources, but transformational metadata can be relevant since the relations between the different elements of the data warehouse can be deduced from this information.

Chaudhuri and Dayal distinguish three kinds of meta data that need to be managed: *administrative*, *business* and *operational*. Administrative contains al metadata that is relevant for the implementation and use of a data warehouse, such as data structures and schema's, transformations, but also user and access profiles. [CD97] With respect to

Veduva and Dittrich and Stöhr et al. this is similar to their definition of technical metadata. Business meta data is similar as to how the other authors defined this, data for end-users such as business definitions and terms and the meaning of information provided by the data warehouse. Operational data is a type not discussed by the other authors, this is information collected during the operational activities of the data warehouse. This includes data about the executed ETL processes, but also about the usage of the data warehouse by end users.

Vaduva and Dittrich specify several applications that meta data is useful for. Regarding DIA it can be useful for both *documentation* and *control information*. Documentation can be done for both business and technical information, but especially technical documentation is interesting when manual work is required for DIA, someone can use it to determine what is impacted if a certain source system is changed or removed. Control information can contain structural information, but also application logic. This type of meta data can be used to be interpreted at runtime and dynamically bound into software execution. In other words, if this meta data is defined in such a structured and specific way it can be used by the data warehouse processes to determine how to extract, transform and load data from the source systems to the data warehouse. [VD01]

Meta data repository

Stöhr et al. discuss the need for good metadata management. They state that intra- and inter-dependencies between the different components of systems can become so complex that it is impossible to manage this if this is represented no where else than in the actual implementation (source code) level. They state three basic requirements for a metadata repository: support collection, provide uniform representation, have the characteristics of being bi-directional. It must support the collection of technical data, i.e. by automated analyzers and user input functionality. The representation of different types of metadata should be uniform and possible to export to a standardized form. Bi-directional characteristics can be achieved to support functionality to interpret the metadata, but also to change the metadata and propagate this to data warehouse. [SMR99]

Stöhr et al. propose a model for metadata management that is both accessible for programmers but also administrators which allows a better way to use metadata for all components of the data warehouse. This information can function as both documentation and control information. Figure 27 shows the UML model of the main classes. For a detailed and more in depth explanation one should refer to the original article. The model defines *entities*, which can be instantiated as a relational table, an object-oriented class, a business concept etc. An entity has one or more *attributes* that define the entity, i.e. the columns of a relational table. *Transformations* of the class non-aggregation or aggregation can be instantiated with a filter and/or a function on one or more related attributes. One or more *transformations* are part of a mapping that defines relations on the entity level. [SMR99]

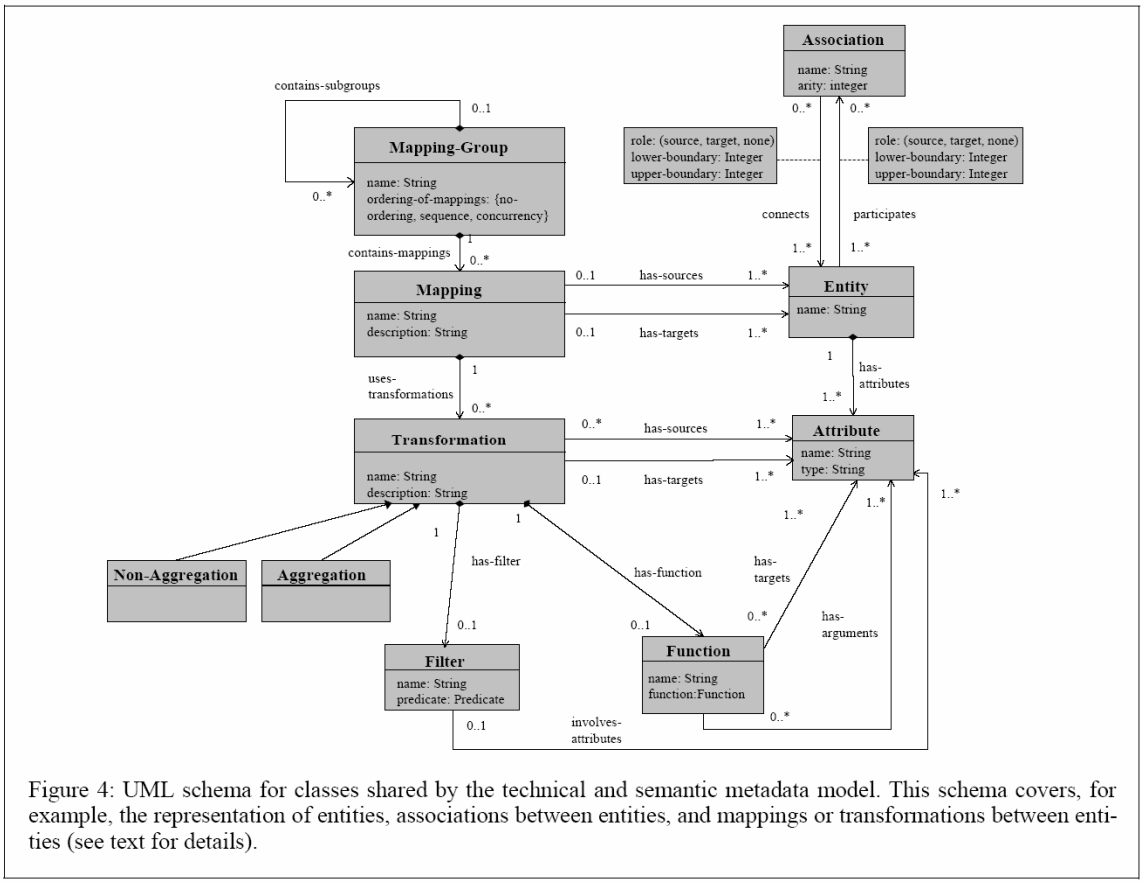


Figure 4: UML schema for classes shared by the technical and semantic metadata model. This schema covers, for example, the representation of entities, associations between entities, and mappings or transformations between entities (see text for details).

Figure 27. UML schema Data Warehouse metadata model [SMR99]

While their model as a whole is a wise and useful approach to a full blown meta warehouse solution, for this research only the perspective of their model that is relevant to support the process of Delta Impact Analysis is discussed. In order to support the process of Delta Impact Analysis the meta warehouse can function as a source for traceability information. As Stöhr et al. discuss the model provides the functionality to both define and describe intra-system and inter-system dependencies, which can be related to the concepts of respectively vertical and horizontal traceability. [SMR99] The model however does not support the approach of schema transformations as suggested by Fan and Poulovassilis [FP03], nor does it support schema evolution. Schema evolution in the meta warehouse could be realized by doing something similar to the concepts of $Add_A()$, $Del_A()$, $Add_F()$ and $Del_F()$ as discussed by Golfarelli (Section 6.2.1) [GLR06]. This could be realized by adding timestamps that mark from when to when the element was in the schema. In order to support augmentation as suggested by Golfarelli [GLR06], another timestamp that indicates till which version the change was augmented. If the change was not augmented to earlier versions of the schema then the timestamp would be equal to the timestamp that indicates when this element was added to the schema.

BI4U designed a data model for the development of a data warehouse. The model shows the components of DSA, ISA, DMSA, Cube and Report, their type of elements, and the relation between the possible elements of the components. It can be recommended that

this model is expanded with support for data warehouse evolution by storing timestamps that mark information concerning the adding, deleting en augmenting of elements in the data warehouse. Also the model defines different entities for every component, which makes it less generic than the model as proposed by Stöhr et al. It is recommended to make the data model more generic and to define the entities of the components on the level of application logic. Concerning the actual development of the meta warehouse the discussed three basic requirements for a metadata repository should be taken into account: support collection, provide uniform representation, have the characteristics of being bi-directional.

References

[AB93] R. S. Arnold, S.A. Bohner; *Impact Analysis - Towards a framework for comparison*; Conference on Software Maintenance, 1993; p292-301

[Art88] L.J. Arthur, *Software Evolution: the software maintenance challenge*; John Wiley; New York; 1988; ISBN 0-471-62871-9

[Ari93] A. Arifoglu; *A methodology for software cost estimation*; ACM SIGSOFT Software Engineering Notes; Volume 18, Issue 2, April 1993: p96-105

[AS06] S. W. Ambler, P.J. Sadalage; *Refactoring Databases*; Addison-Wesley, USA, 2006; ISBN 0-321-29353-3

[Bas99] R.L. Baskerville; *Investigating Information systems with Action Research*; Communications of the Association for Information Systems, Volume 2, Article 19, 1999; p1-32

[BCN91] C. Batini, S. Ceri, S.B. Navathe; *Conceptual Database Design*; The Benjamin/Cummings Publishing Company, Inc., 1991; ISBN 0-8053-0244-1

[BI06a] Business Intelligence For You (BI4U); *Delta Impact Analyse*; internal Document Business Intelligence B.V; 27/12/2006.

[BI06b] Business Intelligence For You (BI4U); *Diagram Software Factory*; internal Document Business Intelligence B.V.

[BI07a] Business Intelligence For You (BI4U); *Naslagwerk Integration services*; internal Document Business Intelligence B.V; February 2007.

[BI07b] Business Intelligence For You (BI4U); *BI4U Website*; visited 03/05/2007; < http://www.bi4u.nl >

[CB05] T. Connolly, C. Begg; *Database Systems: a practical approach to Design, Implementation and Management*; Fourth edition, 2005; Harlow - Pearson Education Limited, ISBN 0-321-21025-5.

[CD97] S. Chaudhuri, U. Dayal; *An Overview of Data Warehousing and OLAP Technology*; ACM Special Interest Group on Management of Data, Volume 26, Issue 1, March 1997; p65-74.

[CZR06] S. Chen, X. Zhang, E.A. Rundensteiner; *A compensation-based approach for View Maintenance in Distributed Environments*; IEEE Transactions on knowledge and data engineering; Vol. 18, No. 8, August 2006; p1068-1081

[EKK02] J. Eder, C. Koncilia, H. Kogler; *Temporal Data Warehousing: Business Cases and Solutions*; Proceedings of the 4th ECEIS, Ciudad Real, Spain, April 2002; p81-88

[EWD06] S.M. Embury, D. Willmor, L. Dang; *Assessing Impacts of Changes to Business Rules through Data Exploration*; International Conference on Software Engineering, 2006; p21

[FP03] H. Fan, A. Poulovassilis; *Using AutoMed Metadata in Data Warehousing Environments*; Proceedings of the 6th ACM international workshop on Data warehousing and OLAP; New Orleans, Louisiana, USA, 2003; p86-93

[FP04] H. Fan, A. Poulovassilis; *Schema Evolution in Data Warehousing Environments – A Schema Transformation-Based Approach*; Lecture Notes In Computer Science; Vol. 3288, Conceptual modeling, 2004; p639-653

[FV99] A.R. Fasolino, G. Visaggio; *Improving Software Comprehension through an Automated Dependency Tracer*; 7th international workshop on program comprehension; 1999; p58-65

[GLR04] M. Golfarelli, J. Lechtenbörger, S. Rizzi, G. Vossen; *Schema versioning in data warehouses*; Lecture Notes In Computer Science; Vol. 3289, Conceptual Modeling for Advanced Application Domains, 2004; p415-428

[GLR06] M. Golfarelli, J. Lechtenbörger, S. Rizzi, G. Vossen; *Schema versioning in data warehouses: Enabling cross-version querying via schema augmentation*; Data & knowledge engineering; Volume 59, Issue 2, 2006; p435-459

[HMP04] A. Hevner, S. March, J. Park; *Design science in Information Systems Research*; MIS Quarterly, Vol. 28, No. 1, March 2004; p75-105.

[HRU96] V. Harinarayan, A. Rajaraman, J. D. Ullman; *Implementing data cubes efficiently*; Proceedings of the 1996 ACM SIGMOD international conference on Management of data; Montreal, Quebec, Canada, 1996; p205-216

[HW06] H. Heerkens, R.J. Wieringa; *The Methodological soundness of requirements engineering papers: A conceptual framework and two case studies*; Requirements Engineering, Vol. 11, No. 4, August 2006; p295-307

[Lee07] A.S. Lee; *Action is an artifact*; Integrated Series in IS: Information Systems Action Research; Volume 13, February 2007; p43-60

[Lin97] M. Lindvall; *Evaluating Impact Analysis - A case study*; Empirical Software Engineering, Volume 2, Issue 2, 1997; p152-158

[Lin03] M. Lindvall; *Impact Analysis in Software Engineering*; Advances in Computers; Volume 59; 2003; p127 –p210

[LL01] Lethbridge and Laganière ; *Object-Oriented Software Engineering*; McGraw-Hill Education, Berkshire, England, 2001; ISBN 0-07-709761-0

[LOC99] S. Lock, A. Rashid, P. Sawyer, G. Kotonya; *Systematic Change Impact Determination in Complex Object Database Schemata;* Lecture Notes In Computer Science; Vol. 1743, Proceedings of the Workshop on Object-Oriented Technology; 1999; ISBN:3-540-66954-X

[LS98] M. Lindvall, K. Sandahl; *Traceability aspects of impact analysis in object-oriented systems*; Software maintenance: research and practice, 1998, Volume 10; p37-57

[Mar00] A. Marotta; *Data warehouse design and maintenance through schema transformations;* Master Thesis, 2000; Instituto de Computacion, Facultad de Ingeneria, Universidad de la Republica. Montevideo – Uruguay.

[Mar07] D. Marco. *Meta Data & Knowledge Management: Managed Meta Data Environment: A Complete Walk-Through*, Part 2, DM Review, 2004

[MJ04] D. Marco, M. Jennings; *Universal Meta Models*; Wiley Publishing Inc. Indianapolis, USA, 2004; ISBN 0-471-08177-9

[MR02] A. Marotta, R. Ruggia; *Data warehouse Design: A schema-transformation approach*; Proceedings of the 22nd International conference of the Chilean Computer Science Society, 2002; p153-161

[MTK06] J. Mundy, W. Thornthwaite, R. Kimball; *The Microsoft Data Warehouse Toolkit*; Wiley, USA, 2006; ISBN 0-471-26715-5;

[OF04] J. Oosterhuis-Geers, L. Ferreira Pires; *Final Project Guide*; University of Twente, Department of Computer Science; 13/07/2004

[PER03] V. Peralta, A. Illarze, R. Ruggia, *Towards the Automation of Data Warehouse Logical Design: a Rule-Based Approach*. CAiSE Short Paper Proceedings 2003; p21-24

[RKZ00] E.A. Rundensteiner, A. Koeller, X. Zhang; *Maintaining Data Warehouses over Changing Information Sources*; Communications of the ACM; Volume 43, No. 6, 2000; p57-62

[SMR99] T. Stöhr, R. Müller, E. Rahm; *An Integrative and Uniform Model for Metadata Management in Data Warehousing Environments*; Proceedings of the International Workshop on Design and Management of Data Warehouses, Germany, Heidelberg, June 1999; p12:1 -12:16

[Sne01] H.M. Sneed; *Impact Analysis of Maintenance Tasks for a Distributed Object-oriented System*; International Conference on Software Maintenance, 2001; p180-189

[SS05] A. Sen, A.P. Sinha; *A Comparison of data warehousing methodologies*; Communications of the ACM, March 2005, Vol. 38, No. 3; p79-84

[VD01] A. Vaduva, K R. Dittrich; *Metadata management for data warehousing: between vision and reality*; International Symposium on Database Engineering & Applications, 2001; p129-135

[Ver98] P. Verschuren; *Het ontwerpen van een onderzoek;* Lemma, Utrecht, The Netherlands, 1998; ISBN 90-5189-707-3

Appendix A.: Orientating interview with Shirly Dijkstra

Interviewee:

Name: Shirly Dijkstra
Organization: BI4U
Function: Information Analyst
Date: 29/05/2007

Introduction

The summary of this interview contains information based on a designed interview with specific questions regarding the following topics:

- Which situations require a DIA
- The process of DIA at BI4U: who is involved, way of working, deliverables
- Discussion of the internal DIA document [BI06a]

Besides the information gained in the structured interview, this summary also contains some more information from other conversations held in the beginning stages of the research project.

Situations that require a DIA

Mainly a DIA is required in case of source system updates where the database structure is changed. But it could also be that a new system is added to the data warehouse.

The implementation of the generic data warehouse model at a new customer could also be considered as something that requires a Delta Impact Analysis. Basically it is an addition of multiple source systems to an existing data warehouse situation, that is the generic model.

The process of DIA at BI4U

In most cases a change in a data sources impacts the part of the data warehouse that is specific for that implementation. The generic part/model of the data warehouse as designed by BI4U for use in multiple implementations is only impacted if 'the world changes'. For instance if new source data becomes available and provides opportunities for new information analysis.

The Delta Impact Analysis is performed by the employees based on their personal insights. The internal DIA document seems to describe the process correctly, but it is not used or referenced to during the actual execution of a DIA.

Discussion of the internal DIA

The internal DIA document lacks concreteness in several aspects. The document contains a model that explains the various tasks of the Delta Impact Analysis process. The following was confirmed regarding several of these tasks:

- "*Has change been sufficiently described*": to the knowledge of the interviewee there are no clearly defined criteria for this task. Usually this task is performed to own insight and by asking one self the question "Can I, or my colleague, continue from this point on with the next task?"
- "*Estimate modifications*" (..to data warehouse): there is no clear work process available or criteria to what this estimate should meet. The employee estimates complexity of the changes based on personal insights.
- "*Estimate consequences":* there is no clear work process available or criteria to what this estimate should meet.

Appendix B.: DIA interview with Marc van der Wielen

Interviewee:

Name: Marc van der Wielen
Organization: BI4U
Function: Consultant
Date: 17/07/2007

Introduction

The interview was held informally without a list of questions. The intention of the interview was to discuss Delta Impact Analysis regarding the possible impact on the various components of the data warehouse. The conversation was based on a very minor DIA that was performed on a data warehouse implementation.

The data warehouse components

The data warehouse consists of various components:

- DSA: this is a one-on-one copy of source tables that are relevant to the data warehouse
- ISA: the data in the data warehouse is transformed to the generic business data warehouse model according to the business rules defined for the information needs.
- BDW: this is the actual data warehouse stored in a relational database
- DMSA: the data warehouse represented in the structure of dimension and fact tables
- Cubes: OLAP cubes that are intelligent views on the data stored in the DMSA in order to optimize generation of reports
- Reports: The actual reports that are generated, based on the information requirements of the customer

The performed DIA

The customer reported a malfunctioning in the loading of the data warehouse, after which it contacted BI4U to solve this malfunctioning. The customer then also mentioned that the source system that is used by the data warehouse was updated.

The procedure followed by Marc van der Wielen then was as follows:

- Red Gate SQL Compare tool was used to analyze the differences in strucutures between the old and the new situation
- Many new tables were added compared to the old structure. These additions were ignored, since this can not result in the malfunctioning of the data warehouse
- Several existing tables were changed compared to the old situation. Looking at the business data model and DSA model of the data warehouse implementation he analyzed which tables were modified that were also present in the DSA. Also the error logs of the loading of the data warehouse were studied.

- This resulted in one table, where one column was added. This column was then added to the DSA and the packages that load the DSA and ISA were modified to be able to deal with this added column.

Types of Delta Impact Analysis

Marc van der Wielen clarified that one can make the distinction between technical and functional DIA. Where technical DIA concerns changes in the technical structure of the source systems and where functional DIA concerns functional changes to the data warehouse. He specified four sorts of changes that can require a functional DIA:

- Business rules: changes in the business rules of how certain information is gathered/calculated from source data
- Method of storing in source: change in how data is stored in the source database that do not involve changing in the structure. For instance if the meaning of a certain column changes, or the way the data is stored in a column is changed.
- Additions to generic business data warehouse model: if the including of more (components of) source systems in the data warehouse is desired or can create more intelligence in the data warehouse, the business data warehouse model needs to be adjusted. The impact of this should also be investigated
- Modify generic business data warehouse model: in some cases the model needs to be reorganized and changed. The impact of this should also be investigated

Technical DIA can always be solved by modifying the DSA-to-ISA components, however possibly it will be interesting to perform a functional DIA to determine if other components should be adjusted to create more intelligence in the data warehouse.

Meta warehouse

Marc van der Wielen stated that for future DIA's it will be useful to research and determine which pieces of information are required to perform a DIA and determine the impact of a specific change. BI4U has already developed a model for a meta warehouse in which the mappings between attributes between the different layers in the data warehouse can be defined. It will be interesting to compare the elements of this model with the identified required information to perform a DIA in the research.

Appendix C.: Survey questions to evaluate DIA

One goal of impact analysis is to gain and communicate a clear perspective the scope of the change. Please indicate how well you would value the detail in which the scope was described, and how complete the information in the described scope was:

	strongly disagree	**disagree**	**neither agree nor disagree**	**agree**	**strongly agree**
The scope of the change is described in sufficient detail.					
The scope of the change is described complete.					

Another goal of an impact analysis is to provide insight into the required resources and costs before implementing a change. Please value the provided overview of resources and costs with regard to how detailed they were described, if the overview was complete and how accurate the overview was compared to the actual resources and costs that were required for implementation:

	strongly disagree	**disagree**	**neither agree nor disagree**	**agree**	**strongly agree**
The required resources and costs are described in sufficient detail.					
The required resources and costs are described complete.					
The described required resources and costs are accurate compared to the actual situation.					

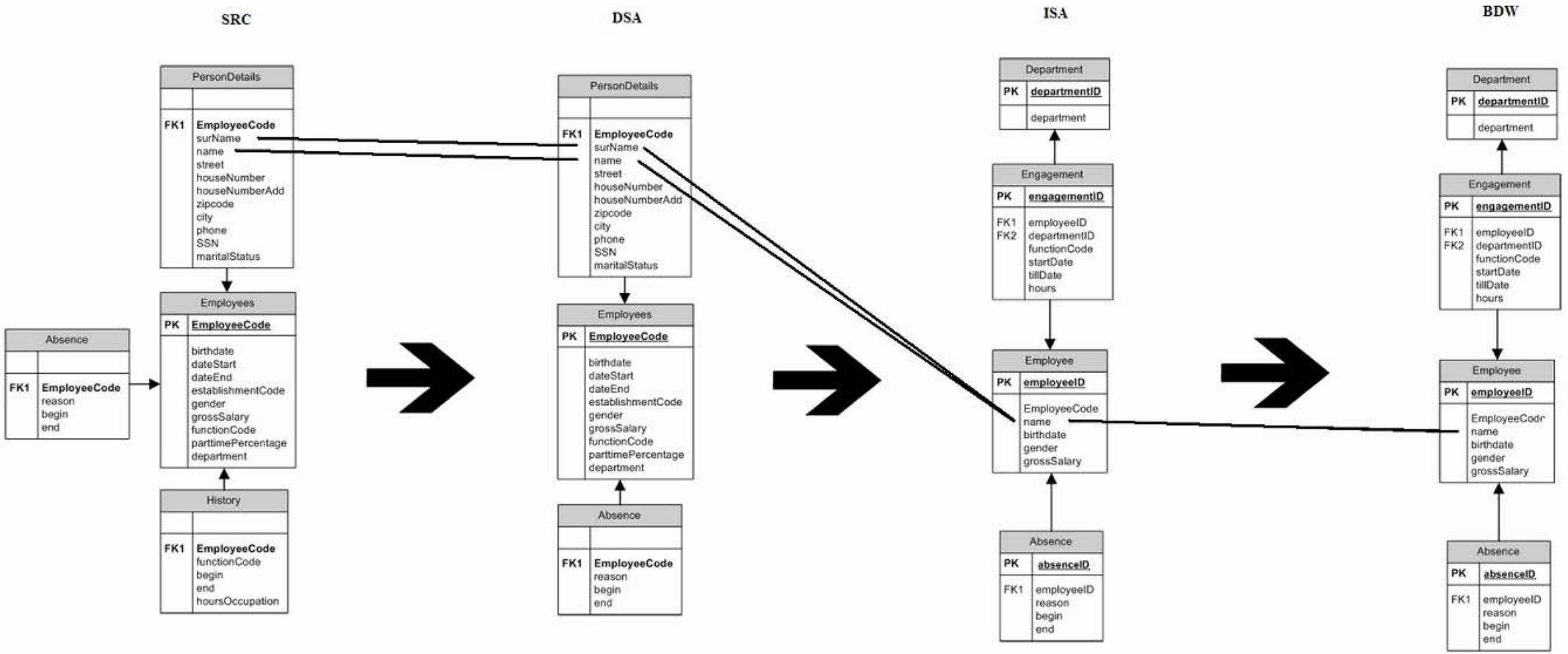
Another goal of an impact analysis is to be able to perform a cost benefit analysis based on the provided information: what are the benefits of implementing a solution, what are the costs, and is it worth it? Please indicate how well you are able to make this consideration using the information from the DIA:

	strongly disagree	**disagree**	**neither agree nor disagree**	**agree**	**strongly agree**
I am able to weigh the cost and benefits of implementing a solution.					

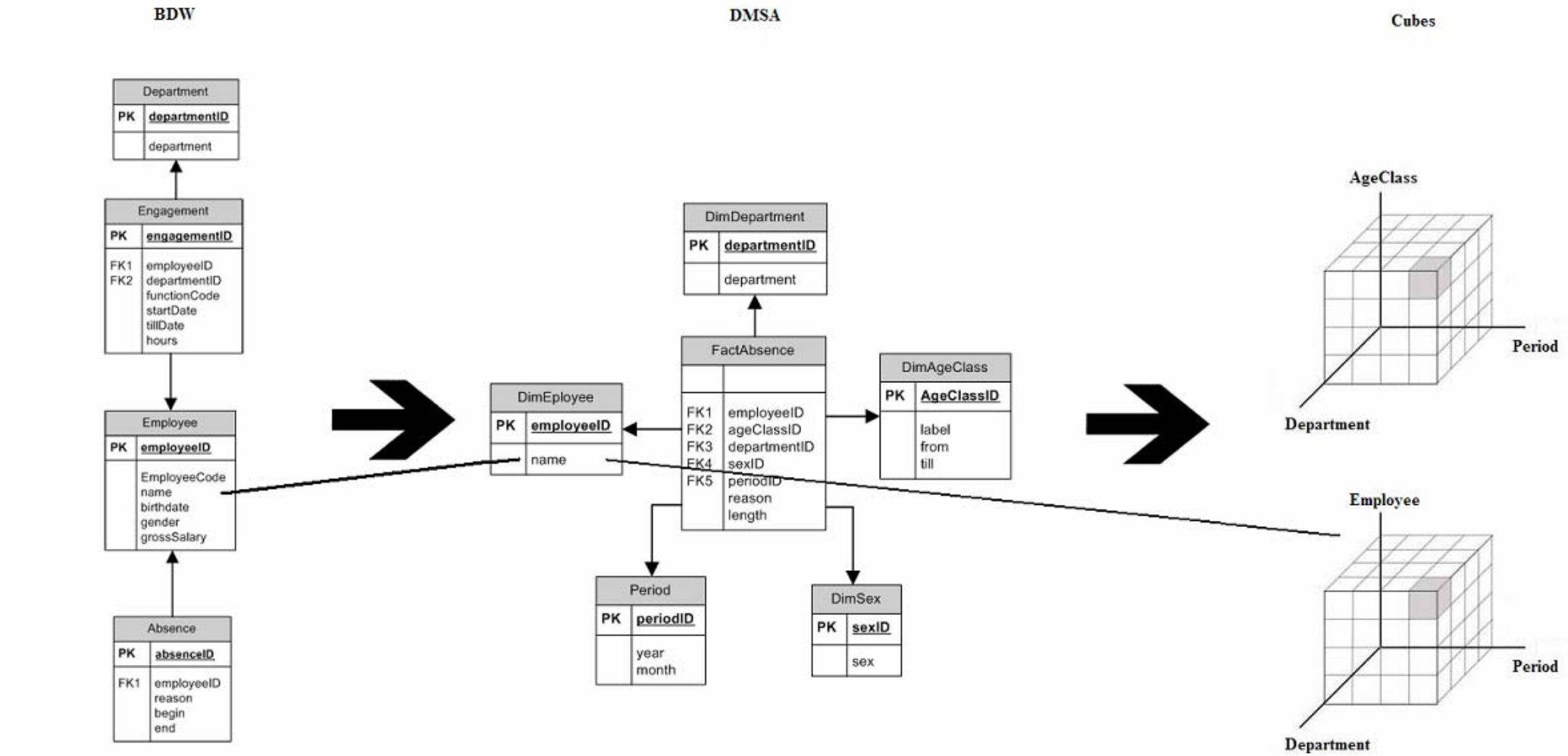
The final goal of impact analysis is to communicate the complexity of a change to others. Please indicate how well you are able to understand the change and its impact that were analyzed with the DIA:

	strongly disagree	**disagree**	**neither agree nor disagree**	**agree**	**strongly agree**
I am able to comprehend the change and its impact.					

Appendix D.: Traceability example (1/2)



Appendix D.: Traceability example (2/2)



Appendix E.: Explanation used in Field Study

The following Dutch explanation was provided to the subject of the field study. This step-plan was provided as a separate document.

UITLEG CASE STUDIE DIA 02/11/07

Wat is de bedoeling van de case studie?

Het is de bedoeling dat er voor een 1e casus een DIA uitgevoerd wordt zoals je deze normaliter zou doen. Je bent geheel vrij in je aanpak en handelingen om de DIA zo goed mogelijk als normaal uit te voeren.

Nadat je de eerste DIA volledig afgerond hebt, wil ik je wat achtergrond informatie geven over mijn onderzoek. Vervolgens is het de bedoeling dat je voor een 2e casus een DIA uitvoert. Bij het uitvoeren van deze DIA zal ik je vragen om op een bepaalde manier te werk te gaan. Verder ben je vrij in je aanpak en handelingen om de DIA uit te voeren zoals jou het beste lijkt. Gedurende de 2^{e} casus kun je mij op elk moment om toelichting vragen m.b.t. mijn onderzoek en hoe ik jou gevraagd heb om de DIA uit te voeren.

Nadat beide DIA's uitgevoerd zijn zullen we de case studie afsluiten met een evaluatie van de uitgevoerde DIA's.

Planning Case Studie

In de tabel hieronder wordt een planning gegeven voor de verschillende activiteiten van de case studie en welke personen er bij betrokken zijn.

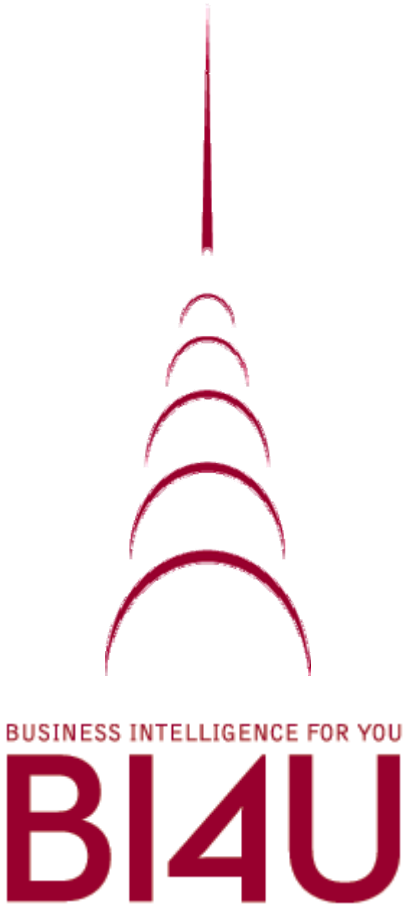
	Tijd	**Activiteit**	**Betrokken personen**
1	08:30-08:45	Toelichten Casus 1	*Shirly -> Martijn*
2	08:45-09:30	Uitvoeren DIA Casus 1	*Martijn*
3	09:30-10:15	Toelichten DIA onderzoek + ter beschikking stellen hulpmiddelen	*Jurriën -> Martijn*
4	10:15-10:30	Toelichten Casus 2	*Shirly -> Martijn*
5	10:30-11:15	Uitvoeren DIA Casus 2	*Martijn*
6	11:15-12:00	Evaluatie	*Martijn, Shirly, Jurriën*

N.b. De planning is een indicatie vooraf en geen strikte eis aan de duur van de verschillende activiteiten.

Appendix F.: Step-plan used in Field Study

The following 26 pages show the Dutch step-plan that was provided to the subject of the field study. This step-plan was provided as a separate document.

Please note that the page numbers in the index of this Appendix do not correspond to the page numbers of the thesis, but to the page numbers of the separate document.



Stappenplan Delta Impact Analyse

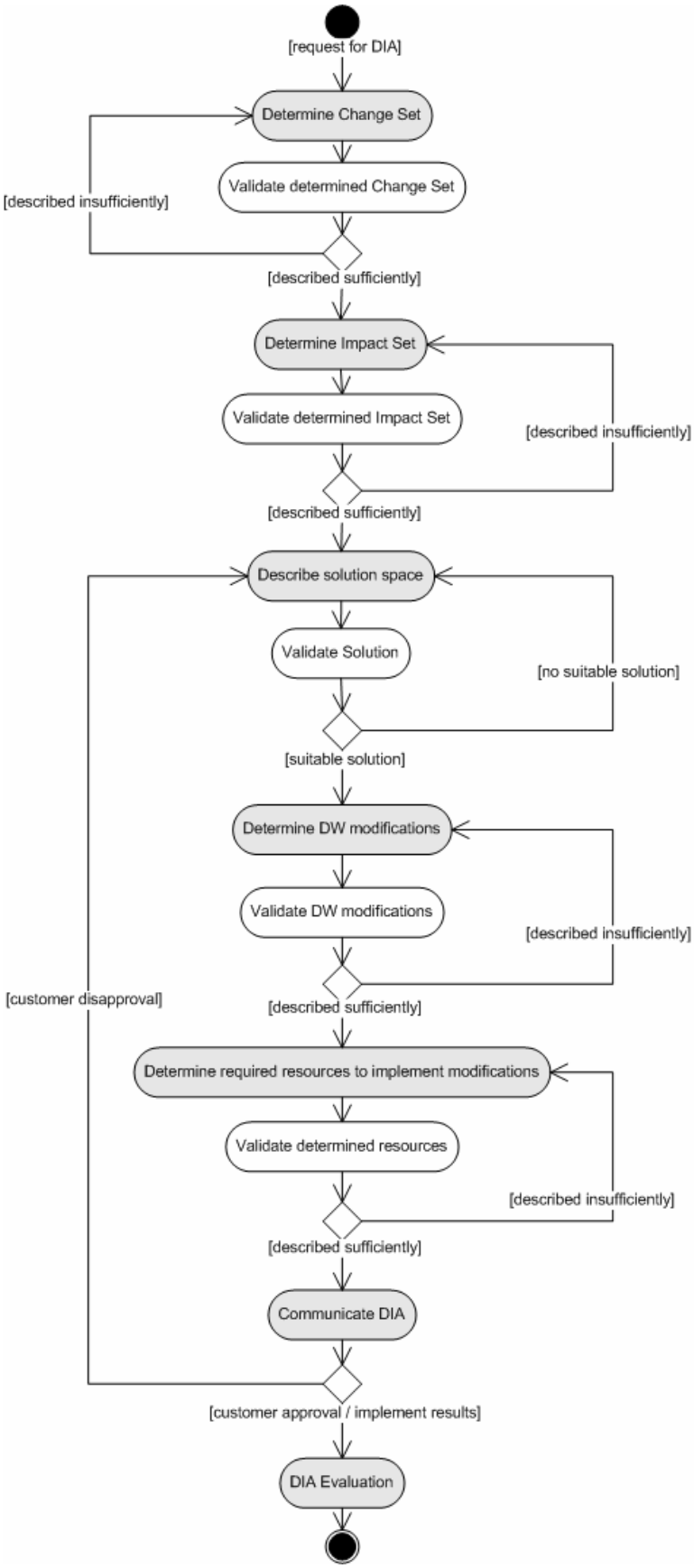
Versie: 1.0 Datum: 30/10/2007

Inhoudsopgave

1 Inleiding

In hoofdstuk 2 zullen in meer detail de zeven stappen van het Delta Impact Analyse proces beschreven worden. De eventueel relevante toelichting om de stappen uit te voeren wordt beschreven in hoofdstuk 3. De stappen zijn gebaseerd op het afstudeeronderzoek "*The design of a Delta Impact Analysis model for Data Warehouses*" door Jurriën Kok.

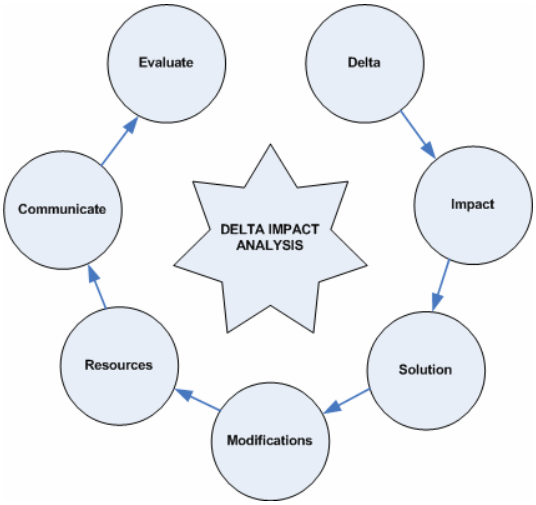
De afbeelding op de volgende pagina geeft het proces van Delta Impact Analyse globaal weer. De zeven stappen die in dit document worden beschreven zijn grijs gekleurd.


[request for DIA]
Determine Change Set
Validate determined Change Set
[described insufficiently]
[described sufficiently]
Determine Impact Set
Validate determined Impact Set
[described insufficiently]
[described sufficiently]
Describe solution space
Validate Solution
[no suitable solution]
[suitable solution]
Determine DW modifications
Validate DW modifications
[described insufficiently]
[customer disapproval]
[described sufficiently]
Determine required resources to implement modifications
Validate determined resources
[described insufficiently]
[described sufficiently]
Communicate DIA
[customer approval / implement results]
DIA Evaluation

2 DIA stappen

In de afbeelding hierboven zijn zeven stappen zichtbaar, namelijk:

1. Bepalen van de change set
2. Bepalen van de impact set
3. Beschrijven van de oplossingsruimte
4. Bepalen benodigde aanpassingen
5. Bepalen benodigde resources en kosten
6. Maak document voor communicatie DIA
7. Evalueer de uitgevoerde DIA



Tussen de zeven stappen staat in sommige gevallen een validatie stap. Dit betekent dat voordat de uitvoerder van de analyse doorgaat naar de volgende stap, hij moet valideren of de huidige stap voldoende is uitgevoerd. Deze validatie kan gedaan worden door de volgende vragen te stellen:

- Is de informatie compleet? (Kan ik of mijn collega op basis van deze informatie de volgende stap uitvoeren)
- Helpt de informatie bij het oplossen van het probleem?

Voor het uitvoeren van de verschillende stappen is de keuze aan de uitvoerder welke bronnen van informatie het beste geraadpleegd kunnen worden om tot een goede analyse te komen. Het gebruik van de volgende verschillende soorten bronnen is gebruikelijk:

- Kennisdragende collega's (m.b.t. het relevante data warehouse).
- Broncode en documentatie van het relevante data warehouse (eventueel meta warehouse).
- Oorspronkelijke ontwerp modellen van het data warehouse.

Suggesties voor mogelijke tools om de impact in kaart te brengen:

- Red Gate SQL Compare: vergelijken database structuren.
- Red Gate SQL Data Compare: vergelijken database inhoud.
- Red Gate Dependency Tracker: inzicht in database afhankelijkheden.

2.1 Stap 1: Bepaal Change Set

Het doel van deze stap is om alle wijzigingen duidelijk te beschrijven.

1.1. Bepaal voor elke delta om wat voor delta het gaat: schema, semantic of business rule
1.2. Indien schema of semantic delta: gebruik een database vergelijk tool om de change set te bepalen.
1.3. a. Indien schema delta: herorganiseer de wijzigingen naar refactorings.
b. Indien semantisch of business rule: specificeer welke elementen veranderen.
1.4. Bepaalde de impact van de delta op de informatie toevoer aan het data warehouse: preserving, reducing, increasing, redefining.

In de vervolgstappen van het model wordt gerefereerd naar de *change set*. Elke gedefinieerde delta aan het einde van stap 1 is een element in de *change set*.

Tussenstap Stap 1 -> Stap 2: valideer de bepaalde Change Set

2.2 Stap 2: Bepaal Impact Set

Het doel van deze stap is om alle afhankelijkheden van de elementen die wijzigen (stap 1) in kaart te brengen om zo inzicht te krijgen in de impact van de wijzigingen.

Onderstaande stappen zijn een tekstuele uitwerking van het diagram in de toelichting. Doorloop de stappen voor elke delta uit de change set. Afhankelijk van het type delta en de impact op de informatie toevoer de stappen voor alle componenten of slechts een deel hiervan uitgevoerd (zie diagram). Met een component wordt bedoeld, één van de verschillend onderdelen in het data warehouse: Bron, DSA, ISA, BDW, DMSA, Kubussen, Rapporten en bijbehorende databases en packages.

2.1. Bepaal het eerst te analyseren component aan de hand van het type delta:
   a. Schema delta in bron: eerst te analyseren component is bron.
   b. Semantic delta in bron: eerst te analyseren component is bron.
   c. Business rule delta: eerst te analyseren component is package dat BR implementeert.
2.2. Bepaal de afhankelijkheden binnen het data warehouse component. Desbetreffende elementen zijn onderdeel van de *impact set*.
2.3. Bepaal de afhankelijkheden naar het opvolgende data warehouse component van de elementen in de change en impact set. Desbetreffende elementen zijn onderdeel van de *impact set*.
2.4. Afhankelijk van de impact op de informatie toevoer, herhaal de stappen vanaf 2.2 zolang relevant voor het opvolgende component.
2.5. Indien de informatie toevoer *increasing* is probeer dan de scope duidelijk te beschrijven rondom dat wat aan het data warehouse toegevoegd kan worden.

Tussenstap Stap 2 -> Stap 3: valideer de bepaalde Impact Set

2.3 Stap 3: Bepaal Oplossingsruimte

Vaak zijn er meerdere mogelijkheden om wijzigingen op te vangen. Het doel van deze stap is om de belangrijkste mogelijkheden te beschrijven en tot een keuze voor de gewenste oplossing te komen.

Doorloop voor elke delta die gedefinieerd is in de change set de volgende stappen:

3.1. Bedenk welke verschillende oplossingen er mogelijk zijn.
3.2. Beschrijf een van de mogelijke oplossingen. Bedenk *oppervlakkig* (dus voer stap 4 en 5 niet daadwerkelijk al uit) :
   a. Welke aanpassingen benodigd zouden zijn (stap 4).
   b. Wat de omvang is m.b.t. benodigde resources en kosten (stap 5).
3.3. Bedenk of er nog een mogelijke oplossing is, zo ja ga terug naar stap 3.2.
3.4. Maak een keuze uit de beschreven oplossingen en beargumenteer de keuze. Dit moment biedt tevens de gelegenheid om anderen, zoals een opdrachtgever of leidinggevende, te betrekken bij het kiezen van de meest geschikte oplossingsrichting.

Tussenstap Stap 3 -> Stap 4: valideer de gekozen oplossing

2.4 Stap 4: Bepaal Benodigde aanpassingen

Na de keuze van een oplossing is het doel van deze stap om vast te stellen welke aanpassingen nodig zijn om de oplossing faciliteren. Er moet duidelijk beschreven worden welke elementen aangepast moeten worden en op wat voor manier.

Uit de gekozen oplossingsrichting moet duidelijk worden wat voor soort aanpassingen gedaan zullen moeten worden. Op basis daarvan zullen sommige stappen wel of niet relevant zijn. Voor elke gekozen oplossing voor de delta's uit de change set moeten onderstaande stappen uitgevoerd worden:

4.1. Beschrijf de benodigde aanpassingen op DSA niveau.
4.2. Beschrijf de benodigde aanpassingen op ISA niveau.
4.3. Beschrijf de benodigde aanpassingen op BDW niveau.
4.4. Beschrijf de benodigde aanpassingen op DMSA niveau.
4.5. Beschrijf de benodigde aanpassingen op Kubus niveau
4.6. Beschrijf de benodigde aanpassingen op Rapport niveau.
4.7. Beschrijf de benodigde aanpassingen m.b.t. historische data in het data warehouse.

Tussenstap Stap 4 -> Stap 5: valideer de bepaalde benodigde aanpassingen

2.5 Stap 5: Bepaal benodigde resources en kosten

Het doel van deze stap is om voor alle gekozen oplossingen van de delta's de benodigde resources en kosten in te schatten. Voor elke gekozen oplossing voor de delta's uit de change set moeten onderstaande stappen uitgevoerd worden:

5.1. Bepaal de omvang van elke aanpassing door de hoeveelheid betrokken elementen in te schatten.
5.2. Bepaal het gewicht van elke aanpassing door de relevante gewichtsfactor op de geschatte hoeveelheid elementen toe te passen.
5.3. Schat de benodigde hoeveelheid werk om elke aanpassing uit te voeren door de inspanningsfactor op het geschatte gewicht toe te passen.
5.4. Gebaseerd op de totale hoeveelheid uren benodigd per aanpassing, maak een schatting van de kosten per aanpassing en de totale kosten, geef eventueel een toelichting m.b.t. de totstandkoming.
5.5. Maak een algehele planning voor alle aanpassingen die gedaan moeten worden voor de geïdentificeerde delta's, geef eventueel een toelichting m.b.t. de totstandkoming.

Tussenstap Stap 5 -> Stap 6: valideer de bepaalde benodigde resources en kosten

2.6 Stap 6: Maak document voor communicatie DIA

Het doel van deze stap is om de geanalyseerde delta's en hun impact voldoende duidelijk te beschrijven zodat deze gecommuniceerd kan worden naar anderen. Dit kan een collega zijn die uiteindelijk de implementatie moet doen, een leidinggevende, of bijvoorbeeld een opdrachtgever die zijn goedkeuring moet geven.

Deze stap wordt voornamelijk reeds gefaciliteerd door de beschikbare template voor de DIA. De volgende stappen moeten daarnaast nog uitgevoerd worden:

6.1. Controleer het document inhoudelijk op:
   a. Staat alle informatie in het document?
   b. Is alles voldoende toegelicht?
   c. Is de spelling en grammatica correct?

6.2. Bedenk of het document geschikt is voor de doelgroep en pas het eventueel aan.

6.3. Bedenk op welke punten de template aangepast kan worden ter verbetering en voer deze verbeteringen door.

6.4. Geef het document aan de persoon/personen naar wie de DIA gecommuniceerd moet worden. Licht, indien nodig, het document persoonlijk toe.

2.7 Stap 7: Evalueer uitgevoerde DIA

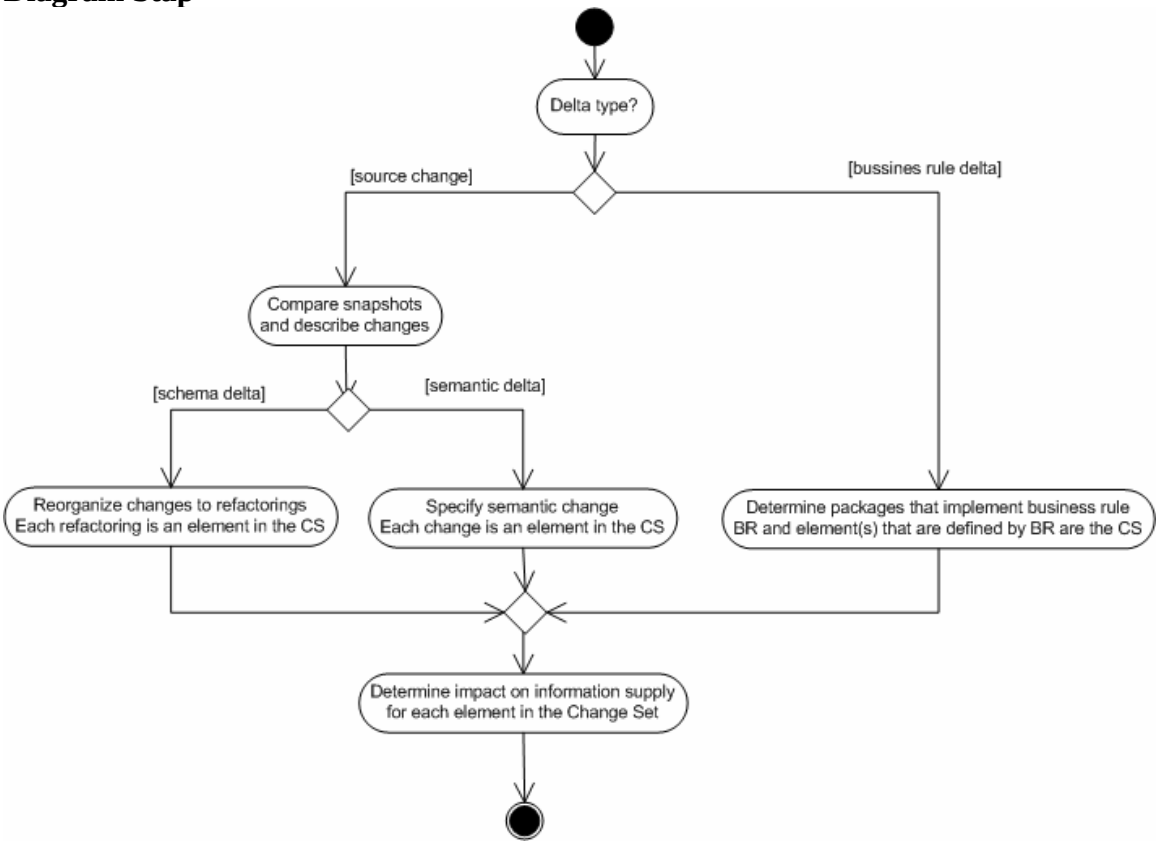
Het doel van de laatste stap in het DIA proces is om de kwaliteit van de uitgevoerde DIA te evalueren, hiervan te leren, en deze leerervaring te vertalen naar aanpassingen in het DIA model.

7.1. Noteer de hoeveelheid besteedde uren aan de DIA.
7.2. Tel de hoeveelheid elementen in de change set, impact set, geschatte impact set en daadwerkelijke impact set.
7.3. Bereken de correct-, verkeerd-, gemist-, amplificatie- en verander ratio's aan de hand van de aantallen in de CS, IS, GIS en DIS.
7.4. Laat de evaluatie vragenlijst invullen door de persoon/personen naar wie de DIA gecommuniceerd is door middel van het DIA document.
7.5. Vul zelf de evaluatie vragenlijst in.
7.6. Documenteer (in het DIA document) de uitkomsten van de evaluatie en de verschillen tussen dat wat geanalyseerd is en dat wat de praktijk was.
7.7. Stel eventueel aspecten aan het DIA model bij, aan de hand van de bevindingen van de evaluatie.

3 Toelichting DIA Stappen

3.1 Stap 1: Bepaal Change Set

Diagram Stap



Algemeen

Onderstaande tabel geeft de mogelijke impact per type delta weer op de informatie toevoer.

Tabel 12. Mogelijke impact op informatie toevoer

	Preserved	Reduced	Increased	Redefined
Schema	x	x	x	
Semantic	x	x	x	x
Business Rule		x	x	x

Uitleg definities:

- **Schema delta:** Een verandering in het database schema van een data warehouse bron
- **Semantic delta:** Een verandering in de semantiek van data in een data warehouse bron
- **Business rule delta:** Een verandering in een business rule die de betekenis van data in een data warehouse definieert.

- **Informatie toevoer**: De set van informatie van de bronsystemen die beschikbaar is voor het datawarehouse
- **Preserved**: De informatietoevoer blijft gelijk.
- **Reduced:** De informatietoevoer neemt af.
- **Increased:** De informatietoevoer neemt toe.
- **Redefined**: De informatietoevoer neemt niet toe of af, maar wordt geherdefinieerd (verandert in betekenis).

Schema delta

Voor schema delta's kunnen de wijzigingen worden beschreven in refactorings die in onderstaande tabel opgesomd staan. In de rechter kolom staat wat de impact is op de informatie toevoer aan het data warehouse.

Tabel 13. Information supply impact of refactorings

Event	Information supply
Merge table Split table Move column Rename column Rename table Replace large object with table Replace one-to-many with associative table Replace column Introduce surrogate key Replace surrogate key with natural key Drop view Rename view	Preserved information supply
Merge column Drop column Drop table	Reduced information supply
Split column Introduce column Introduce table	Increased information supply

Semantic & Business rule delta

Voor semantic en business rule delta's moet de uitvoerder van de DIA naar eigen inzicht bepalen wat de impact van de delta op de informatie toevoer is.

Tools

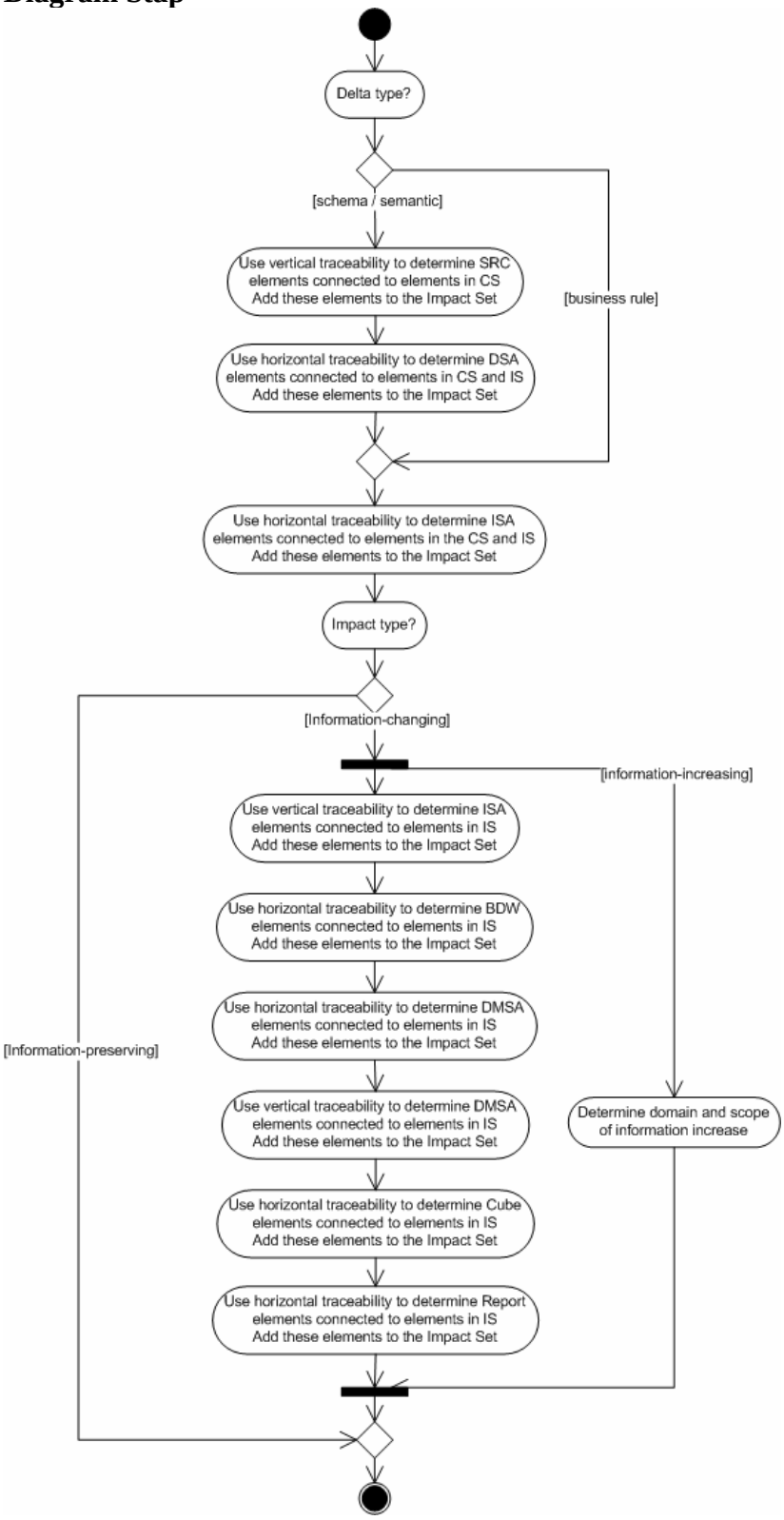
Indien het mogelijk is om twee versies van een bron database, één van voor de door te voeren wijziging, en één van er na, dan kunnen de volgende tools gebruikt worden om schema en semantische delta te identificeren:

- RedGate SQL Compare[4]: vergelijken van database schemata
- RedGate SQL Data Compare[5]: vergelijken van database inhoud

[4] http://www.red-gate.com/products/SQL_Compare/index.htm
[5] http://www.red-gate.com/products/SQL_Data_Compare/index.htm

3.2 Stap 2: Bepaal Impact Set

Diagram Stap



Algemeen

Voor het bepalen van de afhankelijkheden kunnen de reeds genoemde verschillende bronnen worden geraadpleegd:

- Kennisdragende collega's (m.b.t. het relevante data warehouse).
- Broncode en documentatie van het relevante data warehouse (eventueel meta warehouse).
- Oorspronkelijke ontwerp modellen van het data warehouse.

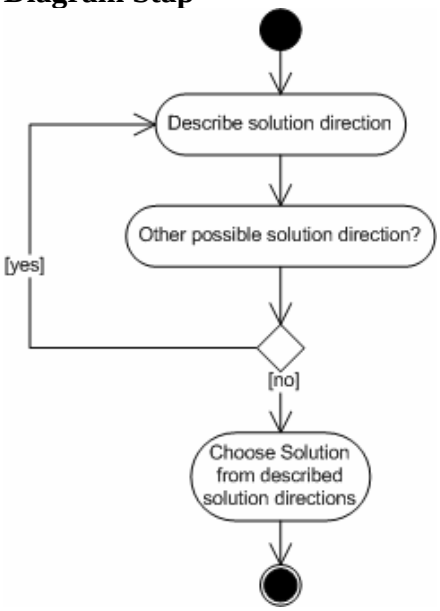
De volgende tool is mogelijk te gebruiken om afhankelijkheden tussen de verschillende componenten in kaart te brengen (nog niet getest in de praktijk):

- RedGate SQL Dependency Tracker[6]: in kaart brengen afhankelijkheden database objecten (cross-database).

[6] http://www.red-gate.com/products/SQL_Dependency_Tracker/index.htm

3.3 Stap 3: Bepaal Oplossingsruimte

Diagram Stap



Algemeen

Mogelijke variatie binnen één oplossingsrichting kan leiden tot meerdere voorgestelde oplossingen. Daarnaast kan de vastgestelde impact op de informatie toevoer als inspiratie voor verschillende oplossingen dienen. Onderstaande tabel geeft een overzicht welke activiteiten onderdeel van een oplossing moeten en kunnen zijn, afhankelijk van de impact op de informatie toevoer.

Tabel 14. Impact-activity matrix

	Preserved	**Reduced**	**Increased**	**Redefined**
SRC -> DSA-to-ISA (ET)				
Aanpassen	must	must	must	must
ISA -> Reports (Core)				
Aanpassen	-	can	can	must
Uitrekenen historische waarden	-	-	can	must

De keuze voor een oplossingsrichting kan gemaakt worden zodra de verschillende mogelijke oplossingen in kaart gebracht zijn. Afhankelijk van de situatie zal de keuze door de uitvoerder van de DIA zelf gemaakt kunnen worden, of vindt hiervoor overleg met derden plaats.

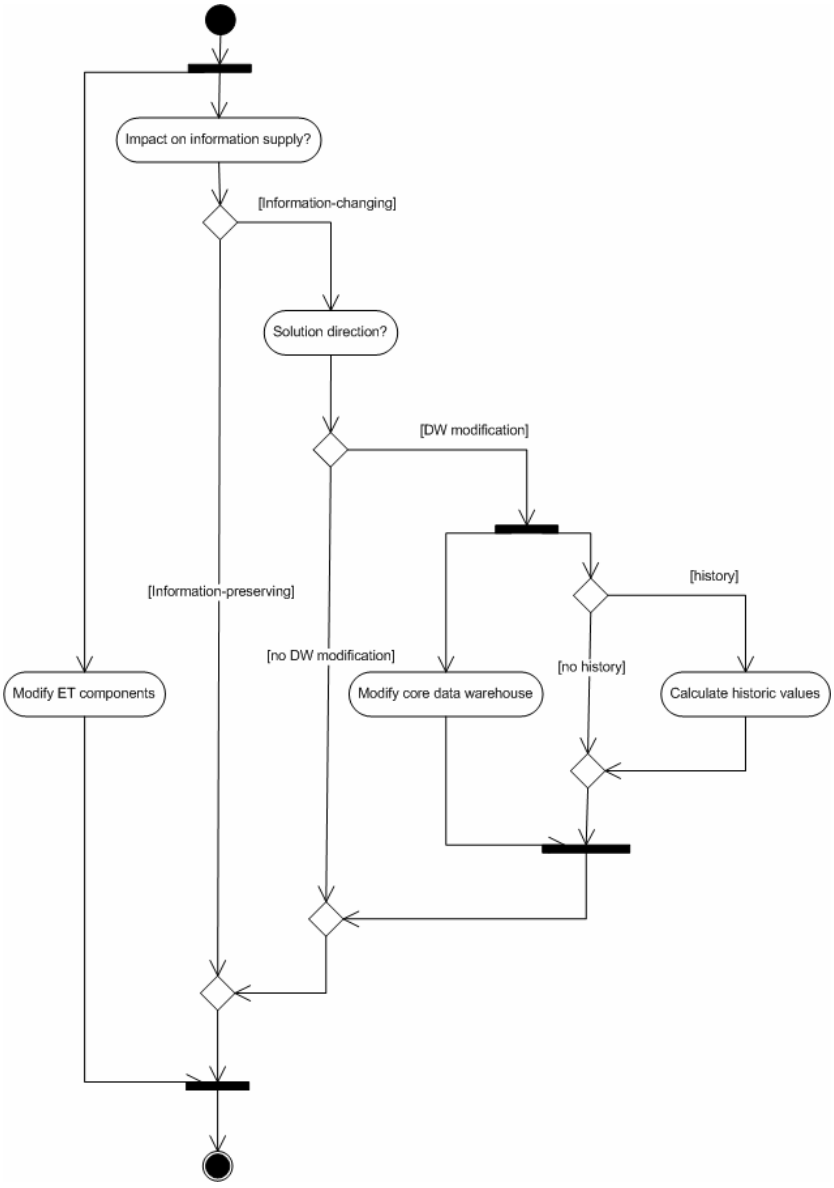
Een toelichting op het soort activiteiten voor de drie categorieën van aanpassingen:

4. *Aanpassen ET components*: aanpassingen aan de elementen van de Source-to-DSA, DSA-to-ISA packages en DSA database.
5. *Aanpassen core data warehouse:* verwijderen of aanpassen van elementen in de ISA, BDW en DMSA databases, de ISA-to-BDW, BDW-to-DMSA, DMSA-to-Cubes packages, de kubussen en de rapporten.

6. *Uitrekenen historische waardes*: activiteiten die benodigd zijn om historische waarden uit te rekenen voor de nieuwe situatie en eventueel de bestaande waardes aan te passen. Dit kan gewenst zijn indien de informatietoevoer *increased* is (historische informatie is ook interessant) en is noodzakelijk indien de informatie toevoer redefined is (historische informatie moet herberekend worden aan de hand van de nieuwe definities).

3.4 Stap 4: Bepaal Benodigde aanpassingen

De verschillende mogelijke benodigde aanpassing zijn afhankelijk van het type delta en de gekozen oplossingsrichting. Onderstaand diagram kan doorlopen worden om te bepalen welk soort aanpassings activiteiten verricht moeten worden voor de gekozen oplossingsrichting.



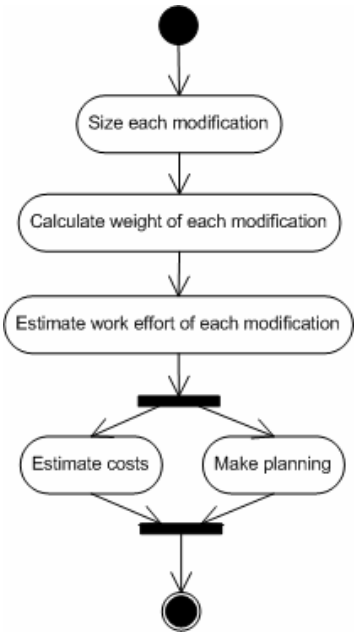
Naast het bepalen van de soort activiteiten die plaats moeten vinden, zal inhoudelijk beschreven moeten worden wat de benodigde aanpassing inhoud. Voor de aanpassingen van componenten zal dit naar inzicht gedaan moeten worden aan de hand van de bepaalde impact. M.b.t het uitrekenen van historische waardes geeft onderstaande tabel meer inzicht in de mogelijk benodigde activiteiten afhankelijk van wat de toepassing van de increased/redefined data is in de DMSA.

Table 15. Uitreken activiteiten

DMSA toepassing	**Activiteit m.b.t. historische data**
Meetwaarde	Schat waardes voor de nieuwe meetwaarde
Afgeleide meetwaarde	Bereken waardes voor die nieuwe meetwaarde
Component (Dimensie Hiërarchie)	Consistent waardes toevoegen voor het nieuwe component in de dimensie
Dimensie	Aanpassen van feitwaardes aan de hand van een op te stellen business rule, indien mogelijk

3.5 Stap 5: Bepaal benodigde resources en kosten

Diagram Stap



Benodigde resources

Voor het inschatten van het gewicht en de benodigde arbeidsuren kan gebruik worden gemaakt van een gewichtsratio en een inspanningsfactor. Op dit moment zijn deze factoren nog niet gedefinieerd en zullen deze naar beste inzicht moeten worden ingeschat. In de toekomst kunnen deze dan bijgesteld worden aan de hand van de evaluatie van het DIA proces.

Tabel 16. Gewichts- en inspanningsfactoren

Aanpassings Type	Gewichts-factor	Inspannings-factor
Aanpassen Source -> DSA-to-ISA components	X_1	Y_1
Aanpassen core Data Warehouse components	X_2	Y_2
Berekenen historische waardes	X_3	Y_3

Kosten

De kosten kunnen simpelweg berekend worden door de geschatte benodigde uren te vermenigvuldigen met het uurtarief van een ontwikkelaar die de oplossing zal gaan implementeren.

Planning

Voor het maken van een planning moet rekening gehouden worden met effectief te besteden uren, weekenden etc. Om persoon-dagen om te rekenen naar kalender-dagen kan een factor toegepast worden die op dit moment is vastgesteld op: *1,487*. Op basis

van ervaring kan deze ratio in de toekomst bijgesteld worden. Deze factor gaat uit van het feit dat een persoon de gehele werkweek beschikbaar is voor implementatie. Er moet dus rekening worden gehouden hoeveel uren van de werkweek de ontwikkelaar beschikbaar is om de oplossing te implementeren:

Kalender-dagen = ([benodigde uren] / [aantal beschikbare uren per week]) * 5 * 1,487

3.6 Stap 6: Maak document voor communicatie DIA

Maak gebruik van de beschikbaar gestelde DIA template.

3.7 Stap 7: Evalueer uitgevoerde DIA

Ratio's

Een aantal ratio's kunnen onderscheiden worden om de adequaatheid en effectiviteit van de uitgevoerde DIA na implementatie te evalueren:

- Correct ratio: $|(DIS \cap GIS)| / |GIS|$
- Vekeerd ratio: $|((DIS \Delta GIS) \cap GIS)| / |GIS|$
- Gemist ratio: $|((DIS \Delta GIS) \cap DIS)| / |GIS|$
- Amplificatie ratio: $|IS| / |CS|$
- Verander ratio: $|GIS| / |Systeem|$

De termen nader toegelicht:

- CS: de Change Set. Zie Stap 1.
- IS: de Impact Set. Zie Stap 2.
- DIS: de Daadwerkelijke Impact Set. De elementen waar na afloop van de implementatie van geconstateerd kan worden dat de delta hier een impact op had.
- GIS: de Geschatte Impact Set. De Change Set (CS) en Impact Set (IS) bij elkaar opgeteld.
- |Systeem|: de hoeveelheid elementen waaruit het systeem bestaat om een relatief perspectief te kunnen bieden op de veranderde elementen.
- Correct ratio: het aandeel correct geschatte impact elementen.
- Verkeerd ratio: het aandeel (incorrect) geschatte impact elementen, waar uiteindelijk geen impact op was.
- Gemist ratio: het aandeel elementen waar wel een impact op was, maar waarvan dit niet ingeschat was.
- Ampl. ratio: de relatieve impact van de change set.
- Verander ratio: hoe groot aandeel van het gehele systeem is geschat impact van de delta te ondervinden.

Vragenlijsten

De vragenlijsten van de volgende 2 pagina's kunnen ook gebruikt worden om de DIA te evalueren.

Vragenlijst evaluatie effectiviteit DIA

Eén doel van impact analyse is om een duidelijk perspectief te vergaren en te communiceren m.b.t. de scope van de verandering. Geef gaarne een waardering voor de mate van detail en compleetheid waarin de scope van de verandering is beschreven:

	Zeer oneens	**Oneens**	**Niet eens, noch oneens**	**Mee eens**	**Zeer eens**
De scope van de verandering is in voldoende detail beschreven.					
De scope van de verandering is voldoende compleet beschreven.					

Een ander doel van impact analyse is om inzicht te vergaren in de benodigde resources en kosten voordat een verandering wordt doorgevoerd. Geef gaarne een waardering voor de mate van detail en compleetheid van het overzicht waarin de benodigde resources en kosten zijn beschreven. Geef tevens een waardering aan de mate van accuraatheid van de geschatte benodigde resources en kosten:

	Zeer oneens	**Oneens**	**Niet eens, noch oneens**	**Mee eens**	**Zeer eens**
De benodigde resources en kosten zijn in voldoende detail beschreven.					
De benodigde resources en kosten zijn voldoende compleet beschreven.					
De benodigde resources en kosten zijn voldoende accuraat beschreven.					

Een ander doel van impact analyse is om in staat te zijn een kosten/baten analyse te maken op basis van de beschreven informatie: wat zijn de voordelen van het implementeren van een oplossing, wat zijn de kosten, is het de moeite waarde? Geef gaarne aan hoe goed het mogelijk is om de kosten/baten afweging te maken op basis van de informatie van de DIA:

	Zeer oneens	Oneens	Niet eens, noch oneens	Mee eens	Zeer eens
Ik ben in staat om een afweging te maken tussen de kosten en de baten van het implementeren van de beschreven oplossing(en).					

Het uiteindelijke doel van impact analyse is om de complexiteit van een verandering naar anderen te communiceren. Geef gaarne aan hoe goed je in staat bent om de verandering en haar impact te begrijpen die d.m.v. de DIA geanalyseerd zijn.

	Zeer oneens	Oneens	Niet eens, noch oneens	Mee eens	Zeer eens
Ik ben voldoende in staat om de verandering en haar impact te begrijpen.					

Vragenlijst evaluatie DIA

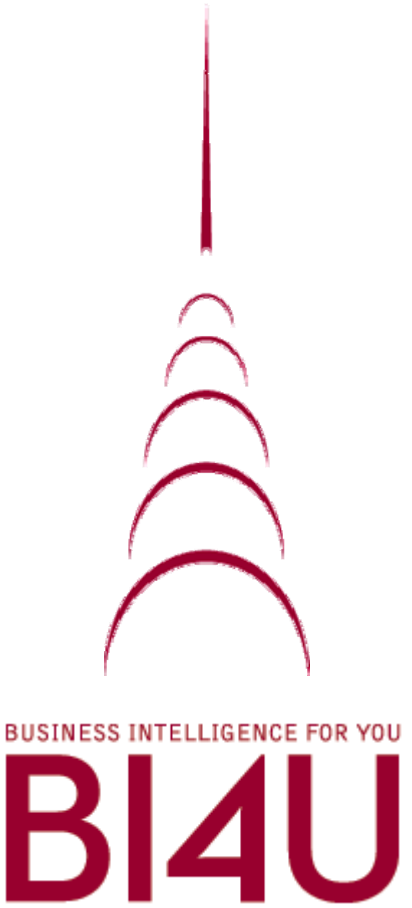
Om de verschillen te beschrijven dienen de volgende vragen beantwoord te worden door de uitvoerder van de DIA:

- Waren de gegeven stappen in het DIA model voldoende om de DIA uit te voeren? Zo niet, wat waren de verschillen tussen de daadwerkelijk uitgevoerde stappen en die van het model?
- Wat zijn de verschillen tussen de geschatte impact en de daadwerkelijke impact?
- Wat zijn de verschillen tussen de geschatte benodigde aanpassingen en de daadwerkelijke benodigde aanpassingen?
- Wat zijn de verschillen tussen de geschatte resources & kosten en de daadwerkelijke resources & kosten?
- Wat zijn de verschillen tussen de planning en de daadwerkelijke uitvoeringstermijn van de DIA?
- Op welke manier kan/moet het DIA model aangepast worden om dergelijke verschillen (voorgaande 4 vragen) in de toekomst te verkomen, dan wel te verminderen?
- Zijn er nog overige opmerkingen m.b.t. de uitgevoerde DIA?

Appendix G.: Template used in Field Study

The following 8 pages show the Dutch template that was provided to the subject of the field study. This step-plan was provided as a separate document.

Please note that the page numbers in the index of this Appendix do not correspond to the page numbers of the thesis, but to the page numbers of the separate document.



Delta Impact Analyse

<omschrijving DIA Case>

Versie: <versie> Datum: <datum>

Inhoudsopgave

1 Change Set

[*Behandel elke geïdentificeerde delta in een aparte paragraaf*]

1.1 Delta A

Type	<schema/semantic/business rule>
Impact op informatie toevoer	<preserving/reducing/increasing/redefining>
Omschrijving	

1.2 Delta B

Type	<schema/semantic/business rule>
Impact op informatie toevoer	<preserving/reducing/increasing/redefining>
Omschrijving	

[*Behandel elke in hoofdstuk 1 beschreven delta in een apart hoofdstuk*]

2 Delta A

<samenvatting beschreven delta>

2.1 Geschatte Impact

2.1.1 DSA

<Indien van toepassing*:* beschrijf>

2.1.2 BDW

<Indien van toepassing: beschrijf>

2.1.3 DMSA

<Indien van toepassing: beschrijf>

2.1.4 Kubussen / Rapporten

<Indien van toepassing: beschrijf>

2.2 Oplossingsruimte

[*Bespreek elke mogelijke oplossingsrichting in een aparte sectie*]

2.2.1 Oplossing <X>

[beschrijf oplossing]

2.2.2 Gekozen Oplossing

[beargumenteer keuze voor oplossing]

2.3 Benodigde aanpassingen

[*Bespreek elke benodigde aanpassing voor de gekozen oplossing in een aparte sectie*]

2.3.1 DSA

<Indien van toepassing*:* beschrijf>

2.3.2 ISA

<Indien van toepassing*:* beschrijf>

2.3.3 BDW

<Indien van toepassing: beschrijf>

2.3.4 DMSA

<Indien van toepassing: beschrijf>

2.3.5 Kubussen

<Indien van toepassing: beschrijf>

2.3.6 Rapporten

<Indien van toepassing: beschrijf>

2.3.7 Historische informatie

<Indien van toepassing: beschrijf>

2.4 Benodigde resources

<Schat voor iedere aanpassing benodigde resources>

Categorie	**# elementen**	**Gewicht**	**Uren**
DSA			
ISA			
BDW			
DMSA			
Kubussen			
Rapporten			
Historische informatie			
Totaal		*<tel op>*	*<tel op>*

3 Kosten

<Maak een overzicht van de verschillende delta's en geef een bereken de kosten voor de geschatte uren>

Delta	**Uren**	**Kosten**
<A>	<neem *Totaal* over>	€ <vul in>
<B>	<neem *Totaal* over>	€ <vul in>
Totaal	<tel op>	<tel op>

<eventuele toelichting>

4 Planning

<Maak een overzicht van de verschillende delta's en geef een planning voor het implementeren van de gedefinieerde aanpassingen>

Delta	**Uren**	**Begin**	**Eind**
<A>	<neem *Totaal* over>	<dd-mm-yy>	<dd-mm-yy>
<B>	<neem *Totaal* over>	<dd-mm-yy>	<dd-mm-yy>
Totaal	*<tel op>*	*<dd-mm-yy>*	*<dd-mm-yy>*

<eventuele toelichting>

5 Evaluatie

<Vul na implementatie van de DIA de volgende gegevens in>

Variabelen	**Waarde**
|CS|	<vul in>
|IS|	<vul in>
|GIS|	<tel op>
|DIS|	<vul in>

Adequaatheid	**Waarde**
Correct ratio	<bereken>
Verkeerd ratio	<bereken>
Gemist ratio	<bereken>

Effectiviteit	**Waarde**
Tijd besteed aan DIA	<vul in>
Amplificatie ratio	<bereken>
Verander ratio	<bereken>
Feedback score Scope	<vul in>
Feedback score resources en kosten	<vul in>
Feedback score kosten/baten afweging	<vul in>
Feedback score communiceren complexiteit	<vul in>

<beschrijvende evaluatie uitgevoerde DIA>

Appendix H.: Management Summary (Dutch/Nederlands)

Motivatie

Het vakgebied van data warehousing is in de afgelopen decennia ontstaan en gegroeid. Een data warehouse wordt op een bepaald moment ontwikkeld ter ondersteuning van business intelligence voor een ongedefinieerde periode. De wereld rondom het data warehouse verandert gedurende zijn leven, inclusief de systemen die fungeren als een bron voor het data warehouse. Om het data warehouse blijvend te laten functioneren en de kwaliteit van data te garanderen, zal deze moeten worden aangepast aan de om zich heen veranderende wereld. Het concept *Delta Impact Analyse* wordt gebruikt door het bedrijf BI4U als omschrijving voor de activiteiten om de impact van specifieke wijzigingen te analyseren. Dit concept is belangrijk omdat het inzicht geeft in hoe een data warehouse aangepast moet worden aan de om zich heen veranderende wereld en de gevolgen hiervan. De motivatie voor dit onderzoek was het ontbreken van een duidelijke definitie voor het concept DIA en wat het omvat.

Doelen

De hoofddoelen van het onderzoek waren om het onderwerp van DIA in de praktijk te onderzoeken, om inzichten te vergaren uit de literatuur en ander onderzoek, om een praktisch model voor DIA te ontwerpen en ontwikkelen, om het DIA model te testen en om aanbevelingen te doen om een betere ondersteuning te bieden voor veranderingen in data warehouse bronsystemen. Deze doelen resulteerden in de volgende algemene probleemstelling: *Hoe kan een Delta Impact Analyse model worden ontworpen dat ondersteuning biedt voor het proces van het analyseren van impact van veranderingen in een data warehouse bronsysteem situatie?*

Aanpak

De onderzoeksaanpak is gebaseerd op het *design science* framework van Hevner et al. in combinatie met *action science* theorie om het, uit het onderzoek resulterende, DIA model te valideren. Het design science perspectief resulteerde in een aanpak die zowel rigor is, door middel van een literatuuronderzoek, als relevant is, door het onderzoek te betrekken op de praktijk. Het onderzoek begon door het concept DIA bij BI4U in de praktijk te onderzoeken, om zo meer inzicht te vergaren in wat DIA omvat en wat relevant was voor de focus van het onderzoeksproject. Vervolgens is een grondig literatuuronderzoek uitgevoerd. Ten slotte is een artefact voorgedragen, een model voor het proces van DIA, en is deze gevalideerd door middel van een field study in de praktijk.

Resultaten

Het onderzoek biedt verschillende bijdrages. De eerste bijdrage is het feit dat er is geconstateerd dat verschillende aspecten van DIA onduidelijk waren bij BI4U. Er ontbraken duidelijke richtlijnen met betrekking tot welke taken precies onderdeel waren van DIA en hoe deze uitgevoerd zouden moeten worden. Tevens was er geen duidelijk raamwerk aanwezig met betrekking tot welk soort veranderingen kunnen voorkomen en welke impact deze zouden kunnen hebben. Er is wel een betere visie omtrent de scope van DIA vastgesteld, wat een betere focus voor het onderzoek bood.

De tweede bijdrage is het feit dat in het literatuuronderzoek inzichten uit het vakgebied van impact analyse zijn toegepast op inzichten uit het vakgebied van data warehousing. Categorisaties zijn voorgesteld om zowel delta als impact te classificeren. De delta kan onderscheiden worden door middel van de oorsprong: schema, semantic of business rule. Een *schema* delta heeft betrekking op de structurele wijzigingen aan een bron database, een *semantic* delta heeft betrekking op wijzigingen in de betekenis of representatie van data in de bron, een *business rule* delta heeft betrekking op wijzigingen in de betekenis van bron data voor het data warehouse. Bij de impact kan onderscheid gemaakt worden tussen: *information-preserving* and *information-changing* impact op de informatietoevoer naar het data warehouse. Binnen het '*changing'* impact type kan nog onderscheid gemaakt worden tussen een informatietoevoer die *reduced, increased* of *redefined* is. Voor de verschillende delta categorieën is gespecificeerd welke impact categorieën van toepassing kunnen zijn, dit voorziet in een brug tussen het definiëren van een delta en het bepalen van de impact in impact analyse. Om met de verschillende information-changing impact types om te kunnen gaan heeft men de keuze om de kern van het data warehouse niet aan te passen, wel aan te passen of aan te passen in combinatie met aanpassingen in de historische data.

De laatste en meest substantiële bijdrage van dit onderzoek is dat de resultaten van de eerste twee bijdrages gecombineerd en toegepast zijn in een artefact. Dit artefact is een model voor het proces van Delta Impact Analyse. Het model beschrijft zeven stappen die uitgevoerd moeten worden om een deugdelijke DIA uit te voeren: van het identificeren van de concrete veranderingen en impact tot het voorstellen van een oplossing inclusief een schatting van de benodigde resources en kosten, communicatie hiervan en een evaluatie. Evaluatie criteria voor het proces zijn bepaald en gebruikt voor een praktische field study als validatie van het voorgestelde model in dit onderzoek. De field study heeft inzicht gegeven in welke aspecten in de toekomst verbeterd kunnen worden.

Conclusies and discussie

Het resultaat van dit onderzoek, het model voor het proces van Delta Impact Analyse, biedt een deugdelijk startpunt voor het tackelen van veranderende bronsysteem situaties voor BI4U en ook voor anderen in het vakgebied van data warehousing.

Verschillende aanbevelingen kunnen worden gegeven met betrekking tot de praktijk. Ten eerste zal het model moeten worden verbeterd door het consistent te gebruiken voor de uitvoering van Delta Impact Analyses en het aan te passen aan de hand van de uitkomsten van de evaluatie van het proces. Tevens is het interessant om toekomstig onderzoek te doen naar de bruikbaarheid van bestaande tools en/of het ontwikkelen van een tool ter ondersteuning van stappen in het DIA proces. Met betrekking tot het wetenschappelijk vakgebied kunnen ook enige aanbevelingen gedaan worden voor toekomstig onderzoek. Er kan bijvoorbeeld meer onderzoek gedaan worden naar resource- en kostenschatting modellen en technieken die bruikbaar zijn voor DIA. Tevens is het interessant om te onderzoeken hoe het DIA model uitgebreid kan worden tot een compleet model, dat toegepast kan worden om met elk soort verandering om te gaan omtrent een geïmplementeerde data warehouse oplossing. De onderwerpen van *data warehouse*

evolutie en *meta warehouse* zijn aangemerkt als interessante onderwerpen die het proces van DIA kunnen ondersteunen.

Appendix I.: Glossary

Business Rule: The logic applied to calculate or otherwise derive a business-related value.

Business Rule Delta: A change in a business rule that defines the meaning of data in the data warehouse.

Change Set: *see Primary Impact Set.*

Data Warehousing: A subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.

Delta: The gap between two instances of an object at different moments in time.

Impact Analysis: The process of predicting the impact caused by a change.

Impact Set: *see Secondary Impact Set.*

Delta Impact Analysis: All activities performed to identify the consequences and required counter actions of a planned or performed change in the source system situation that influences, or even obstructs, the correct extracting, loading, transformation and/or analysis of the data by the data warehouse.

Estimated Impact Set: The union of the *Primary Impact Set* and the *Secondary Impact Set.*

Intelligence Factory: A collection of knowledge, methods, models and modules developed by BI4U to develop data warehouse solutions.

Information-changing: An event that changes the *information supply.* Three sub classifications of information-changing are: information-increasing, information-reducing, information-redefining.

Information-increasing: An event that increases the *information supply*; more information is available for the data warehouse.

Information-preserving: An event that keeps the *information supply* intact.

Information-redefining: An event that changes the meaning of the *information supply*.

Information-reducing: An event that reduces the *information supply*; less information is available for the data warehouse.

Information supply: The total set of information provided by sources that is available. for the data warehouse.

Primary Impact Set: The set of identified elements that form the change (delta). This term is also referred to as *Change Set* in the proposed DIA model.

Refactoring: changes to the structure of a database that are of a limited extent, do not impact the semantics and do not break the functionality of the system using the database.

Schema: The definition of the structure of a database.

Schema Delta: A change in the database schema of a data warehouse source.

Secondary Impact Set: The set of identified elements as being impacted by the *Primary Impact Set.* This term is also referred to as *Impact Set* in the proposed DIA model.

Semantic Delta: a change in the semantics of the data of a data warehouse data source.