Eliciting Information Requirements for DW systems

Deepika Prakash
Delhi Technological University, Delhi, India
dpka.prakash@gmail.com
Supervisor: Dr. Daya Gupta

Abstract: Data Warehouse technology addresses business process analysis. However, it ignores upstream decision-making like formulating policies and policy enforcement rules. We provide a requirements engineering approach for building an integrated data for all types of decisions in an organization. To do this, we develop a two level generic platform with the bottom level having generic models of decisions, information, and decision-information association as well as information elicitation techniques for eliciting information for decisions. The higher level is the source of decision for the lower layer and is exemplified by policy enforcement rule decisions as well as operational decisions for managing the business process. Each source produces its own data warehouse requirements specification and these are integrated together using our integration technique.

Keywords: Data Warehouse, Requirements Engineering, Early Information, Decision, Policy enforcement rule, Data Warehouse Integration

1 Introduction

Data Warehouse (DW) failure statistics highlight the crucial role of RE in mitigating system failure [1]. Hayen [2] refers to studies that indicate the typical cost of a DW project to be one million dollars in the very first year. However, one-half to two-third of these projects fail. One of the causes of this failure [1] is inadequate determination of the relationship of the DW with strategic business requirements. These statistics reinforce the need for RE for DW systems.

DWRE techniques identify DW structures like facts, dimensions and finally arrive at star schema. DW structures are identified from existing systems, information gathered from users' of the DW or a combination of the two. DWRE techniques are classified into three broad categories based on the nature and phases of the process. By nature they can be top-down, bottom-up or mixed and by phases they can have a single-derivation phase or multiple phases in which the process is performed.

A more detailed comparison of top-down, bottom-up and mixed approach is as follows:

I. **Bottom-Up or Supply driven approaches** assume that existing, already available, information, needs to only be converted into the multi-dimensional form. Thus, the starting point is existing data bases and data sources. Desired facts and dimensions are then imposed on these available sources. There are two basic approaches, (a) Database Driven approach that starts from existing databases [3], and (b) ER schema driven approach that starts from ER schema [4].

ER driven techniques have been criticised on several grounds. According to [5] "Entity relation models cannot be used for enterprise data warehouses". Information is limited to what has been captured by the ER diagram. These techniques do not give primary importance to the users' perspective [6].

II. **Top-Down or Demand driven approaches** determine information contents of a DW To-Be from scratch. These approaches directly adopted model driven techniques developed in software/information systems RE like goal-orientation and scenario-orientation. User driven and Goal driven approaches of DWRE belong to this category.

Some User driven approaches include techniques developed by [7, 8]. However, it has been observed that some users may not be able to describe their requirements [9]. Users do not see their organization from a "broad angle" and so the requirements are "narrow" [10].

Goal driven approaches [9], [13] suffer from the inherent limitation of goal orientation. Firstly, goals are fuzzy concepts. [11] points out that goal are "informal and incomplete" and "difficult to precisely define". GORE is subjective, dependent on the requirements engineer view of the real world from where goals are identified [12]. Further, the process of goal reduction is unguided.

III. **Mixed Driven Approaches** In purely demand driven techniques, the information needed for decision making may not necessarily be available in existing data sources, whereas in purely supply driven techniques decision making may require information outside of that available in existing data sources. This led to the development of **mixed driven techniques** where the needed information was identified and the available data was determined.

According to the approach of [6], there is a change of perspective required that views nodes of a goal hierarchy as goals in the first perspective and as decisional alternatives later. This treats all alternatives uniformly and deals with 'what is to be achieved'. In the approach of [14], there is little guidance on what questions to ask even though the metrics determined are critically dependent on these questions.

We notice the following drawbacks currently facing DWRE:

1. Lack of DW support for upstream decision-making in an organization:
   According to [15], the primary concern of data warehouse technology is to provide support to decision makers for managing business processes better. Thus, the focus is on "what to do next" type of decisions that are operational in nature. Information support for operational decision-making is provided to all levels in an organization.

   In [16], there is compartmentalization of operational, business analytics and content analytics in separate modules. The authors recognize multiple levels of

decision makers for long term goals. In terms of decision making, the nature continues to be operational.

OMG, in its Business Motivation Model [17] conceptualizes a business in terms of policies and directives that govern their enforcement. This suggests yet another classification of decisions that is based on the nature of the task to be carried out, namely, policy formulation, determination of policy enforcement rules, operational decisions. Notice that the first two of these are **upstream** to the third and not supported by DW technology. Thus, there is a need to develop specific techniques for these as well.

2. Limited understanding of the Decision-Information Link :

Decisional and Information perspectives have been introduced by [6] [14] respectively. However, we find that

(a) The relationship between the notions of decision and information is not fully explored. Thus, the decision-information association is left un-articulated and remains implicit. This inhibits a full investigation into what information is needed for which decision and vice-versa.

(b) DWRE does not take into account the structure of a decision and the semantic notions underlying decisions. The former means that it is not possible to adopt model driven requirements engineering leading to relatively poor guidance in the elicitation task. The latter implies that the conceptual basis for adopting the notion of a decision itself remains weak.

(c) Information models are assumed to be multi-dimensional in nature. This leads to an emphasis on determining facts and dimensions at the expense of determining information properties like required aggregations and historical information needs. As for decisions, this implies that only partial guidance can be provided in the information elicitation task.

There is a need to explicitly model the decision-information relationship and treat both decision and information as first class concepts of DWRE.

3. Limited techniques specific to Information Elicitation:

DWRE techniques are highly oriented towards arriving at information in the form of Facts, Dimensions and Measures. This is either done directly without analyzing information and or without sufficiently exploring information before structuring it. Techniques like [6] belong to the former class and techniques like [3], [13, 14], [18] to the latter. Even though some investigation of information was done with Information scenarios of [13] there is no guidance provided for developing these scenarios.

While arriving at MD structures is essential, it is equally important to elicit, examine and analyze information that is unstructured and also to elicit information in a guided manner.

To sum up, there is *need to treat decision and information as first class concepts of RE models, develop decision and information models for conceptual clarity and*

effective guidance, and to lay emphasis on eliciting early, unstructured information before arriving at multi-dimensional structures.

With three inter-related DW systems, policy formulation, policy enforcement, and operational, there may be common data across them. Therefore, need for integrating these arises. Existing approaches do star schema/data mart integration by identifying conformed dimensions. However, this implies long lead times due to first arriving at the star schemas and then integrating them. Requirements may change during this period and the integrated system may be out of step with desired one.

We can now define the following research questions:

1. What are the different kinds of decisions, the applications from which they originate and their inter-relationship?
2. Can we define information elicitation techniques that are neutral to the types of decisions?
3. Having determined information relevant to decisions of each application, do we keep separate Data Warehouses application-wise or do we maintain one integrated form?

2. Solution Approach

In order to answer the research questions, we propose a solution divided into the following steps:

2.1. Defining upstream and operational decisions and the decision continuum

We start by establishing the 'Decision Environment' and derive a typology of decisions. We define two broad categories of decisions: Imperative decisions and Managerial decisions. Managerial decision-making is upstream and is of two kinds. One kind of Managerial decisions' is those that deal with formulation of norms and standards that are to be followed in organizations. These decisions are Policy Decisions. The other kind are concerned with the enforcement of given policies. **The decision problem here is that of defining an appropriate set of rules that the organization will follow during its operations.** These decisions are Policy Enforcement Rule (PER) Decisions**.** Imperative decisions are derived from policy enforcement rules and consist of operational actions. **The imperative decision making problem is that of selecting the most appropriate action in a given situation such that it also does not violate policy enforcement rules.**

Imperative decisions can be taken once Policy enforcement rule decisions are taken. Policy enforcement rule decisions are taken once policy decisions have been taken. Thus, there is a continuum between policy, policy enforcement rule decisions and operational decisions.

Policy formulation is done in a number of contexts with varied stakeholders [19, 20], shows dependence on related policies [21] and consensus building [20]. Since policy formulation is a many-facetted and complex task, we have left it for a separate

investigation. Therefore, in this thesis, we assume a policy representation system and consider DWRE for PER decisions and operational decisions only.

2.2. Developing a Generic Platform

Our RE process is rooted in Decision Requirement which we model as a tuple <decision, information>. Decision Requirement implies that RE process has two steps, first to determine the choice set of decisions and then to elicit information to choose one from the choice set.

This and our treatment of decision and information as first class concepts leads us a two level generic platform. The bottom of Fig 1 shows the generic platform having generic models of Decision Requirement, decision and information. Information elicitation techniques for decisions are defined in this layer. The higher layer is the source of decisions for this lower layer. Decisions are obtained from PER formulation and operational decisions. This means that the lower layer is neutral to the source of decisions and can be used for any kind of decisions. It is generic.

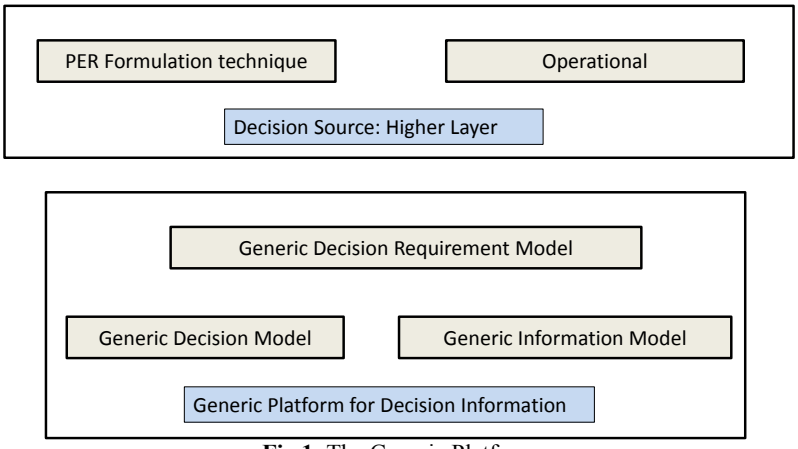

Fig 1: The Generic Platform

For the bottom layer, we propose three information elicitation techniques namely, CSFI, ENDSI and MEANSI as a set of generic techniques for eliciting information for the DW to-be. The elicited information has to be converted to multi-dimensional form and conforms to a *generic information model* developed in the Generic Platform.

For the higher layer, we propose to develop RE approaches for our two sources of decisions. Once this is done, an integration process has to be developed to arrive at the integrated DW.

2.3. Building an integrated DW

Our process begins with first developing Policy Enforcement Rule DW and operational DW. For the former, we propose a PER life cycle thus address the issue of providing support for strategic decision making. For the latter we propose the operational life cycle. Now the remaining question is to integrate the two DWs into one comprehensive DW for the entire organization. .This makes the integrated DW a single enterprise resource that supports ALL forms of decision making. The integration is done using our Vertical integration technique.

Let us consider the RE process in each individual life cycle.

PER Life Cycle

PER life cycle creates a DW for PER decision making, DWper, with a rule base and information in its own operational early information base. The input to the PER life cycle is operational policies. The PER life cycle has two parts; one is to formulate rules and the other to elicit relevant information. We assume that organizational policies are represented in our extended first order logic. Thereafter, we propose guidelines to arrive at PERs. PERs are of the form "WHEN triggering action IF condition THEN correcting action". In general there can be more than one correcting action for a given condition and the decision maker must formulate the appropriate set of correcting actions. In order words, the decision maker works with the choice set {select action, modify action, reject action}. Choosing one of these constitutes the decision problem.

We now need to elicit information that the decision maker will refer to in the decision making task of choosing the appropriate correcting action from the choice set. The generic early information elicitation approaches namely ENDSI, MEANSI and CSFI, defined in the platform of Fig 1 are used. This constitutes early information, $EI_{per,}$

EI_{per} can be:

1. a source of information for the vertical integration life cycle
2. converted to late, structured information

For the latter we have proposed guidelines to convert early information into ER diagram. Subsequently, ER can be converted to star schema by applying existing algorithms of [4]. DWper thus obtained can be used by PER level decision maker.

We applied our process to AYUSH policies consisting of 151 policies. The RE methodology was implemented in a tool called ELISPE and it was used to elicit the required information.

Operational Life Cycle

Operational life cycle creates a DW for operational decision making, DWop. Actions are first extracted from PERs and each action, a, of the PER layer is treated as a decision, d, to be taken at the operational level. This yields the initial set of decisions and is the input to the operational life cycle.

Applying the generic decision model mentioned in Fig 1, each decision is subjected to AND/OR decomposition and generalization-specialization process and leaf nodes are determined. The next task is to elicit information for every decision using CSFI, ENDSI, and MEANSI. The elicited information is $EI_{op.}$

Again, EI_{op}, can be either be converted to ER diagram, for which we propose guidelines, or can be used as a source of early information for vertical integration. In the case of the former, once the ER diagram is converted to star schema we obtain a stand-alone DWop.

Vertical Integration Life Cycle

We found that there are two problems with keeping separate Data Warehouses. These arise if there is common data in them.

- Difference in refresh cycles between DWper and DWop cause common data to have differen values in the two DWs. Thus, rule formulators and operational decision makers end up taking decisions on different data in this temporal window. The larger this window, the longer this inconsistency exists.
- Loss of business control occurs when data of an operational DW calls for decision makers of the policy enforcement DW to take decisions, but the decisions are not taken because daa in the latter do not suggest this need.

Thus, integration is required to maintain compatibility between PER and operational level.

We show that there are in fact two forms of integration that can exist, horizontal and vertical. While the former integrates data marts at the same level of decision-making, the latter integrates data marts across PER and operational levels. For vertically integrating DWper and DWop, we propose a 'build by integrating' approach that works pair-wise pair-wise. When a new data mart is to be built, its requirements specification is integrated with an existing one. The integrated requirements specification then goes through the development cycle. Thus, the point of integration is moved upstream into the requirements stage. In other words, early information is integrated. The advantages of integrating upstream and in a pair-wise fashion are:

- downstream development effort is minimized.
- it never allows un-integrated data marts to be built. Thus, pre-empting our two problems.
- A complete logical DW is available for decision making.

Integration is done as a four-step process, Metadata reading, Correspondence Drafting, Information Mapping and Conflict Resolving. The integrated early information obtained is then converted into ER schema and finally into star schema.

Through vertical integration, an integrated DW is obtained that can be used for both forms of decision making.

3 Contribution of research work

A summary of the contributions made is as follows:

1. Addressing full decisional making continuum: DW support has been extended from providing just operational support to providing policy enforcement and operational support.

2. Elicited Information can be traced back to members of the choice set thereby facilitating decision making. In the case of PERs the choice set is {select A, modify A, delete A} where A is an action and {select A1, select A2, select A3} for operationa decisions. For each alternative, information is elicited, thus relating information to a particular member of the choice set.

3. Discovery of early information: Information is obtained in a two-step process, early information elicitation step, using ENDSI, MEANSI and CSFI and late information elicitation step where early information is converted into ER diagram which is subsequently converted to a star schema.

4. Development of a requirements integration approach that pre-empts the problems of inconsistency in decision-making and loss of business control.

4 Conclusion and future work

This work addresses the issue of providing support to different kinds of decisional needs in a unified, enterprise wide DW system. A two level generic platform is proposed with generic models at the bottom level and decision sources at the higher level.

We develop RE process to arrive at separate DWs for the PER and operational decisions respectively. The two DWs are integrated upstream in the requirements engineering phase. The integrated requirements specification is then converted into multi-dimensional form.

Future work includes developing a policy life cycle for eliciting information for policy decisions and integrating it with 'lower' levels of the decision continuum.

5 References

1. Alshboul, R. (2012). Data Warehouse Explorative Study. *Applied Mathematical Sciences*, *6*(61), 3015-3024
2. Hayen R., Rutashobya C., Vetter D., (2007), An Investigation Of The Factors Affecting Data Warehousing Success, Issues In Information Systems, Volume VIII, No. 2, 547-553, 2007
3. Golfarelli M., Maio D., Rizzi S. (1998, January). Conceptual Design of Data Warehouses from E/R schemes. In System Sciences, 1998., Proceedings of the Thirty-First Hawaii International Conference on (Vol. 7, pp. 334-343). IEEE.

4. Moody L.D., and Kortink M.A.R. (2000), From Enterprise Models to Dimensional Models: A Methodology for Data Warehouses and Data Mart Design, Proc. of the Intl Workshop on Design and Management of Data Warehouses, Stockholm, Sweden, (pp. 5.1-5.12)
5. Kimball, R. (1996): The Data Warehouse Toolkit, New York: J. Wiley & Sons.
6. Giorgini, P., Rizzi, S., Garzetti, M. (2005). Goal-oriented requirement analysis for data warehouse design. In Proceedings of the 8th ACM international workshop on Data warehousing and OLAP (pp. 47-56). ACM.
7. Winter, R., and Strauch, B. (2003, January). A method for demand-driven information requirements analysis in data warehousing projects. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on* (pp. 9-pp). IEEE.
8. Bruckner, R., List, B., &Scheifer, J. (2001). Developing requirements for data warehouse systems with use cases. *AMCIS 2001 Proceedings*, 66.
9. Boehnlein, M., and Ulbrich vom Ende, A. (2000). Business Process Oriented Development of Data Warehouse Structures. In Proceedings of Data Warehousing 2000 (pp. 3- 21). PhysicaVerlag HD
10. List, B., Bruckner, R. M., Machaczek, K., & Schiefer, J. (2002, January). A comparison of data warehouse development methodologies case study of the process warehouse. In *Database and Expert Systems Applications* (pp. 203-215). Springer Berlin Heidelberg.
11. Horkoff, J., and Yu, E. (2010). Interactive analysis of agent-goal models in enterprise modeling. *International Journal of Information System Modeling and Design (IJISMD)*, *1*(4), 1-23.
12. Haumer, P., Pohl, K., and Weidenhaupt, K. (1998). Requirements elicitation and validation with real world scenes. *Software Engineering, IEEE Transactions on*,*24*(12), 1036-1054.
13. Prakash N., and Gosain A. (2008). An Approach to Engineering the Requirements of Data Warehouses. Requirements Engineering Journal, Springer, 13 (1), 49-72
14. Bonifati A.. Cattaneo F., CeriS., A. Fuggetta, and S. Paraboschi (2001). Designing Data Marts for Data Warehouses. ACM Trans. Software. Eng. Methodology, 10(4). (pp. 452–483).
15. Adamson (2010) The complete reference: Star Schema. Tata McGraw-Hill
16. Imhoff, C., & White, C. (2008). Full Circle: Decision Intelligence (DSS 2.0). *B-Eye-Network, Published: August*, *27*.
17. BRG, 2010 BRG, Business Rules Group (2010), The Business Motivation Model: Business governance in a volatile world, Release 1.4, July 2010
18. Prakash, N., and Bhardwaj, H. (2014). Functionality for Business Indicators in Data Warehouse Requirements Engineering. In *Advances in Conceptual Modeling* (pp. 39-48). Springer International Publishing.
19. Lindbloom C.E. (1993), Woodhouse E.J., 3rd edition, Prentice Hall, 1993
20. Ritchie J.R.B. (1988), Consensus policy formulation in tourism: Measuring resident views via survey research, Tourism Management, 9,3, 199-212, 1988
21. Park Y. T. (2000), National systems of Advanced Manufacturing Technology (AMT): hierarchical classification scheme and policy formulation process, Technovation, 20,3, 151-159, 2000