Data Warehouse Requirements Engineering

Dr. Vinay Kumar[1], Reema Thareja[2*]

ABSTRACT

Data warehousing projects are failing at an alarming rate because of poorly defined objectives, incorrect assumptions, inadequate planning and a wide gap between user's expectations and the final product. This ultimately results in wastage of time, money and effort spent on the project. Most of these factors are a result of inadequate requirements engineering activities. Since data warehouse is a large database with an absolutely different scope of usage, the information requirements analysis for such systems is entirely different from that of traditional databases. This became evident while implementing National Population Register (NPR) project. Therefore, we have proposed a requirements engineering framework which supports a combination of incremental and iterative development of data warehouse project. The framework will help the development team to develop a set of complete and verifiable requirements.

KEYWORDS

Requirements Management, Requirements Engineering, Verifiable Requirements, Data Warehouse Framework, NPR.

INTRODUCTION

Data Warehousing has emerged as a powerful technology for integrating distributed data into a comprehensive analytical tool to enable decision-making. Bill Inman has defined data warehouse as *a collection of consistent, subject-oriented, integrated, time-variant, non-volatile data in support of management's decisions.* A data warehouse is not just a collection of data, but it entails in it a comprehensive process of extracting relevant data from operational source, transforming and aggregating the extracted data and allowing users to access it for making strategic decisions [19].

A data warehouse stores historical data from multiple sources including sources outside the organization [11]. The data is cleansed and validated before storing in standard format that may be used to analyze the data and business processes in a comprehensive and flexible way. For this purpose data warehouse is modeled multidimensional to support users to view the data from different requirement perspectives and at different levels of aggregation.

Many data warehouse systems fail because they are developed based on an incomplete and / or inconsistent set of information requirements [7]. Moreover, user‟s requirements keep changing with new product launches, new promotional strategies, new pricing schemes, new customers, etc. For this purpose, the paper proposes a framework of activities that will be performed for homogenizing and validating information requirements for data warehouse systems using the Goal Structured Requirement Engineering and Traceability Model [13]. Requirements form the backbone of the model and in this paper we discuss requirement engineering for successful implementation of the goal model in data warehouse systems.

Requirements engineering is complex task because the users who understand the problem to be solved does not know how to develop the system and the developers who are well versed with the tools and techniques to develop the system finds it difficult to understand what problems have to be solved. It is therefore important to document the user‟s requirements clearly and concisely [4].

Requirements are statements that describe what the system must do, how it must do, what properties it must exhibit, and what constraints it must satisfy. Accordingly, requirements engineering is a process that emphasizes the use of a systematic technique to ensure the completeness, consistency, and relevance of the system requirements [17][18]. Activities that are part of requirements engineering are as follows:

- **Requirements elicitation** for identifying user's needs and system's constraints
- **Requirements analysis** for refining the user's needs and constraints.
- **Requirements specification** for documenting user's needs and constraints
- **Requirements Validation** to ensure that requirements are complete, correct and consistent.
- **Requirements management** for scheduling, coordinating, and documenting the above activities

1. Guide, Department of IT, VIPS, Guru Gobind Singh Indraprastha University, Delhi. Email: vinay5861@gmail.com

2. Research Scholar, Department of Computer Science, SPM College, Delhi University, Delhi. Email: reema_thareja@yahoo.com

This journal is cited as : **JIMS 8i-Int'l J. of Inf. Comm. & Computing Technology(IJICCT)**

The other major problem in requirements engineering is that there is a diverse group of users who have their own set of needs and expectations from the system. Many times these requirements may be conflicting. All these decisions need to be recorded to know the reasons in case the system fails. During development, dynamic requirements are managed through defined requirements-change-management process.

While preparing the implementation plan for National Population Register (NPR) project, it has been realized that the existing approaches for requirement engineering for data warehouse are not sufficient. Considering many aspects of good governance, the Government of India took decision to create the National Population Register (NPR) which will be a comprehensive repository of information of all usual residents of the country. Biometric data like photographs, finger-print and iris data are also be collected for residents above 15 years of age. For children between 5 and 15 years of age, only iris data is to be captured. The NPR data is expected to serve as the most comprehensive identity data of all residents of the country and it is to be linked with various social welfare and developmental plan in the country. It is therefore essential to do the requirement engineering in such a way that it will cater the need of diverse group of user at different level of administrative hierarchy.

The paper is divided into five sections. Section 2 describes existing approach for requirement engineering in data warehouse. Proposed approach is discussed in Section 3. The implementation of proposed approach is described in Section 4 using a real life project on National Population Register. The paper is finally concluded in Section 5.

EXISTING APPROACH FOR DWRE

The user‟s requirements determines applications, sources of data, the structure of fact and dimension tables, quality of information, frequency of data updates, etc. The key issue in data warehouse development is to match information requirements of user who will be using data warehouse in future with present available information supply. There are only a few approaches that deal with such issues specifically [22]. Two most widely used approaches are demand driven approach and supply driven approach [3]. The two approaches differ on whether the matching process is guided by information demand or information supply. Another basic difference lies in the implementation of ETL process.

You need to load your data warehouse regularly so that it can serve its purpose of facilitating business analysis. Similar is the case in NPR wherein data is required to be used for verification at various levels and in decision making process. To do this, data from one or more operational systems needs to be extracted and copied into the data warehouse. The challenge in data warehouse environments is to integrate, rearrange and consolidate large volumes of data over many systems, thereby providing a new unified information base for business intelligence. [19]

Table 1: Comparison of Demand Driven and Supply Driven Approaches

Demand Driven	Supply Driven
1. Information requirements are collected from the data warehouse users[20][21] 2. But end users are unable to specify their requirements precisely and accurately as they do not have sufficient knowledge of all available information sources available in the organization. 3. Complicated ETL process 4. Gives importance to user‟s requirements	1. Analysis of source systems is done to reengineer their logical data schemas. The end users are then be asked to specify their requirements from such a consolidated data schema 2. Simplifies ETL process 3. Gives less importance to user‟s requirements 4. Based on what is available is data repository

The process of extracting data from source systems and bringing it into the data warehouse is commonly called ETL, which stands for extraction, transformation, and loading. ETL refers to a broad process and not simply three well defined steps. The acronym ETL is perhaps too simplistic, because it omits the transportation phase and implies that each of the other phases of the process is distinct. Nevertheless, the entire process is known as ETL.

The methodology and tasks of ETL are not necessarily unique to data warehouse environments: a wide variety of proprietary applications and database systems are the IT backbone of any enterprise. Data has to be shared between applications or systems, trying to integrate them, giving at least two applications the same picture of the world. This data sharing was mostly addressed by mechanisms similar to what we now call ETL.

A variant of demand driven approach derives information requirements by analyzing business processes and transforming the relevant data structures

of business processes into data structures of the data warehouse [3]. A brief comparison of demand driven approach and supply driven approach is presented in the Table I.

PROPOSED APPROACH

The approach begins with identification of users, their dominant application type followed by collection, documentation and validation of identified requirements. Once the requirements are reviewed and approved, conceptual model is developed for the data warehouse. Thereafter, requirements management activities are performed for establishing and maintaining the integrity and accuracy of the requirements as the data warehouse project evolves. Therefore requirements engineering can be thought of comprising the following steps:

- Requirements Identification
- Generation of Conceptual Model
- Requirements Management

REQUIREMENTS IDENTIFICATION

Requirements identification process begins with identification of actual users of the data warehouse. Such users specify the tasks they perform for achieving objectives established by the business requirements. This leads to identify what different types of applications they are or will be using for their analytical work and in decision making process. More often the analysts use OLAP tools for flexible information analysis, while upper management prefer for analyzing business trends. The identification of dominant application is an issue of concern as it has implications on which data models are used, which source systems are relevant, and which decision processes have to be considered.

GENERATION OF A CONCEPTUAL DATA MODEL

Regularly used reports are analyzed to match available information with the requirement of information. Data sources are analyzed to ensure quality and usability of data in the data warehouse. For matching information supply and information demand, an aggregate information map is made which specifies a data schema of information subjects on an aggregate level. The information map specifies the source of data, users who use that data, concepts related with that data, terms that are homonyms or synonyms, etc. on an aggregate level. It then helps in determining unsatisfied information requirements. Unsatisfied information requirements are prioritized on factors like: data granularity demanded, refresh frequency, data privacy and security, implementation time, overall costs of the project etc. Data sources are analyzed to understand the existing

The requirements elicitation is iterative process. First iteration provides a set of initial and prioritized high-level requirements that describes the objectives, scope, opportunities, and vision for the data warehouse. Then end user‟s analytical needs are refined from high level concepts to low level details. The second iteration considers the perceived implications of the requirements on future data warehouse usage & requirements. This leads to consolidation of initial requirements by making adjustment with new requirements. In the next iteration, all the requirements are reanalyzed to determine relevance of these requirements from perspective of end-user. Relevance implies whether these requirements add value to the system or are unnecessary. In the last iteration, it is ensured that all the requirements are considered and it is approved by the end user‟s perspective. These iterations are repeated a couple of times before the final requirements document is approved and signed. Requirements is collected from the users through interviews, workshops, prototyping, use cases, Goal Decision Information model [15], easy REMOTEDWH [16] and DWARF [14].

The requirements that have been elicited are then clearly documented specifying functional as well as non-functional requirements of the users. A complete and useful requirements document includes information about the requirements management plan, terminology related to problem domains, project scope, use case specification, non functional requirements and business rules. Documentation is followed by requirements validation. It is done to identify and correct any pitfalls in requirements identification. In review sessions, documented requirements are shared with all the involved parties for their comments and feedbacks followed by development of an OLAP interface prototype for validation of the identified requirements. transformation rules. Maintaining uniformity in data semantics is done to control and manage any inconsistencies and discrepancies in data to meet unsatisfied requirements by the existing data.

Once uniformity in data semantics is finalized, information requirements are re-assessed to form a final priority sequence of detailed, homogenized information that should be provided by the data warehouse system. These requirements are categorized as “must have,” “want,” and “wish to have” to provide flexibility in development process to enable the system to meet technical limitations, if any, or cost in future. In order to complete the schema design, it is required to identify facts and dimensions and to determine the extent to which a metric is *additive* along dimensional hierarchies. Additionally, dimensions that help to analyze facts along different viewpoints are identified. A schematic diagram of the arrangements of facts and dimensions tables is depicted in the Figure 1.

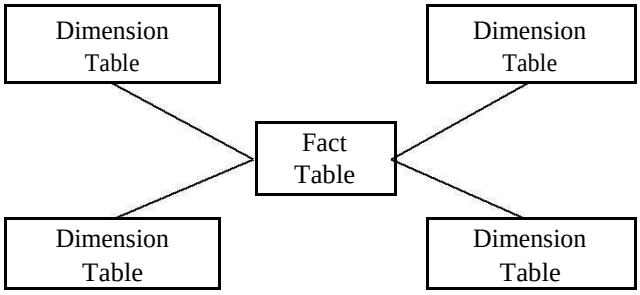

Figure 1: Schematic diagram for fact and dimension table relationship

Every dimension table has a direct relationship with the fact table in the middle. The arrangement allows every dimension table with its attributes to have an equal chance of participating in a query to analyze the attributes in the fact table. Any change suggested by the user in the data model triggers backtrack to a previous step of analyzing information demands. A schematic diagram for entire activities to be carried out to generate conceptual data model is shown in the Figure 2.

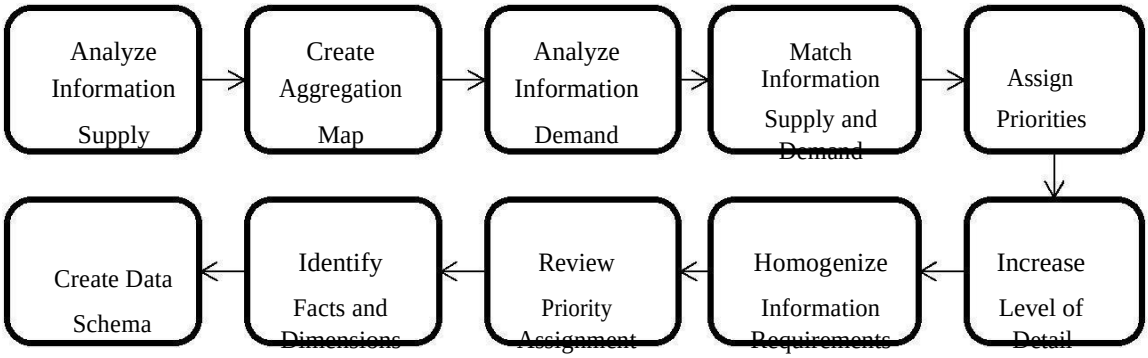

Figure 2: Schematic diagram for creating conceptual data model in the proposed approach. **REQUIREMENTS MANAGEMENT** then corresponding field are shown and rest are hidden.

Each change in requirements results in an amendment to the requirements specification or an entirely new version of the requirements document. It is therefore decided to maintain revision history in the requirements documents according to the standard format by including change made, date of change, individual who proposed the change, and justification for the change. It facilitates users and development team to track the status of each requirement throughout the development process.

For improved requirements traceability, a requirements form is developed. For example, if use cases are used to define requirements then the form has different fields, if the requirements are defined using prototyping then use case fields are hidden and relevant fields are displayed. Similarly if the requirements are defined through interview or questionnaire or any other the technique

Such dynamic forms support a different requirements structure; ensures consistency for the requirements, and allow automatic checking of the completeness of each requirement. The process ultimately leads to a well designed and organized metadata that stores definition of general terms, which are used in the requirements specification.

Under this phase of requirement engineering, impact of any requirement change on the overall data warehouse system is assessed using traceability links[12] that would help to follow the life of a requirement both in the forward as well as the backward direction. This is illustrated in the Figure 3. Forward link from users to requirement represents „change in requirement proposed by user and its consequent impact on data warehouse. Backward link represents traceability of any change carried out in the data warehouse. The traceability links avoids unnecessary design, code and

1. Guide, Department of IT, VIPS, Guru Gobind Singh Indraprastha University, Delhi. Email: vinay5861@gmail.com

2. Research Scholar, Department of Computer Science, SPM College, Delhi University, Delhi. Email: reema_thareja@yahoo.com

documentation of elements that are not related to user-specified requirements[10]. To follow the change control process, a change control board is formed that considers all proposed changes at regular intervals and decides which proposed changes to approve.

Figure 3: Forward and backward traceability links from data warehouse users to the data warehouse

4. DWRE for NPR

NPR Database creation exercise has been undertaken by the Registrar General of India (RGI) for the Country as a whole. After digital conversion of the demographic data, the same will be integrated with the corresponding biometric data from Unique Identity Authority of India (UADAI). The consolidated demographic and biometric data is supposed to form baseline of this NPR project. Each record is is of 5MB (demographic and biometric data together). Since population data is dynamic in nature, provision of scalability is to be taken care of. The NPR is initiated with the purpose of monitoring security due to infiltration and for monitoring and controlling the proper utilization of funds under various developmental and social welfare schemes. Providing information about the usual residents as and when needed for authentication and validation is one of the primary goals of creating this database on priority. The information related to resident may be required by the concerned authorities. We describe here two of its applications and requirement of information from NPR data warehouse: authentication of residents with NPR Card by security agency and authentication of residents with NPR for various services integration.

AUTHENTICATION OF RESIDENTS WITH NPR CARD BY SECURITY AGENCY

As a part of the NPR project, every registered resident is to get a NPR smartcard equipped with a chip containing sufficient information to identify the resident uniquely. The card helps in further verifying with biometric data of the resident. When a security agency wants to verify the identity of a resident, he asks the resident for the NPR Smart card. If the person is a usual resident of the area and has registered himself for a NPR card, he will have a valid NPR card, assuming the resident is above 18 years of age.

When a security agency is required to authenticate a person, the person may or may not have a NPR card. If NPR card is available, the security personnel take the NPR card and put the card into the smartcard reader. If the card is genuine then the card reader will be able to read the contents of the chip and display the chip contents. If the security personnel want to establish the identity of the resident using biometric, he can use the device to compare the finger-print minutiae stored in the card chip with the resident finger-print besides the visual comparison of the resident with the photograph on the card.

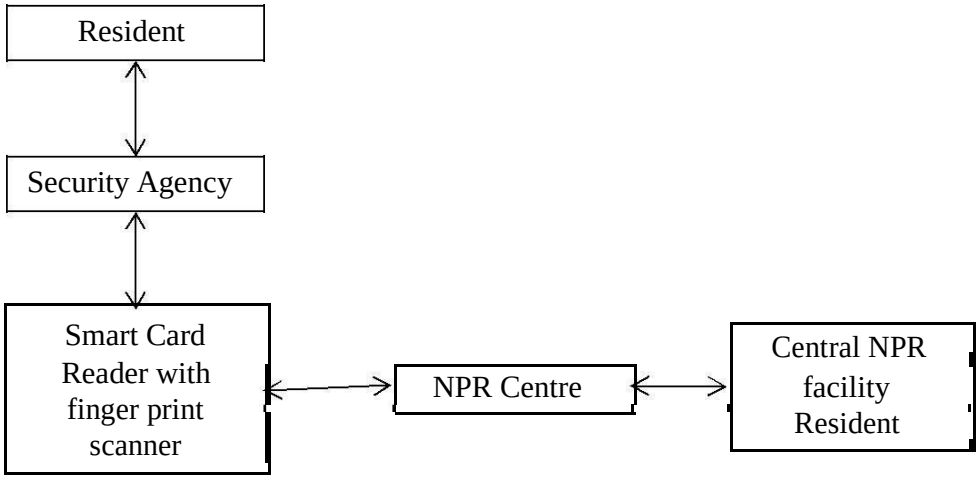

Figure 4: Authentication process of residents with NPR smart card

Thus it is important to identify the requirement of information in such a way that the NPR card helps security agency to verify
that the card is genuine
name and address on the spot
biometric details on the spot

In case the resident does not have NPR smart card, the person is detained by the security agency for verification. The police take the finger-prints of the person using a scanning device and sent the same as images to the nearest NPR center. At the NPR center, a finger-print minutia matching is done and the result (yes/no) is sent back to the agency. Additional information like resident name; address etc and matching information may be found by the NPR center to help further investigation. The entire process of NPR verification is depicted in the Figure 4.

In the case of authentication of residents with NPR for various services integration, the local government agency verifies the authenticity of claimant of various welfare schemes run by government. It helps in managing the project in such a way that the benefits of the scheme reach to the target group and the pilferage of the fund is minimized to zero level. Any beneficiary will have to first show its UID number embedded smartcard issued to the individual. If the card is not available, the person is to get the card issued before availing the benefits of the scheme.

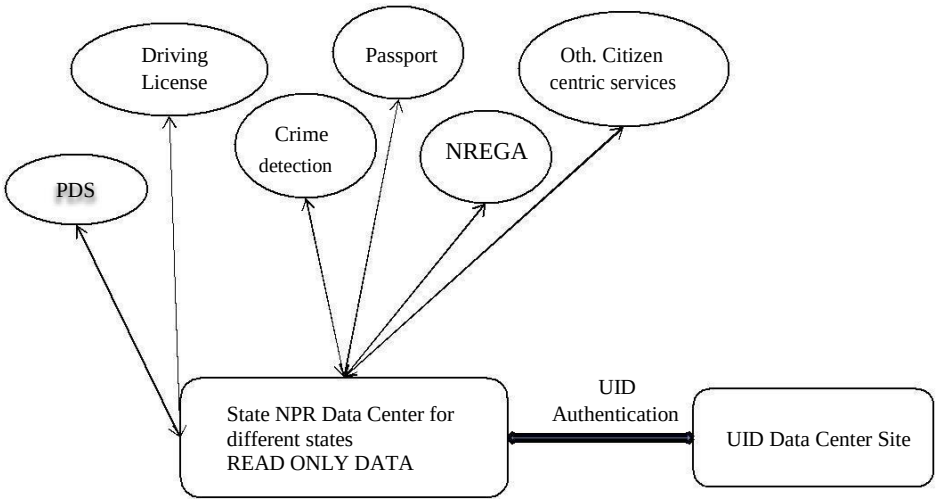

Figure 5: Authentication of residents with NPR for various services integration

The logical architecture of the NPR system is depicted in the figure 6. This is designed taking into consideration possible requirements in future. Once the core NPR database is in place, we essentially need to define the various layers for the NPR application system which will have to be developed to realize the envisaged objectives of the NPR Data warehouse updation and maintenance.

The NPR web access interface acts as a common interface for all stakeholders, and it would provide interface for various stakeholders of the system and robust role based access controls for access to appropriate information and/or applications. Its **p**resentation layer acts as an interface with users which will include static and dynamic content based on templates for various stakeholders. Every user request has to be routed through front controller component, which provides the access for various NPR functionality using view component of MVC design pattern (Model, View and Controller and Front Controller).

1. Guide, Department of IT, VIPS, Guru Gobind Singh Indraprastha University, Delhi. Email: vinay5861@gmail.com

2. Research Scholar, Department of Computer Science, SPM College, Delhi University, Delhi. Email: reema_thareja@yahoo.com

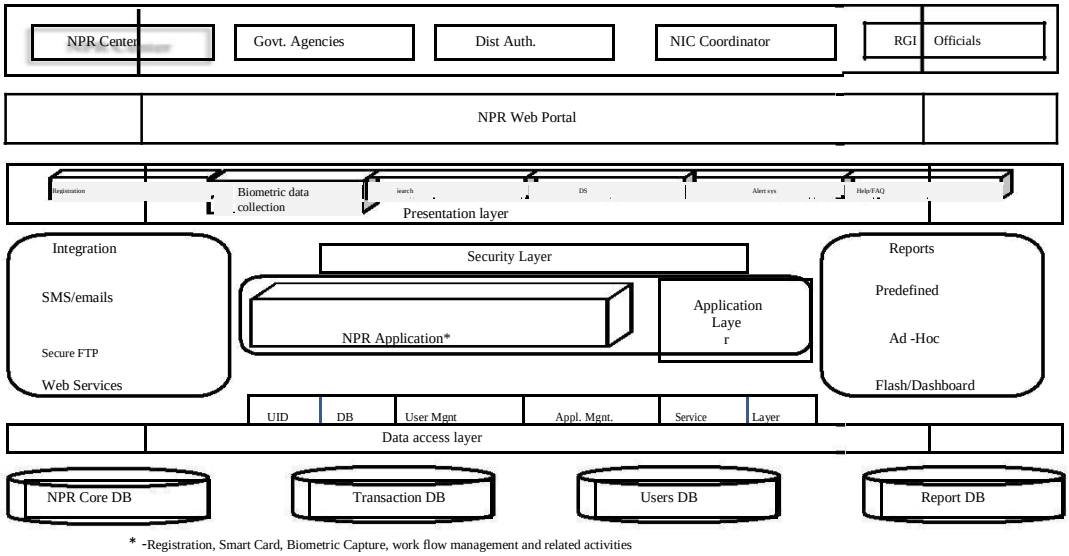

Figure 6: Logical Architecture of the NPR System

Reporting layer facilitates the users in generation of standard pre-defined reports, queries and refreshing the dashboard

data. Information requirements from data warehouse need to be made available for

Dashboards to provide rich, interactive user interface to make the presentation of data intuitive, relevant, and easy to understand. And also to provide users with information filtered for their identity, function, predefined security rules, etc.

System should facilitate a set of standard pre-defined reports.

System should facilitate users in generation of ad-hoc reports to meet their requirements.

Security layer ensures access by authenticating system users against a securely stored set of credentials from the database. NPR applications like Registration, De-duplication, Biometric Capturing, Biometric Identification, NIN Generation etc. to be part of application layer. The applications will be accessible from the internet through published URL to authorized users over a web-server.

Service Layer of the NPR system facilitates routing and transformation, standards based interaction, redefined ready-for-use service types. Integration layer will ensure standards based protocols, web service, APIs, generic connectors etc for meeting current and future NPR system requirements. Data Access layer will provide simplified and secured access to the core NPR database, the "temporary" transactional database and the user database.

5. CONCLUSION

The requirements elicitation and management process for traditional databases is fundamentally different from that for data warehouse systems. The process for data warehouse systems is a large and complex. The paper presents a methodological support for information requirements analysis, elicitation and management. The methodology is based on demand-oriented approach that supports an iterative, priority oriented approach for requirements engineering. As primary deliverables, a star schema based requirements specification document is created. The star schema is a conceptual model of data warehouse. The proposed framework encourages active role of end users in data warehouse development activities. End users are involved in interviews, scenario discussions, filling questionnaires, giving feedback on prototypes, validating requirements and evaluating the schema thus formed.

Moreover, requirements management activities help in improving user"s perception of requirements changing consequences, and the effect of frequently changing needs to the data warehouse schedule. The proposed framework acknowledges the gap in current practices based on the established best project management practices and tries to fill it in a logical manner. The framework of activities is similar to those taken during most data warehouse requirements collection and definition efforts. However, our framework also considers the *Requirements Elicitation* process and *Requirements Management* process which is often ignored for data warehouse systems.

ACKNOWLEDGEMENT

We are grateful to all those who have been constantly encouraging us to go for such application oriented study work besides the regular work which we are doing at our respective departments.

REFERENCES

[1] Ballard, C., Herreman, D., Schau, D., Bell, R., Kim, E., Valencic, A., Data modeling techniques for data warehousing, redbooks.ibm.com, 1998.

[2] Birk, A.; Heller, G.; John, I.; Joos, S.; Muller, K.; Schmid, K.; & von der Massen, T.*Report of the GI Work Group "Requirements Engineering for Product Lines"* (IESE-Report No. 121.03/E, V1.0). Kaiserslautern, Germany: Fraunhofer Institut Experimentelles Software Engineering, 2003.

[3] Boehnlein M., Vom Ende U., 2000, "A Business Process Oriented Development of Data Warehouse Structures", Proceedings of Data Warehousing, Physica Verlag.

[4] Brooks, F. "No Silver Bullet: Essence and Accidents of Software Engineering."*Computer 20*, 4 (April 1987): 10-19.

[5] Bruckner R.M., List B., Schiefer J., 2001,"Developing requirements for data warehouse systems with use cases", proceedings of 7th Americas Conference on information systems.

[6] Davis, A. M. *Software Requirements: Analysis and Specification*. Englewood Cliffs, NJ: Prentice-Hall, 1990.

[7] Dorfman, M. & Thayer, R. H. *Software Requirements Engineering*. Los Alamitos, CA: IEEE Computer Society Press, 1997

[8] Faulk, S. R. "Software Requirements: A Tutorial," *Software Requirements Engineering*. Los Alamitos, CA: IEEE Computer Society Press, 1997, 128-149. .

[9] Faulk, S.; Harmon, R.; & Raffo, D. "Value-Based Software Engineering (VBSE): A Value-DrivenApproachtoProduct-Line Engineering", *Software Product Lines: Proceedings of the First Software Product Line Conference (SPLC1)*. Denver, Colorado, August 28-31, 2000. 205-224..

[10] Gardner, S. „Building the Data Warehouse," Communications of the ACM, 1998, vol. 41, no. 9, 52-60.

[11] Inmon, W.H., Building the data warehouse, Wiley, New York, 1996.

[12] Jukic N. , Nicholas J., 2010, "A Framework for Collecting andDefining Requirements for DataWarehousing Projects", Journal of Computing and Information Technology, Vol-18, pg 377-384.

[13] Kumar V., Thareja R, "Goal Structured Requirement Engineering and Traceability Model for Data Warehouses", International Journal of Information Technology and Computer Science, 2013, vol. 12, pg. 78-85.

[14] Paim F.R., Castro J.B., 2003, "DWARF: An Approach for Requirements Definition and Management of Data Warehouse System", 11thIEEE International Requirements Engineering Conference (RE'03), Monterey Bay, California, USA.

[15] Prakash N., Gosain A., 2003, "A Requirements Engineering for Data warehouse Development", Proceedings of CAiSE03 Forum, pg 13-16.

[16] Shiefer, J., List, B., Bruckner, R.M., 2002, "A holistic approach for managing requirements of data warehouse systems", proceedings of 8th Americas Conference on information systems

[17] Sommerville, I. & Sawyer, P. *Requirements Engineering: A Good Practice Guide*. New York, NY: John Wiley & Sons, 1997.

[18] Sommerville, I.; Kotonya, G., *Requirements Engineering: Processes and Techniques*, Horizon Pubs & Distributors Inc., 1998.

[19] Thareja R., Data Warehousing, Oxford University Press, India 2009.

[20] Winter R. and Strauch B., 2003, "A method for demand driven information requirements analysis in data warehousing project", proceedings of the 36thHawaii International Conference on System Sciences, USA.

[21] Winter R., Strauch B., "Information Requirements Engineering for Data Warehouse Systems".*ACM Symposium on Applied Computing (SAC'04) Nicosia*, Cyprus.2004.

[22] Yu, E.S.K., 1997, "Towards Modeling and Reasoning Support for Early-Phase Requirements Engineering", proceedings of IEEE International Symposium on Requirements Engineering ,226-235