Dimension Identification in Data Warehouse Based on Activity Theory

Yuan Gao
Dimension Identification in Data Warehouse Based on Activity Theory

Student;
Yuan Gao
Examiner;
Mathias Hedenborg
Supervisors;
Peng Zhang
Marcelo Milrad
Preface

Summer comes again. The city of Växjö is as beautiful as the one I saw last August when I just arrived. After study ten-month here, I have a memorable impression about this beautiful country. This Master’s thesis in Computer Science has been written at the School of Mathematics and Systems Engineering (MSI) at Vaxjo University. The research efforts are carried out in cooperation with the Blekinge Institute of Technology(BTH). My supervisor at BTH has been Mr Peng Zhang while at Vaxjo University my supervisor has been Dr. Marcelo Milrad.

I would like to thank Marcelo Milrad at Vaxjo University for his recommendation about the IMIS project at BTH University and for his guidance. I would also like to thank my supervisor at BTH Peng Zhang for his patient guidance and good advice. I also would like to thank Mathias Hedenborg and David Nadel who provide me with valuable help.
Abstract

Nowadays, business intelligence techniques are applied more and more often in different settings including corporations and organizations both in the private and public sector. It is really a broad field which can assist business people to realize the state of their organization and make profitable decisions.

In this thesis, I will focus on one of its components, data warehouse, by proposing activity theory as the method to solve the dimension identification problem in data warehouse. Under the background of project IMIS and the involved personnel, who determine the dimension, firstly I study how to use the ER method, “bottom up” method, and activity theory method to identify the dimension in data warehouse, and some relevant knowledge about the three methods. Then, we apply the three methods to identify the dimension. After that, I evaluate the dimension identification results of the three methods according to the feedback from the healthcare organization to get their veracity and integrality. Finally, based on the results of my efforts, I arrive to the conclusion that the activity theory method can be applied to identify the dimension in data warehouse, and with the comparison to the other two traditional methods (ER model and “bottom up”), the activity theory method is more easy and natural to identify the dimension of a dimensional model.
1. INTRODUCTION

1.1 BACKGROUND  7

1.2 RESEARCH QUESTIONS  8

1.3 HYPOTHESIS  8

1.4 METHODOLOGY  9

1.5 HOW TO MEASURE SUCCESS  9

1.6 EXPECTED RESULTS  9

1.7 LIMITATIONS  10

1.8 OUTLINE  10

2. THEORY

2.1 BUSINESS INTELLIGENCE

2.1.1 Business intelligence and one of its study cases  11

2.1.2 BI components  13

2.1.4 Process of Business intelligence  14

2.2 DATA WAREHOUSE

2.2.1 What is Data Warehouse, its chief characteristics and architecture  14

2.2.2 Dimensional model in data warehouse  16

2.2.3 Design Methodology of Data Warehouse  18

2.3 SOME METHODS TO DETERMINE DIMENSION OF DIMENSIONAL MODEL

2.4 ACTIVITY THEORY AND ENGSTRÖM’S TRIANGLE MODEL FROM ACTIVITY THEORY  21

2.5 AT METHOD TO IDENTIFY THE DIMENSION  23

3. CASE STUDY

3.1 THE IMIS PROJECT  26

3.2 USE CASE IN IMIS

3.2.1 One use case  28

3.2.2 Another use case  33

3.3 BRIEF SUMMARY  36

4. DISCUSSION

4.1 CRITERIA AND THEIR MEASUREMENT

4.1.1 Method feasibility criteria  37

4.1.2 Method learning time criteria  37

4.1.3 Method applying time criteria  38

4.1.4 Dimension veracity criteria  38

4.1.5 Dimension integrality criteria  38

4.2 EVALUATION

4.2.1 Feasibility criteria  39

4.2.2 Time needing criteria  40

4.2.3 Veracity criteria  41

4.2.4 Integrality criteria  42

4.3. COMPARISON  43
1. Introduction

Nowadays computer science is one of the most popular and active disciplines in the world, and it has made a tremendous contribution to human society. From just simple computation in the very beginning to business intelligence, intelligent systems to assisting human thought, computer science plays a crucial role in the life of human being and it can not be easily replaced. A computer is of course a powerful tool, and how to use it correctly and efficiently seems to be significant to the users. As a tool, a computer could not thrust itself in human society without any “lube”, there must be something that goes with it to make the combination with social usage seamless. In order to achieve this goal, computer scientists cannot just focus on the field of computer science but also other disciplines that intersect with it if these researches want to design more effective and efficient computer systems to support human activity.

Some human social psychology theories such as Activity Theory, is like a medium that can assist human using computer, which connects computer and human society closer together. According to the Engström’s triangle model from Activity Theory, the sketch map can be like this in figure 1.1,

![Figure 1.1 Sketch map of human apply Activity Theory into computer use to enhance the usage to get better effect.](image)

In this chapter, I introduce the background of this thesis (business intelligence, in the context of the healthcare project IMIS), the research questions I focus on (data warehouse, dimensional modeling and Activity Theory), the hypothesis I propose. This chapter ends with a presentation of the expected result and the outline of the thesis.

1.1 Background

During last decade, we have seen an increasing interest in business intelligence (BI), which is defined as “the processes, technologies, and tools needed to turn data into information, information into knowledge, and knowledge into plans that drive profitable business action” [1] BI encompasses data warehousing, business analytic tools, and content/knowledge management,” [1] by the Data Warehousing Institute. More and more business organizations apply business intelligence technology into their work in order to further realize the state of their organizations and to assess the business environment, and to make corresponding decisions to do profitable actions.
Recent development in the field of business intelligence, makes it possible that BI techniques are used more and more by different industries including healthcare organization, such as the project Integrated Mobile Information System (IMIS). IMIS is developing a computer system to aid the daily care for diabetic patients. And The efforts of this thesis are carried out in connection to the project IMIS.

According to the definition given by Kelley [2], BI is a gathering of process, technology and tools, including data warehouse (DW) which is defined as “a repository of integrated information, available for queries and analysis (e.g. decision support, or data mining)” by W.H. Inmon and C. Kelley [2]. To fully appreciate the inherent potential of BI, people should offer as good data warehouse as possible which is an important element of BI process.

Steve Hoberman states in his paper “The Importance of Data Modeling as a Foundation for Business Insight” [3], people deploy diverse techniques to fill various contextual usages, including denormalization, surrogate key, dimensionality, indexing, partitioning, and views in data warehouse. The particular efforts of my work are focused on dimensionality techniques.

1.2 Research questions

In dimensionality technique, the point is building the dimensional model which is “A type of data modeling suited for data warehousing. In a dimensional model, there are two types of tables: dimensional tables and fact tables. Dimensional table records information on each dimension, and fact table records all the "fact", or measures.” [4]. And I will give more details about it in chapter 2.

The problem is that in the most popular methodologies of data warehouse designing that are applied today, such as “nine-step methodology” proposed by Ralph Kimball, there is no standard method to identify the dimension for a dimension model. However some traditional methods, such as Entity-Relationship (ER) model method (ER model is “a way of graphically representing the logical relationships of entities/objects in order to create a database” [5]), and “bottom-up” method, are usually applied, they have some limitations or poor quality to some extent in the criteria of understandability and time needed, requirement and feasibility, integrality, and veracity. As encountered in the project IMIS, we could not identify the dimension of a dimensional model completely and accurately with these two methods in some case. And even sometimes, these two methods could not work at all.

Due to the inadequacy and unfeasibility of the ER model method and “bottom-up” method, in this thesis, I propose the activity theory method (AT method) to solve the dimension identification problem in data warehouse design. I try to apply Engström’s triangle model from Activity Theory to represent the activities in this healthcare organization, and identify the dimension of a dimensional model based on the activity theory models.

1.3 Hypothesis

Activity theory method can be applied to solve the dimension identification problem in data warehouse to support data intelligent manipulations in business intelligence. That
is, in dimensional modeling of data warehouse, dimensions can be identified by using Engström’s triangle model from Activity Theory to analyze the activities that involved.

AT method should lead to a better-fit effect compared to the other two traditional methods in time needs, feasibility, veracity and integrality criteria

1.4 Methodology

In the thesis, the method to conduct my research is;

Firstly, let the involved personnel, who determine the dimension in the project IMIS, study how to use the ER method, “bottom up” method, AT method to identify the dimension in data warehouse, and some relevant knowledge about the three methods (in this step, we record the time needing of learning time criteria).

Secondly, I apply the three methods to identify the dimension in data warehouse design in the project IMIS (in this step, we record the time needing of applying time criteria).

After dimension identification, I evaluate the results of the three methods according to the feedback from the healthcare organization, which is collected by using questionnaire and interview, to get their feasibility, veracity and integrality.

1.5 How to measure success

I measure the success of my proposed method according to the feedback from the healthcare organization. If the results of applying the AT method can meet the great mass of dimension needed of the healthcare organization, it is successful. And I compare the AT method with the other two traditional methods (ER method and “bottom up” method) in the criteria of time needs (learning time and applying time), feasibility, integrality, and veracity. If the AT method leads a best fit model method compared to the other two traditional methods in the criteria above, I understand that the AT method is more suitable/recommended. I will particularly clarify the reason I choose these four criteria and the measurement of them at the beginning of chapter 4.

1.6 Expected Results

Based on theoretical discussions and testing the AT method in the use case of project IMIS, I expect to get the result that can easily identify the dimensions of a dimensional model in data warehouse by applying the AT method. In comparison to the two other traditional methods (ER model method and “bottom up” method), AT should lead to a better performance based on the four criteria (understandability and time needing, requirement and feasibility, integrality and veracity), that is the AT method is more natural for human thinking, easier to apply without so much requirements about other components of IMIS (e.g. source databases) or tools (e.g. CASE Tools), and to get better qualified results. It is expected that the AT method should perform better than the ER model method and the “bottom up” method, the data warehouse designed with it would be better.
1.7 Limitations

In this paper, I propose a theoretical method for identifying the dimension of dimensional model in data warehouse design. It is based on a theoretical argumentation and the results derived from the application of these ideas within the framework of the IMIS project.

1.8 Outline

In this thesis, I first give a description of the theories, concepts and methodologies involved, and then propose the activity theory method in chapter 2. I test this activity theory method in the use cases of IMIS in chapter 3. Then I compare the three methods (ER model, “bottom up” and activity theory) with the aspects (time needs, feasibility, integrality, and veracity) in chapter 4 and give the my conclusions in chapter 5 and future work in chapter 6.
2. Theory

In this chapter, I firstly give a brief description about business intelligence, its main components and its process with a study case in 2.1. Then in section 2.2 I focus on data warehouse, I introduce its main characteristics, architecture in 2.2.1, dimensional model in it in 2.2.2, and its design methodology in 2.2.3. Besides, in the end of section 2.2, I also point out the problem that there is no standard method of dimension identification circumstantiated in the “nine-step” methodology proposed by Ralph Kimball. In 2.3, I introduce some traditional methods to solve the problem of dimension identification in data warehouse, and Activity Theory (AT) in 2.4, then propose the AT method to solve the dimension identification problem in 2.5.

2.1 Business Intelligence

This section is a description about business intelligence in 2.1.1, its components in 2.1.2 and its process in 2.1.3.

2.1.1 Business intelligence and one of its study cases.

Business Intelligence is neither a product nor a system but a process that converts the organizational data into information, and converts information into knowledge that can assist people in the organization to determine their business decisions. It is a approach to make good use of the organizational data asset to further realize the state of the organization and to assess the business environment, to make corresponding decisions to do profitable actions. “It is an umbrella term that combines architectures, applications, and databases. It enables the real-time, interactive access, analysis, and manipulation of information, which provides the business community with easy access to business data.” [6] With BI techniques, people analyzes the historical data, which are generated through the past business transactions or other kinds of activities, to further realize the past and present business situation and performances. “By giving this valuable insight, BI helps decision-makers make more informed decisions and supplies end-users with critical business information on their customers or partners, including information on behaviors and trends." [6]

At a word, “BI makes the right information available in the right format to the right person at the right time.” [7]

A case study\(^1\) of business intelligence

As one of the largest home improvement multinational retailers in Europe, OBI Inc. (Germany), which was founded in 1970, generated expected revenues of more than 5.6B DM in 1997. In the intensely competitive market nowadays, the enterprise, which wants to expand and succeed, cannot afford to stand still. Thus OBI shaped a series of business strategies, like expansion and one-to-one relationship with each customer, in order to develop and exceed. Until 1998, OBI had about 500 stores in most European

\(^1\) Based on information published in DM Review Magazine, July 1998, by OBI's CEO, and publically available corporate information (in Germany, 2000)
countries and several in China. And following its most important business strategy, to develop a one-to-one relationship with each customer, OBI has been learning as much as possible about their needs and purchasing patterns in order to serve them better. This has been a pressing requirement, because the customer base is changing dramatically. A significant increase in the number of single-adult households and dual-income families, combined with changing family dynamics, is bringing more women into OBI's stores. As a result, OBI has to incorporate women's styles and preferences into store displays and product assortments--without turning away loyal male customers.

In the mid-1990s, OBI recognized that managing the franchise expansion, the growth in SKUs, the wider assortment of products and the analysis about customers demanded the help of business intelligence techniques. So OBI and Tandem, a Compaq company, have begun implementation of an enterprise-wide information system with BI techniques that links transaction detail (where, when and which items) with customer information (demographics, history and purchasing patterns). This wealth of information opens the door to product, service and customer profitability analysis, assortment planning, one-to-one marketing programs, replenishment management and other day-to-day management and service improvements.

And OBI's decision support solution from Tandem is expected to generate a number of cost-saving results that will help strengthen our company's market leadership position in Europe and improve the profitability:

**Launch one-to-one marketing.** As part of its one-to-one marketing program, OBI is tracking customer purchases through the use of "virtual loyalty" cards. Once we identify customers at the point of sale and understand their purchasing patterns--which stores are frequented and how often, and which items are purchased in which quantities--we can use that information to refine our marketing strategies and optimize our in-store assortments for the specific store demographics.

**Improved inventory management.** OBI expects to see improvements in inventory management in three areas. First, "no sales" from out-of-stock products should drop to an absolute minimum, increasing store turnover 10 to 15 percent. Second, OBI expects to shrink item overstocks, reducing inventory costs and clearance markdowns across the company's franchises. In addition, OBI intends to transfer much of the responsibility for on-time deliveries to vendors that will be linked to the system. OBI then would be able to negotiate minimum guaranteed-profit contracts, increasing vendor loyalty while yielding dramatic improvements in reduced inventory costs and on-time delivery.

**Enhanced customer service.** The near real-time system should help retail staffs respond quickly to customer queries. If an item is out of stock, an employee would be able to check the status using a simple Web-browser interface and tell the customer immediately whether the item has been ordered, when it will arrive or whether it is available at another store.

In the study case, OBI make good use of its organizational data asset to launch one-to-one marketing, improve inventory management, and enhance customer service by applying various business intelligence components. And I will introduce them in the next section.
2.1.2 BI components

**Source database** is mainly composed of internal database and external database.

Internal databases are generally the ones used for operational data store (ODS), which is designed to quickly perform (such as querying the price of an ice-cream), rather than complex queries on large numbers of data as Data warehouse supplies. In OBI, people store the operational data/information, which is generated through the current business transactions or other kind of business activities, in the operational databases.

External databases mainly refers to the data/information that we can get from external data sources, such as the data/information from web, the databases of other organizations. In OBI, people can collect information from the external databases like the Wall street journal or other business organizations.

**Online Analytic Tools (OLAP)** is part of the broader category business intelligence which also includes extraction, transformation and loading (ETL) in data warehouse, data mining, and relational reporting. By applying it, users can quickly get answers to the analytical queries which are dimensional in real world. Typically, the applications of OLAP are applied in business reporting for management reporting, marketing, financial and so on. “**MULTIDIMENSIONAL** is our key requirement. If we had to pick a one-word definition of OLAP, this is it.” [8] The databases for OLAP employ a multidimensional data model, for it allows for complex analytical and ad-hoc queries within a rapid execution time.

In OBI, by applying various kinds of OLAP tools, people can get answers to the analytical queries about the data/information of the organization, such as “what brand of flooring sells best in Vaxjo in May 2006?”, “reporting the purchasing record of Mr Milrad in the last 3 years?”, and “in the category of window, which three brands have the least repairs in the second half of 2005?” With the report results, the decision-makers of OBI will make some business response, such as sending more flooring of brand A that sells best in Vaxjo to the local store, and sending advertisements of brand B to Mr Milrad, which he bought most in last 3 years.

**Data mining** can be defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [9] and "the science of extracting useful information from large data sets or databases.” [10]

It is a kind of discovery with analysis of large databases in order to find previously unsuspected relationships, in which some relevant people are interested.

In OBI, with application of data mining techniques, people can discover the potential and undiscovered relationships between data. And with the data mining tools, they may get the relationships like, “most of the customers who earn over 100,000 dollars/year would like to buy the product of brand A”, and “in May, June, July and August, the sale of caliduct always fall to the neap”. And after confirmation of these relationships, decision-makers of OBI can send more advertisements of brand A to the customers who earns over 100,000 dollars/year, and reduce the inventory of caliduct in May, June, July and August.

The component **Data warehouse** will be particularly discussed in 2.2.
2.1.4 Process of Business intelligence

Figure 2.1 below illustrates the general process of business intelligence and the factors (data, information and knowledge) involved. In the process of business intelligence, data warehouse collects correlative data/information from many different data sources. With the extraction, integration and loading (ETL), large numbers of current and historical information is stored in this repository. The decision makers or some other apply some techniques, such as OLAP, data mining, to discover the underlying knowledge that hide in the information. With the know ledges derived from data, it is likely to be more easy for decision-maker to make profitable business decisions.

In the case of OBI, people firstly collects a mass of data from different data sources including internal databases and external databases into data warehouse with extraction, transformation, loading in order to get data integrated and qualified. And then, people apply various business intelligence techniques, like OLAP, data mining, to make use of these historical, integrated and qualified data/information in data warehouse to get the outcome that can assist decision-makers to make profitable business decisions.

![Diagram of BI process](image)

*Figure 2.1 The process of BI and the factors to go with*

2.2 Data warehouse

In this section, I introduce its main characteristics, architecture in 2.2.1, dimensional model in it in 2.2.2, and its design methodology in 2.2.3.

2.2.1 What is Data Warehouse, its chief characteristics and architecture

“Warehousing is an emerging technique for retrieval and integration of data from distributed, autonomous, possibly heterogeneous, information sources. A data warehouse is a repository of integrated information, available for queries and analysis (e.g. decision support, or data mining)” [2]

Data warehouse is a repository which collects data/information from different data sources, including internal ones (e.g. ODS of an organization) and external ones (e.g. information from web or ODS of other organizations), current ones and historical ones (maybe 5 years ago). I think it can be regarded as the base of business intelligence
which supplies data/information for the queries and analysis of business analysis activities and decision-making tasks.

“A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.” [11]

Here are the four chief characteristics of data warehouse as follows;

Subject-oriented,

Data warehouse is oriented to the major subject areas of an organization, which have been identified in the data model, such as customer, furniture, insurance premium, and counterclaim in an insurance agent.

Integrated,

Because the information in data warehouse is from Heterogeneous data sources, the data has their own format in those data sources. When data is moved into the warehouse, it is converted to be in integration.

Time –variant,

The time horizon for the data warehouse (5~10 years) is significantly longer than that of operational systems (generally 60~90 days).

Nonvolatile,

Data warehouse is accessed and loaded, but the update of data does not occur in it.

Here is the architecture of data warehouse in figure 2.2 which also shows the process of how data warehouse works. At the very beginning, data are in diverse source databases including internal ones like operational databases of an organization and external ones like databases of other organizations or from web. The data warehouse collects the information from the source databases with extraction, transformation, loading (ETL) in order to get data integrated and qualified. Then Data warehouse builds data marks for particular subject matters. And by applying business analytic tools (Reports, OLAP, Data mining) to process the data / information, we can get some knowledge that is helpful for business decision making.
2.2.2 Dimensional model in data warehouse

The logical design technique often applied in data warehouse is called dimensional modeling, which is different from and contrast with ER modeling usually applied in relational database design. And “the dimensional model has a number of important data warehouse advantages that the ER model lacks.” [13]

Let us back to the case study of business intelligence in section 2.1. In OBI, there is a series of common business activities that involved the customer, product (various furnishings), vendor (furnishings producing factories), order form and consignment. They are, the customer gets order form offered by vendor, and the consigner (chain stores of OBI) implements the consignment according to the order form to send product to the customer. In order to more realize the organizational state about these business activities, OBI collects relevant data/information of these entities into data warehouse for further analysis.

If people still apply ER modeling in the data warehouse design of OBI about the case above, it leads to a kind of complanation. All the entities seem to be “equal” in the ER model like in figure 2.3. But in fact, because of some practical reasons, some entities in data warehouse are not “equal” to others, and they require particular management.
Then let us have a look at a three-dimensions view of these entities according to theirs record quantity in figure 2.4.

As we see in this three-dimensions view, the entities stand for customer, product, vendor and consignment are sparsely loaded, whereas the order entity is largely loaded. That means there will be a large number of data in order entity table, however less data will be in the entities of customer, product, vendor and consignment tables. We should apply a particular way to deal with the entity order, which holds a large amount of data differs from others.

Because of the inadequacy of ER modeling, designers of data warehouse apply dimensional modeling to solve this problem.

“A dimensional model is a form of data modeling that packages data according to specific business queries and processes. The goals are business user understandability and multidimensional query performance.” [14] It is a technique generated because of the way human thinking and the using context of data (multidimensional query
performance, such as OLAP, reporting).

“Each dimensional model is composed of one table with a multipart key, called the fact table, and a set of smaller tables called dimension tables. Every dimension table has a single-part primary key that corresponds exactly to one of the components of the multipart key in the fact table.” [13]

A fact table records something tangible that can be reported, and has a multipart primary key made up of two or more foreign keys. It contains one or more numerical measures, or "facts," which occur for the combination of keys that define each record, such as the order entity in figure 2.3.

As Baya Pavliashvili mentioned in his book [15], a dimension table is typically very simple which contain the levels on which you want to group your reports. And according to the paper “Development of a Prototype System For Accessing Linked NCES Data” by Sameena Salvucci, Stephen Wenck and James Tyson, dimension often contains descriptive textual information, whose attributes are used as the source of most of the interesting constraints in data warehouse queries, and they are virtually always the source of the row headers in the SQL answer set. [16]

In figure 2.3, the entities of customer, product, consignment and vendor are dimensions of fact entity order.

2.2.3 Design Methodology of Data Warehouse

In project IMIS, we apply the nine-step data warehouse design methodology proposed by Ralph Kimball 1996 to design the data warehouse, which are as follows and I just give descriptions of the first four steps, for they are concerned with this paper;

1. Choosing the Process

The process refers to the subject matter of a particular data mart. Data mart is the element of data warehouse, "...The data warehouse is nothing more than the union of all the data marts..."[17]. "A data mart is a collection of subject areas organized for decision support based on the needs of a given department." [18] A data mart is the unit of data warehouse that focus on particular subject matter of the whole enterprise, such as finance data mart, marketing data mart.

Good beginning leads a half of success. If you want to go on data warehousing, you should exhibit the first data mart to be built as successful as possible. It should be the one which is to be delivered on time, within budget, and can answer the most ad hoc business questions within its field.

2. Choosing the Grain

“Choosing the grain means deciding exactly what a fact table record represents. Only when you have chosen the grain can you have a coherent discussion of what the dimensions of the data mart's fact table are.” [19] That means after the grain choosing of fact table, we can identify the dimensions of the fact table. The grain decision for the fact table also determines the grain of each of the dimension tables.

3. Identifying and Conforming the Dimensions

“The dimensions are the drivers of the data mart. The dimensions are the platforms for browsing the allowable constraint values and launching these constraints.” [19]

As Maciej Matysiak mentioned in his presentation [12], dimensions set the context for formulating queries about the facts in the fact table. If one dimension comes about in
two different data marts, they must be the same dimension, or one must be a subset of the other, for it is the only way that two data marts can share one or more dimensions in the same application. When a dimension is used in more than one data mart, the dimension is referred to as being conformed.

4. Choosing the Facts
“The grain of the fact table determines which facts you can use in the data mart. All of the facts must be expressed at the uniform level implied by the grain.” [19]

5. Storing Precalculations in the Fact Table
6. Rounding Out the Dimension Tables
7. Choosing the Duration of the Database
8. The Need to Track Slowly Changing Dimensions
9. Deciding the Query Priorities and the Query Mode

According to this “nine-step” methodology, the dimensions of the fact table are determined during step 2 and step 3, however there isn’t any detail about how to determine dimensions for the dimensional model. In step 2, choosing the grain means deciding exactly what a fact table record represents, such as sales amount by store by day and the foreign keys pointing to the dimension tables, and then in step 3, identifying dimensions according to the grain, such as store table (Region dimension) and day table (Time dimension). But how do the designers know to record the sales amount by store by day, which will be determined as the dimensions of this dimensional model, in step 2 in advance? It seems that the designers of data warehouse identify the dimensions because the dimensions have been already identified. Maybe you can just turn to empiricism to determine dimensions ahead, but it is apparently suspect here, for there are diverse kinds of industries in this world, and each of them has their own characteristics of dimension identification which is also various in different situations. And the purpose of business intelligence is to find out the potential knowledge underlying your organization, if you still apply experience here in this process, it doesn’t make sense. So there should be an approach applied in step 2 to determine the dimension of the dimensional model at the beginning, and then identify the grain of fact table according to what dimensions it takes.

Nowadays, some traditional methods are applied to solve this problem, but they have some limitations or poor quality to some extent. And I will introduce them in the next section.

2.3 Some methods to determine dimension of dimensional model.

Identifying the dimension of a dimensional model is not easy, especially in some complicated and capricious business industries. Although people know BI can give them potent support, how to use it effective and efficient should be much accounted of. As the following article published on DM Review [20], many people do not know dimensional model clearly, and they may be hesitating to determine the dimension.

“Question: A data warehouse is subject-oriented. Are there any rules to go about defining the subjects? ... How many subject areas should an organization typically have?

Answer: The rule of thumb is that subject areas are what the business wants to talk
"about" or the "nouns" of the business, e.g. finance, manufacturing, marketing, sales, HR, legal, shipping, etc. If you are creating an Enterprise Data Mart Architecture (EDMA) for a "bottom up" approach, you will also be defining the "dimensions" of the business at the same time as the subject areas. The dimensions are the "adverbs/adjectives" of the business. They are how the business wants to talk "about" the subject areas. The dimensions represent the ways the business wants to segment and relate the transactions/activities associated with the "nouns" such as time, geography, sales geography, product, customer, division, location, etc.

But in fact, there is no standard approach to determine the dimension of a dimensional model. Generally many of data warehouse designers determine dimension with ER models of source databases, or applying an indirect approach called “bottom up logical data modeling” to get an ER model first and then determine the dimension for fact table.

**The ER model method**
Let us use the previous case of customer, product, order and vendor of OBI again.

In the source database, the entities of customer, vendor, product, order, and consignment can be represented in ER model like this in figure 2.3, such as entity vendor has the relationship “offer” to entity order and entity customer has the relationship “get” to order.

The ER model method is intuitionistic, it determine the entity as dimension which has direct relationship to the fact entity in ER model. By viewing the ER model of the entities related to the entity (order), which is the fact table, we can know that customer, vendor, product and consignment have direct relationships to entity order, and all of them will be determined as the dimensions of this dimensional model.

But as we know, the data warehouse is an integration of many different source databases and the ER models of the source databases’, which will be applied to determine the dimension, may not remain because some of them are really old. And the designers of data warehouse may be not the ones who design the source databases, so it is not feasible to apply ER model in this situation.

**The “Bottom up logical data model” method**
Due to the absence of ER models and designers of source databases, someone applies the “bottom up logical data modeling which refers to the painstaking task of normalizing existing process-oriented (denormalized) data structures into a “best-guess” logical data model.” [3]

This method has 4 steps, which are as follows:
1. Review record layouts and DDL to locate potentially embedded entities.
2. Trace through primary and foreign keys to identify inferred entities and their underlying relationships among others.
3. Painstakingly apply normalization rules to every data element.
4. Convert technical column and table names into the ones that are qualified for business usage.

In fact, this method is first to apply some computer tools (like CASE tools) to produce a “best-guess” logical data model which is similar to the real logical model that
does not remain, and then identify the dimensions with the “best-guess” logical model. If we apply this method to the example in figure 2.3, the process is that using some tools like CASE tools to try to recover the logical model of the entities customer, vendor, product, order and consignment, and then determine the dimension according to the “guessed” logical model. In addition, not all the use cases can apply this method, if the source database is object-oriented, network database or hierarchical database. this method does not work (this reason will be presented later in the discussion of chapter 4). In this thesis, “bottom up” method is regarded as an theoretical method without real implement.

According to brief theoretical discussion above about these two traditional methods, we can find that both of them have some limitations (ER model method’s failure because of the ER model’s absence, and “bottom up” method’s failure because of the multiform data models of source databases) or poor quality (ER model method identify dimension according to part ER models but not integrated organizational one, and “bottom up” method identify dimension according to “guessed” logical model) to some extent. So we change our view to Activity Theory, which is a theory from human society, in order to discover another method to identify the dimension of dimensional model in data warehouse.

2.4 Activity Theory and Engström’s triangle model from Activity Theory

The following paragraph is written by consulting “A conceptual framework based on Activity Theory for mobile CSCL”. [21]

Activity Theory is a theoretical framework for analyzing human practices as developmental processes with both individual and social levels interlinked at the same time [22]. It uses “activity” as the basic unit to analyze behaviors of human beings. AT has made significant contributions to the fields of human–computer interactions [22], and network communication and education [23].

AT is not a methodology [24] but a theoretical framework for analyzing human practices in a given context, which cannot be understood or analyzed outside the context in which it occurs. Activity, which means, “what people do”, is reflected through people’s actions as they interact with their environment. Both individual and social levels are interlinked while at the same time providing an alternative way of viewing human thinking and activity.

The AT framework uses “activity” as a basic unit for studying the behavior of human beings and it has the idea that the relationship between the subject and the object is not direct but rather mediated through the use of a tool.

A tool can be something physical (e.g., X-Ray machine, the monitor device) or intellectual (e.g., the Database SQL aggregate functions). Physical tools are used to handle or manipulate objects while intellectual tools can be used to influence behavior in one way or another.

Vygotsky (1978) originally introduced the idea that human beings’ interactions with their environment are not direct but instead are mediated through the use of tools and signs, and it was developed further by Leont’ev (1981), who created a hierarchical model for analyzing an activity. Inspired by this analysis, Engeström (1987) extended Vygotsky’s original conceptualization for the mediated relationship between the subject
and the object by introducing an expanded version of the activity triangle model that also incorporates Leont’ev’s concepts. Thus, Engeström offers a general model of human activity that reflects its collaborative nature. The model’s components, shown in Figure 2.5, are:

1. **Object in the activity** (‘raw material’ or ‘problem space’, i.e.),
2. **Subject in the activity** (i.e., doctors, nurses),
3. **Tool mediating the activity** (anything physical, e.g., computers; or mental, e.g., models or heuristics used in the transformation process),
4. **Rules and regulations** (norms that circumscribe the activity),
5. **Division of labor** (e.g., actions undertaken by individuals within the group versus tasks that are a group responsibility),
6. **Community** (individuals directly or indirectly involve in the tasks)
7. **Outcome** (i.e., the results and final products of the defined objectives).

![Figure 2.5: The triangle model of activity [25].](image1)

Figure 2.6 describes a medical activity example represented with Engström’s triangle model, it is a work activity of a physician working in a care clinic. In this activity, the patient with health problems and illness is the object. The outcomes include health improvements and recoveries which we expect, and the unpleasant outcomes such as poor-qualified treatment, disrecovery, and dissatisfaction of both physician and patient. The tools include powerful instruments such as medical devices, medicine, operating rooms, medical-record as well as some related methods and concepts. The community consists of the clinic and staff. The division of labour shows the tasks and potency of the staff involved in. Finally, the rules are criteria that regulate the time used, outcome measurement, and rewards.

![Figure 2.6 An medical activity example](image2)
### 2.5 AT method to identify the dimension

If we wish, we can apply Engström’s triangle model from activity theory to represent all the activities of an organization, which will also represent all the elements involved in this organization at the same time, such as staff, regulations, customer, and devices.

According to these Engström’s triangle models, it is very easy to see what entities compose these activities, the relationships among entities, and which entities are the factors that can influence others.

The activity theory method can be executed in two ways, which are:

**Way one**, which is like “top-down”;

*Step 1:* Representing all the activities relevant to this subject matter of the organization.
*Step 2:* Filtering the activity model, and picking out the ones in which entity of fact table involved.
*Step 3:* Identifying the dimensions according to the components that are related to the fact table entity in these activity models.

But sometimes, the source databases may not have record for all the identified dimension, so we have to abandon some dimensions. Or apply the other way at the beginning.

**Way two,** which is like “bottom up”,

*Step 1:* Representing all the activity models relevant to this subject matter of the organization
*Step 2:* Filtering the activity models, remaining the ones in which entity of fact table involved
*Step 3:* Identifying all the entities relevant to the subject matter in source databases
*Step 4:* Introducing the entities into the filtered activities models, remaining the ones that are components of these activity models.
*Step 5:* identifying these entities as dimensions.

In fact, both of the above two have the same clue that is applying activity theory model to identify the dimension of dimensional model. You can choose them based on the context. As the combination of “top-down” and “bottom-up” in other fields, these two ways can be both applied to be interactants.

Let’s use Engström’s triangle model from Activity Theory to analyze the case of OBI mentioned above in section 2.2.2.

Figure 2.7 is the activity theory model represents the activity that vendor offers and gains order by using the tools like computer and telephone in the environment of net and telecom. The vendor must obey the laws of the environment net and telecom, and different roles have different task.

![Figure 2.7 Vendor uses computer / tell to offer order](image)
Figure 2.8 is the activity theory model represents the activity that customer gets order by using the tools like computer and telephone in the environment of net and telecom. The customer must obey the laws of the environment net and telecom, and different roles have different task.

![Network and Telecom Model](image)

*Figure 2.8 Customer uses computer/tell to get order*

Figure 2.9 is the activity theory model represents the activity that vendor does consignment according to order in the environment of storehouse or some other place. Everyone must obey correlative laws and different roles have different task.

![Vendor and Consignment Model](image)

*Figure 2.9 Vendor uses order to do consignment*

Figure 2.10 is the activity theory model represents the activity that customer uses order as evidence to get products in the environment of storehouse or some other places. In the same way, everyone involved must obey correlative laws and different roles have different task.

![Customer and Product Model](image)

*Figure 2.10 Customer uses order to get product*

Because I just want to analyze the activities whose components present in the example, I only present these 4 activities in the organization.

According to the activities, we can see that customer, product, vendor, consignment have direct relationships with order which has large amount of data. And these 4 entities
can influence order in real world. So we can determine that vendor, product, customer and consignment are the dimensions of order.

By the way, I just omit the environment component which also has direct relationship with order form, and in fact, this component can stand for the dimensions such as time, geography, weather which involved. This component will be detailed discussed in the use case.

And we can get the dimensional model of the case in figure 2.11

![Dimensional Model](image)

*Figure 2.11 The dimensional model of the case in OBI*

By and large, this chapter is an introduction about the theories and concepts involved in this thesis. I first introduce business intelligence with a case study of OBI about its function, components and process. Especially in the component of data warehouse, which is the research domain of this thesis, I propose the AT method to solve the dimension identification. And with brief theoretical discussion about the AT method and two other traditional methods (ER method and “bottom up” method), I find that in some aspects like feasibility, the quality of identification and time needing, the ER method and “bottom up” method have some deficiency to some extent, whereas the AT method leads better performance. And in the next chapter, I will apply the AT method into the use case of the project IMIS, which is the background of this thesis, to test it and to compare with the other two traditional methods.
3. Case study

In this chapter, I will introduce the core ideas of this thesis as I will present the IMIS project. IMIS applies BI techniques to support and provide the right information to people working in the health care sector. In the data warehouse design of the project IMIS, we also encounter the dimension identification of dimensional modeling, and we find that we could not get complete and high efficient result if just applying the traditional methods mentioned in chapter 2 (ER method and “bottom up” method). So we apply the AT method into the project, to test and compare it with the traditional methods in the use case.

3.1 The IMIS project

As one of the ten worst healthy problems in the world, diabetic is becoming a fearful health killer in our life [25]. It is a terrible chronic disease which may bring so many bad complications, such as cardiopathy, high blood pressure, nephropathy and hypoglycemia. Any kind of these complications can do a lot of harm to human being, even death.

The project Integrated Mobile Information System (IMIS) financed by VINNOVA\(^2\) is developing a computer system to aid the daily care and medical treatment for diabetic patients by adopting computer intelligence. The IMIS is to provide the diabetic patients with a mobile-network communication platform for homecare supervision, self-treatment, and preparation before face-to-face diagnoses.

Figure 3.1 give an overview of the IMIS’s architecture

![Figure 3.1: The view of IMIS’s architecture [27]](image)

---

2 Swedish Governmental Agency for Innovation Systems, Http://publiceng.vinnova.se/
In order to enhance the service quality, the project IMIS applies business intelligence technology to make good use of the data asset of the healthcare organization. With BI, we can further realize the states of the care receivers in the healthcare organization and make better remedial decisions for them.

And the structure of IMIS can be predigested in figure 3.2 according to the relevancy of this thesis.

In IMIS, we apply operational database systems (ODS) to store the operational data/information (such as the physiological information about diabetics, and the personal information about care provider) from care receivers and care providers, and care providers and care receivers can also query the data/information from the ODSs. Data warehouse is applied to collect the data/information from the source databases (ODSs) in this healthcare organization with extraction, transformation, loading (ETL) in order to get integrated and qualified data/information. Then Data warehouse builds data marks for particular subject matters. And by applying business analytic tools (Reports, OLAP, Data mining) to process the data/information, we can get some helpful knowledge, which can be stored in knowledge management system (KMS). And the knowledge can assist the managers of this healthcare organization to make profitable business decisions and assist care providers further realize the states of the care receivers to make better remedial decisions.

![Figure 3.2 Business intelligence in IMIS](image)

### 3.2 Use case in IMIS.

In this section, I present two use cases of the project IMIS to test AT method, ER model method and “bottom up” method.

First of all, people, who determine dimension (care providers and managers in this healthcare organizations), have a study about these three methods (ER method, “bottom up” method and AT method). And on average, people spent several hours learning ER method, for none of them has any preknowledge about computer or data modeling.
before, while they spent just less than one hour learning AT method, for as a human social theory, AT method is easy to understand and link to practice. Nobody had interesting to study “bottom up” method, for CASE tools and normalization seems to be baffling to them.

And then, we come into use case.

3.2.1 One use case

One purpose of IMIS project is to enhance the quality of face-face treatment of care provider (doctor, nurse, etc) and care receiver (diabetic). In order to achieve this final purpose, the managers of the healthcare organization should know the face-face treatment and its relevancy in depth firstly, and then make relevant decisions to improve the state.

As I mentioned in introduction, project designers of IMIS apply business intelligence into this project in order to make good use of data asset in organization to get profitable actions. And they apply this technology into the subject face-to-face treatment in order to improve its quality. They collect relevant information of face-to-face treatment into data warehouse, and apply OLAP, report tools to get some high performance queries, such as how many treatments in June 2006 (Time dimension), how many treatments in Ronneby (Region dimension). In the component of data warehouse, we apply the “nine-step” methodology to design the data warehouse. And we design the data mart about the subject matter of face-to-face treatment.

According to the characteristics of fact table, it is apparent that the entity face-to-face treatment is the fact entity, for it has a large amount of tangible numeric data that can be reported. And then, in step 2 and step 3 of the “nine-step” methodology, we should determine the dimensions of this fact.

In project IMIS, we firstly apply the two traditional methods mentioned in 2.2.4 to this use case.

ER model method.

Viewing ER model is the core of this method, so let us have a look at the ER models of the source databases in this case. Because the whole ER models in the healthcare organization are very large and complex, I just represent part of the ER model that involves the fact entity. The part ER model is in figure 3.1, which represents the relationships between the fact entity face-to-face treatment to the entities care provider, care receiver, medicine and device.
From this part ER model of the IMIS, we can see that the entities care provider, care receiver, medicine and device have direct relationships to the fact entity, and according to the ER model approach we identify the entities care provider, care receiver, medicine, and device as the dimensions of the fact table face-to-face treatment entity. In project IMIS, this ER model, which would be the only stuff for us to identify the dimension, is just one part of a source database in the big healthcare organization. It seems that the other existing ER models of the source databases have nothing to do with the dimension identification. That is not logical, for this method would lead to dimension missing. The even more terrible situation is that the ER models of some source databases do not remain. So we cannot complete dimension identification with this method. And with ER model method, we spent several days to get the result. We first spent days collecting all the source databases’ ER models, then pick out the ones that relate to fact entity for hours, and identify dimension according to the ER models within an hour.

*The “bottom up” method.*

In the “bottom up logical data modeling” method, we can apply CASE tools which have a common feature with this approach, because both of them have something like reverse engineering. CASE tools produce a physical data model of the underlying database by tracing through the primary and foreign key relationships between tables in database (step 2), and then calling the tables “entities” and showing the traced key relationships as data relationships among entities. There is no analysis, normalization, or business name qualification in the course. We can regard this physical data model produced by CASE tool as a starting point to approximate “real” logical data model with normalization process.

We just apply this method as a theoretical method but not implement it, because it costs too much to really implement, and some of the source databases in this healthcare organization are object-oriented database, network database and hierarchical database but not all relational databases (this reason will be particularly discussed in chapter 4). And the result is an ER model that is similar to the real ER model but not really the same. Because it is doubtful that tools will ever be able to mimic the human reasoning and judgment processes that are intrinsic to bottom-up logical data modeling.
Due to the inadequacy and unfeasibility of the ER model method and “bottom-up” method behave in the project IMIS, we turn to applying AT method to solve the dimension identification of data warehouse.

**Dimensional model of IMIS based on AT**

In IMIS, we apply the method based on Activity Theory in 2.5.

First of all, we represent all the activities relevant to the subject matter face-to-face treatment in this healthcare organization with Engström’s triangle model from Activity Theory, such as the model in figure 3.2. The care provider (doctor, nurse) uses telephone to contact with care receiver to inform the face-to-face treatment in the community of telecom, obeying the law of network.

![Figure 3.2 An example activity model in IMIS](image)

(It sounds time needing to represent all the activity model of the healthcare organization, but when you are familiar with this method, you can only represent the ones that involves to the fact entity.)

Second we filter all the activity models, remaining the ones that contain fact table entity face-to-face treatment, such as the models express the activities as follows in figure 3.3;

*The care provider (doctor, nurse, technician) applies medical device (X-Ray, blood-pressure meter) to face-face treatment (FTFT) in the community (hospital, clinic) in figure a.*

*The care provider uses face-to-face treatment to cure care receiver in the hospital in figure b.*

*The care provider applies medicine to face-to-face treatment in the hospital in figure c.*

*The care provider determines the face-to-face treatment according to the level (low, middle, high) of care receiver (diabetic) in figure d.*
In fact, there may be more activities related to this subject, but I am not the employee of this organization who is much more familiar with this medical case, so I just represent main ones, but I think it is enough to identify the dimension.

According to the activity models in figure 3.3, we can clearly see that the entities device, care provider, care receiver, hospital, medicine, and level of diabetic have direct relationships to the fact table entity face-to-face treatment. And according to the Activity Theory method in section 2.3.1, in an activity model the components which are related to the entity fact table can be determined as the dimensions of the dimensional model. So the entities care provider, care receiver, device, medicine, level of care receiver and hospital can be the dimensions of this dimensional model.

But this is not the end, we just collect the skin-deep stuff of this method, in fact there would be something more to dig in this activity theory. As general knowledge, activity must happen in a context or environment which stands for something such as weather, time and region. And these issues are very important dimension for a fact table. So
when viewing the Activity Theory models, these two issues will be spontaneously kept in mind.

So we get the dimensional model like this in figure 3.4

![Figure 3.4 Dimension model of use case one](attachment:image)

With the AT method, we spent hours to get the result in this use case. Firstly, we spent hours to all together relevant members of the healthcare organization to present all the relevant activity theory models for hours, and then identify the dimension of face-to-face treatment according to the model within an hour.

**Feedback**

This is the feedback from the healthcare organization about the results of the methods. I evaluate the success of the method according to its feedback from the healthcare organization. In this process, I designed and used a questionnaire in order to know that if the method result meets the business dimension needs of this organization. In this questionnaire, the care providers and managers of this organization show that they have the business dimension needs as follows:

*How many face-to-face treatments in each month of 2005? (To evaluate the workload of the months)*

*How many face-to-face treatments did doctor Gao participate in last week? (To evaluate the care provider's working state)*

*How many face-to-face treatments did Mr. Smith (diabetic) have in each last four months? (To evaluate the trend of this care receiver)*

*How many face-to-face treatments did each hospital have in Ronneby last month? (To evaluate the workload of each hospital)*

*How many face-to-face treatments use brand “A” X-Ray device? (To evaluate the device and decide future order)*

*How many face-to-face treatments use brand “C” medicine? (To evaluate the medicine and decide future order)*

According to the feedback from healthcare organization, it is possible to see that ER method result just covers round 60% of business dimension needing of this use case, whereas AT method not only covers all the dimension needs of the organization, but also
propose other 2 potential dimensions (weather and level). And these 2 dimensions will be used to solve the report like; how many face-to-face treatments happened in sunny day, or windy day, or rainy day? (To research the relationship between weather and diabetic), how many face-to-face treatments did early diabetics, medium-term diabetics, or terminal diabetics have last month? (To research the state of different level diabetics). And these two dimensions have been validated by the organization.

3.2.2 Another use case

In the healthcare organization, there is another fact entity, which has a lot of data record, called prescription. If care providers want to well understand their patients, the entity prescription is really a good fact to investigate. In project IMIS, we also apply BI to process this subject matter. We apply data warehouse to collect relevant information to entity prescription from different source databases, and apply OLAP, report tools to get some high performance queries, such as how many prescriptions did doctor Gao make (Care provider dimension), how many prescriptions are made in June 2006 (Time dimension). In the component of data warehouse, we apply the “nine-step” methodology to design the data warehouse. And we design the data mart about the subject matter of prescription.

At the beginning, we also apply ER model method and “bottom up” method to solve the problem, but there is no ER model still remaining which involves prescription entity. So we can’t apply this method here.

And then we turn to the AT method.

First of all, we represent all the activities relevant to the subject matter prescription in this healthcare organization with Engström’s triangle model from Activity Theory.

Second we filter all the activity models, remaining the ones that contain prescription entity, such as the models express the activities as follows in figure 3.5;

- The care provider (doctor) makes prescription with the care receiver’s state of illness in figure a.
- The care receiver (diabetic) get/ buy medicine with the prescription in figure b.
- The care provider (technician) applies medical device to implement the prescription in the hospital in figure c.
- The care receiver gets therapy (X-Ray test, injection) according to prescription in figure d.

![Diagram of care processes](image-url)
Figure 3.5

(a) Care provider make prescription according to the state of illness.
(b) Care receiver get medicine according to prescription.
(c) Care provider applies medical device to implement prescription
(d) Care receiver gets therapy according to prescription.

According to the Activity Theory method in section 2.3.1, in an activity model the components which are related to the entity fact table can be determined as the dimensions of the dimensional model. So the entities care provider, care receiver, medical device, medicine, state of illness, therapy and hospital can be the dimensions of this dimensional model. And with the inherent characteristics of activity mentioned in last section, that is activity must happen in a context or environment which stands for something such as weather, time and region. So we would not miss the dimension like time, weather and region. And we can get the dimension of entity prescription in figure 3.6.
With the AT method, we spent hours to get the result in this use case. Firstly, we spent hours to all together relevant members of the healthcare organization to present all the relevant activity theory models for hours, and then identify the dimension of prescription according to the model within an hour.

**Feedback**

This is the feedback about the results of these three methods from the healthcare organization in this use case. And we evaluate the success of the method according to its feedback from the healthcare organization. In this feedback, I apply questionnaire to know that if the method result meets the business dimension needs of this organization.

In this questionnaire, the care providers and managers of this organization show that they have the business dimension needs as follows:

*How many prescriptions made in each month of 2005?* (To evaluate the workload of the months)

*How many prescriptions did doctor Gao make last week?* (To evaluate the care provider’s working state)

*How many prescriptions did Mr. Smith (diabetic) have in each last four months?* (To evaluate the state and trend of this care receiver)

*How many prescriptions did each hospital have in Ronneby last month?* (To evaluate the workload of each hospital)

*How many prescriptions recommended diabetic to apply device ‘A’ treatment?* (To evaluate the device and decide future order)

*How many prescriptions recommended brand ‘C’ medicine?* (To evaluate the medicine and its curative effect)

*How many prescriptions recommended applying therapy ‘I’?* (To evaluate the efficacy of therapies)

According to the feedback from healthcare organization, it is possible to see AT method not only covers the entire dimension needs of the organization, but also propose other 2 potential dimensions (weather and state of illness). These 2 dimensions will be
used to solve the report like; how many prescriptions happened in sunny day, or windy day, or rainy day? (To research the relationship between weather and diabetic), how many prescriptions did different states of illness diabetics have last month? (To research the state of diabetic who is in different illness of state). And dimension state of illness has been validated by the organization.

3.3 Brief summary

With use cases in project IMIS, we can get the result that AT method has mostly fulfilled the dimension needing of the healthcare organization, and it is a successful method according to the criteria described in section 1.4. In the test use cases above, it is possible to see that AT method leads a more understandable way and gets much more integrated and better quality results, whereas ER model method and “bottom up” method do not perform well or even fail. I will compare them in the next chapter according to the relevant aspects in relation to the problems we encounter in project IMIS.
4. Discussion

In this chapter, I first propose the criteria and their measurements in section 4.1 which will be applied in the evaluation and comparison of ER method, “bottom up” method and AT method. And then, give detailed evaluation and comparison according to the four criteria in section 4.2.

4.1 Criteria and their measurement

In this section, I propose the criteria to evaluate the ER model method, “bottom up” method, and AT method with measurement of each criteria.

4.1.1 Method feasibility criteria

Before applying one method to solve a problem, it is necessary to make sure that the requirements of this method has been satisfied, and if the requirement cannot be contented, this method will fail sooner or later.

In other words, the satisfaction of method requirement determines the feasibility of the method.

And the measurement of the criteria is:

To what extent requirements this method has, and the scale numbers corresponds to the following values: (The following estimation scale is made by consulting the chapter 7 “Design for particular purposes: Evaluation, Action and Change” in book “Real world research” [28], and so do the next 4 criteria in section 4.1.2, section 4.1.3, section 4.1.4 and section 4.1.5.)

1 – very bad (requirements haven’t been met at all)
2 – bad (only some requirements have been met, and need extra tools)
3 – medium (some requirements have been met)
4 – good (requirements have been fully met)
5 – very good (this method has no requirement at all, or requirements can be easily fulfilled)

By quality measurement of the method’s requirement, we can evaluate the feasibility of this method.

4.1.2 Method learning time criteria

Before applying one method to solve problem, people should firstly study how to apply the method. An understandable method is of course popular for the users, whereas an impenetrable method is embarrassing. And the method learning time is the criteria to evaluate whether the method is easy for user to study or not.

And the measurement of the criteria is:

Before applying, how much time do the people, who determine dimension, spend to learn this dimension identification method. The measurements is in hour, and the scale numbers corresponds to the following values:

1 – very bad (long time to study and have to accumulate experience)
2 – bad (long time study, for couples of week)
3 – medium (receivable time, for several days)
4 – good (learn within a very short time, for several hours)
5 – very good (do not need to learn or learn within one hour) [28]

By time measurement for the method learning, we evaluate if this method is easy, medium or hard to understand for its users.

4.1.3 Method applying time criteria

When applying one method to solve problem, an issue that people usually care about is the time applying. An easy and timesaving method is of course popular for the users, whereas a hard and time-consuming method is embarrassing.

And the measurement of the criteria is;

During applying, how much time the dimension determiners spend to identify dimension with the method.

The measurements is in hour, and the scale numbers corresponds to the following values:
1 – very bad (get result for very long time, for months)
2 – bad (get result for long time, for weeks)
3 – medium (get result for receivable time, for days)
4 – good (get result for short time, for hours)
5 – very good (get result immediately within an hour) [28]

By time measurement for the method applying, we evaluate if this method is easy, medium or hard to apply for its users.

4.1.4 Dimension veracity criteria

After applying one method to solve problem, the veracity of the result is what people care about. A good method usually leads veracious result, whereas an bad method usually results in poor quality result.

And the measurement of the criteria is;

Correctness of the dimension identification with this method according to the business dimension needs.

The measurements are in percentage, and the scale numbers corresponds to the following values:
1 – very bad (0%~20%)
2 – bad (20%~40%)
3 – medium (40%~60%)
4 – good (60%~80%)
5 – very good (80%~100%) [28]

By percentage measurement for the method applying, we evaluate if this method leads a veracious result or results in a poor quality result.

4.1.5 Dimension integrality criteria

As the reference at the beginning of section 2.3, “the dimensions are the "adverbs/adjectives" of the business. They are how the business wants to talk "about" the subject areas. The dimensions represent the ways the business wants to segment and relate the transactions/activities associated with the "nouns"”[16].

According to this reference, we can see that the dimension is the point of view from
business, which want to analyze their organizations in their business subjective perspectives.

So we evaluate the method by its result, to see if the result of this method fulfills the business dimension needing. And the measurement of the criteria is;

According to the business dimension needing, how many percentage dimension we get with this method, and the scale numbers corresponds to the following values:

1 – very bad (0~20%)
2 – bad (20%~40%)
3 – medium (40%~60%)
4 – good (60%~80%)
5 – very good (80%~100%) [28]
6 – perfect (not only mostly fulfill the business dimension (80%~100%) right now, but also propose some potential dimension that can be used in future)

By percentage measurement of the methods dimension identification result to the business dimension needing, we evaluate the integrality of this method.

4.2 Evaluation

In this section, I will evaluate the ER method, “bottom up” method, and AT method according to the criteria mentioned in section 4.1 with their measurements.

4.2.1 Feasibility criteria

ER model method;
The ER model method of course requires the ER models of the source databases. But as we mentioned before, data warehouse is an integrated repository based on diverse source databases. If we encounter some of the source databases which are very old and complex, it is very possible that the ER models of theirs do not remain. If we do not have the ER models, how can this method works. It must fail in this way. In project IMIS, some involved source databases in this healthcare organization are really old and complicated, and the ER model of theirs do not remain for long, even some of them had no ER model when they were designed. So it is impossible for us to completely identify the dimension according to this method.

With the use cases in chapter 3, and according to the measurement of method feasibility criteria in 4.1.1, ER method gets 3 – medium (some requirements have been met).

“Bottom up” method;
The requirement of the “bottom-up logical data model” method seems even more complicated. In this method, we do not just apply CASE tools but also do normalization. That is really difficult for businessmen to manage. “Because this process is sometimes referred to as reverse engineering, which is a common feature in CASE tools. However, the reverse-engineering capabilities of CASE tools only cover a small part of the process in bottom-up logical data modeling – and only if the existing data structure is a relational database.” (Larissa Moss. 2004) [16] That means not all the use case can apply this method, if the source database is object-oriented or some other kind, this
method does not work. In project IMIS, some of the source databases in this healthcare organization are object-oriented database, network database and hierarchical database. So it is impossible to completely identify dimension according to this method.

And according to the measurement of method feasibility criteria in 4.1.1, “bottom up” method gets 2 – bad (only some requirements have been met, and need extra tools)

**AT method;**
The method of Activity Theory totally depends on people in the organization but not on any existing systems relate to computer. It has no requirement about any other elements except human being. In a word, this approach just needs people in this organization. So its feasibility is the best in these three. In project IMIS, I am honored to invite so many members in this healthcare organization to participate dimension identification with AT method.

With the use cases in chapter 3, and according to the measurement of method feasibility criteria in 4.1.1, AT method gets 5 – very good (this method has no requirement at all, or can be easily fulfilled)

**4.2.2 Time needing criteria**

In this section, I will combine the learning time criteria and applying time criteria together as time needing criteria. And discuss learning time criteria and applying time criteria separately.

**ER model method;**
ER model is a logical data model which is intuitionistic and pellucid. It is a graphical way to represent all the entities in an organization and the relationships among them. Generally it is no trouble to understand an ER model for someone who has professional knowledge about database. But businessman like care providers and managers in this healthcare organization mainly does the dimension identification, and ER model is not so easy for them to understand and identify dimension according to them. In project IMIS, we also encounter the problem that the care providers and managers in the healthcare organization could not understand the complex and large ER model. But with the study and some help from professional data modeler, this problem can be conquered.

With the use cases in project IMIS, such as the ones we present in chapter 3, and according to the measurement of method learning time criteria in 4.1.2, ER method gets 4 – good (learn within a very short time, for several hours)

With the use cases in chapter 3, such as the ones we present in chapter 3, and according to the measurement of method applying time criteria in 4.1.3, ER method gets 3 – medium (get result for receivable time, for days)

**“Bottom up” method;**
The “bottom-up logical data model” method seems much more complex. It needs to apply CASE tools to discover and infer entities and the relationships among them. And apply normalization to the process-oriented data structure. It is really time-consuming and need professional knowledge. So I think business men do not like to apply this
method. In project IMIS, not only members in this healthcare organization but also data warehouse designers do not want to apply this method at all.

Although we haven’t really implement “bottom up” method in IMIS, we can get the theoretical evaluation.

According to the measurement of method learning time criteria in 4.1.2, “bottom up” method gets 2 – bad (long time study, for couples of week).

According to the measurement of method applying time criteria in 4.1.3, “bottom up” method gets 2 – bad (get result for long time, for weeks).

**AT method;**
The method based on Engström’s triangle model from Activity Theory is a exert of applying human social psychological theory into computer science. It is concise, pellucid and can easily relate the facts in real world. There is no trouble to understand and manage it. People who know the activity happen in real world can easily make out the model. And after getting the model, it is very easy to determine the dimension in a short time. In project IMIS, we first represent all the relevant activity models to the subject matter of data mart, and identify dimension according to them. Although this method needs a little more time than ER model method which can utilize existing ER model conveniently, it is ok enough.

With the use cases in project IMIS, such as the ones we present in chapter 3, and according to the measurement of method learning time criteria in 4.1.2, AT method gets 4 – good (learn within a very short time, for several hours).

With the use cases in chapter 3, such as the ones we present in chapter 3, and according to the measurement of method applying time criteria in 4.1.3, AT method gets 4 – good (get result for short time, for hours).

**4.2.3 Veracity criteria**

**ER model method;**
In ER model method, we determine the entities which have direct relationships with fact entity as the dimensions. And according the use cases of project IMIS, such as the use cases we present in chapter 3, the dimensions captured by ER method, are in the business dimension needing.

According to the measurement of method veracity criteria in 4.1.4, ER method gets 5 – very good (80%~100%).

**“Bottom up” method;**
The “bottom-up logical data model” method is the worst of these three in this aspect. For all it produces is just a guest by computer tools. The core of dimensional model is from human being view. It is really doubtful that tools will ever be able to mimic the human reasoning and judgment processes. But sometimes people combine this approach with top-down, so its products may be validated. And the veracity will be improved. In the project IMIS, when we identify dimension with “bottom up” method, we have to invite people relevant to the process to validate the result. Generally, most of the dimensions identified with “bottom up” method are denied.

According to the measurement of method veracity criteria in 4.1.4, “bottom up” get a
theoretical result 2 – bad (20%~40%).

AT method;
As ER model method does, the Activity Theory method takes a high quality in this aspect. For it represent all the relationships between the fact entity and other entities. So all it identifies can be the views of how human being to see the fact. And all of them can be determined as dimensions. And according the use cases of project IMIS, such as the ones we present in chapter 3, most of the dimensions captured by AT method, are in the business dimension needing.

According to the measurement of method veracity criteria in 4.1.4, AT method gets 5 – very good (80%~100%).

4.2.4 Integrality criteria

ER model method;
In ER model method, we determine the entities which have direct relationship with fact entity as dimensions, but sometimes some entities which have indirect relationships or have no relationships represented on one ER model may also be the dimensions of this fact entity. As we know, data warehouse is the repository of information from many different source databases. You can get ER models of their own database, but not an integrated one for the whole organization. In the project IMIS, we come across the situation like this: one source database does not contain all the entities of the whole process, they are stored in multiple databases. So when you get an ER model, you just get one part view of this process, this leads to something like localization to the overall determination. Figure 4.1 is an instance of this situation we encounter in project IMIS. And with this method, it is very possible to omit the relationships between some entities to fact entity.

Back to the result we got from use cases in chapter 3, ER method just captured around 60% dimension of business dimension needing.

So according to the measurement of method integrality criteria in 4.1.5, ER method gets 4 – good (60%~80%).

“Bottom up” method;
The “bottom-up logical data model” method just produces a guess of logical model by using some computer tools. It is apparently impossible to get the integrality just by itself.

And according to the measurement of method integrality criteria in 4.1.5, “bottom up” get a theoretical result 2 – bad (20%~40%).

**AT method;**
The method based on Engström’s triangle model from Activity Theory does much better to the other two methods in this aspect. With AT method, we first represent all the activity models in an organization, and all the stuff involved in one activity will be exactly represented too. So all the elements represented in models that directly or indirectly relates to fact entity will be easily captured, and no one can be escaped. But because of the lack of human being’s cognition to real world, some of underlying activities may not be realized, and some dimensions will be missed. In project IMIS, we have represented all activities we can cognize in this healthcare organization, maybe some activities have been missed, but I think all we got is enough for dimension identification. And according to the use cases in IMIS, we get the result that generally AT can mostly capture all the dimension of business dimension needing, and can propose some potential and constructive dimension that may be used in future.

And according to the measurement of method integrality criteria in 4.1.5, AT method gets 6 – perfect (not only mostly fulfill the business dimension (80%~100%) right now, but also propose some potential dimensions that can be used in future).

**4.3. Comparison**

In this section I will compare the ER method, “bottom up” method, and AT method according to the evaluation in section 4.2.

In table 4.2, I represent all the states of the three methods in each criteria. According to the theoretical argument and evaluation comparison among the three methods, we can get the result that AT method leads a better overall performance compared with the other two traditional methods in the criteria of feasibility, learning time, applying time, dimension veracity and dimension integrality.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Method</th>
<th>ER model method</th>
<th>“Bottom up” method</th>
<th>AT method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method feasibility</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Method learning time</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Method applying time</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Dimension veracity</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Dimension integrality</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

*Table 4.2 The comparison of the three methods*
5. Conclusion

In the IMIS project, I faced the dimension identification problem of data warehouse with traditional methods (ER model method and “bottom up” method). As an alternative solution, I proposed AT method to solve this problem. By providing a theoretical argumentation and testing these ideas with a couple of use cases connected to the project IMIS, I arrived to the conclusion that AT method can be applied to solve the dimension identification of data warehouse in the healthcare organization. According to those theoretical arguments, evaluation and comparison among the three methods in the criteria of feasibility, learning time, applying time, dimension veracity and dimension integrality, we can get the result that AT method leads a better overall performance compared with the other two traditional methods. So the hypothesis formulated in section 1.3 can be verified using the results presented in the former sections as evidence.

Regarding further thinking about the usage of the AT method, it can be claimed that it performs better than the other two traditional methods, the quality of dimensional modeling has been improved and the data warehouse enhance as well.
6. Future work

In this thesis, I mainly apply AT method to identify dimension of dimensional model in the field of medical industry. But in fact, this approach could be applied to any other branches including market, finance, tourism and so on (in section 2.5, I have applied AT method into the mark/company industry for a try). For all of the industries are the organizations that consist of a set of activities, as the AT method just relies on human activity but not others. Thus, I think there will be a bright future for AT to be used in the context of these industries.
7. Reference

[1] Steve Hoberman, 2005, Intelligently design: Building a physical data model that works, Teradata magazine online.
[7] Sage Software Inc., 2005, Understanding Business Intelligence and Your Bottom Line,
[20] Douglas Hackney, August 1 1999, A data warehouse is subject-oriented. Are there any rules to go about defining the subjects, DMReview.

Some involved books and papers


The Data Warehousing Information Center, www.dwinfocenter.org

Data mining, Kurt Thearling, Http://www.thearling.com/

Center for Activity Theory and Developmental Work Research of the University of Helsinki, Http://www.edu.helsinki.fi/activity/