ISSN 2278-3091
Volume 9, No.3, May - June 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse13932020.pdf
https://doi.org/10.30534/ijatcse/2020/13932020

Conceptual Design of Data Warehouse using Hybrid Methodology

G. Sekhar Reddy[1], Dr Ch. Suneetha[2]
[1]Research scholar, Acharya Nagarjuna University, Guntur, India, golamari.sekhar@gmail.com.
[2]Associate Professor, Department of Computer Applications, RVR&JC college of Engineering, Guntur, India, suneethachittineni@gmail.com

ABSTRACT

In recent times, data warehousing achieved tremendous attention in various organizations including universities to analyze important aspects of their academic environment. In this paper, we present an automatic design system to integrate the functionality of both the requirement-driven and data-driven approaches. In addition, it is established on the basis of i* framework and Dimensional Fact Model (DFM) which is used to design the actions of actors, and the relationship existing among the agents in DW (Data Warehouse) environment. Furthermore, the proposed system introduces and illustrates an automatic design technique based on a logical programming to perform the integration of different data sources using UML multidimensional schemas that are reconciled with data sources to improvise the conceptual quality.

Key words : Data-driven approach, DFM model, i* framework, Logical programming, Requirement driven approach.

1. INTRODUCTION

Data warehousing environment is greatly differ from all the other operational environments, which are used to design database for various organizations. Uniqueness in DW stands for its integrity, non-volatility and efficiency, which would raise several differences in the modeling and designing schemas of DW. In early 1990's different approaches were used to build data warehouse. Some of them are: bottom-up method, top-down method, hybrid method and federated. With huge storage of information and its relevant historical data, DW tends towards the necessity demanding for a big data space [1]. This in turn leads to the introduction of slowly changing dimension method [2]. For this reason Kimball's in [3] proposed a dimensional modeling technique, which stores the data in Multi-dimensional Store (MD) in the form of facts and dimensions. Furthermore [4] develops a conceptual design for the multidimensional systems, it outlines the user's initial requirements [5], availability of data sources, system behavior and the related issues regarding the database schemas. Still, the design schema in MD model fails to address the required information, so that poor communications has been established between the DW developers and the decision makers [6, 7]. To overcome this failure Goal Oriented Requirement Engineering (GORE) framework was proposed, which uses Unified Modeling language (UML). [8, 9] use conceptual design for the implementation of MD model, where the term conceptual modeling includes the models based on Entity/Relationships. Furthermore logical modeling is a representation of schemas, which includes star schema, snowflake schema, and constellation schema.

Requirement analysis plays a major role in developing an effective and efficient data warehouse project. For analyzing various user necessities, four approaches were commonly used in [10] like data-driven approach (supply-driven), user-driven, goal-driven (demand driven/requirement driven) and mixed driven approach. [11] uses an user focused method to extract the end-user necessities and designing a DW as multi-dimensional task. [12] Automatically validate the user requirements to conciliate with data sources for designing the DW. [13-15] use data-driven approach with DFM model to identify the multi-dimensional concepts from domain ontology, where the viability of data warehouse is guaranteed. However it fails to analyze the user needs and hence there may arise some possibilities for the occurrence of integrity problem, additionally the approach wasn't a traditionally automated method. Therefore [16] focuses on designing a data warehouse automatically from relational schemas. [17] focuses on the requirement-driven approach, it start with processing the business goals which is then followed by building a multidimensional schema. It achieves user necessities but it would not support effectiveness of all data present in the data source.

[18] defines research and implementation issues of hybrid methods for the design of DW. [19] uses a hybrid Model Driven Architecture (MDA) approach to design the conceptual hybrid MD model and also it automatically derives

the logical representation of the MD model of a DW. Yet it fails to deal with the complex multidimensional model structure. [20] uses semi-automatic approach which comprises the process of identifying the facts, dimensions and dimension hierarchies. [21] proposes the data-driven approach, introduced by Inmon in the year 1996, which uses the bottom-up technique. This approach analyzes and remodels the data sources to obtain the multidimensional schema to design a DW. It manually gathers and maps the requirements onto the relational schema. Also it generates the attribute tree for each identified facts semi-automatically. It then identifies the measures, dimensions and the dimension hierarchies, which is carried out by pruning the attribute tree followed by giving raise to the multi-dimensional schema even though the method fail to identify the dimensional concepts automatically. Hence to deal with all these drawbacks faced by the conventional methods, a new approach is used, which incorporates the advantages of both requirement driven and data driven approaches to preserve user needs by defining the hybrid methodologies to design a data warehouse conceptually.

2. HYBRID DESIGN METHODOLOGIES

2.1 Data-Driven Approach

It uses dimensional fact model to construct a conceptual model in semi-automatic manner. Once the user requirements are detected from the end user through several interviews, the data sources (requirements) gathered are integrated as a global schema. During conceptual design the method will generate an attribute tree, having the multi-dimensional elements (fact, measure and dimensions).

Where, fact represents the root, dimensions depicts the child nodes of root and measures represent the leaf node. Finally based on the designer requirements, the tree has been remodeled, to define the multidimensional concepts. At the end of remodeling, the final tree can be viewed in the form of cube. It is then transformed into star or snowflake schemas. Here, it uses snowflake schema as its logical representation, it is a normalized schema and is well suited for constructing data warehouse. Some drawbacks of this approach are as follows:

- Difficulties in understanding the user needs
- Leads to wastage of time because the designer does not focus on essential portion of data source, instead they are forced to focus towards entire data source to detect the multi-dimensional elements.

2.2 Requirement-Driven Approach

It is implemented according to i* framework, used to design the actions of actors, and the relationship existing between the agents in DW environment. This context helps the designers to accomplish an in-depth analysis in order to offer a formal system of the decisional infrastructure.

Strategic dependency model is the one which outlines how DW aids the users to attain business objectives presenting internal dependencies. The Strategic rationale prototype illustrates particular business objectives as well as task. Finally multidimensional schema is created through UML for DW. Here it's challenging to settle the correspondence amongst multidimensional features and its units. To overcome this drawback, reconciling UML schemas are introduced. Finally, the reconciled multidimensional schema is transformed into snowflake/star schema.

2.3 Semi-Automatic Approach

Semi-automatic modeling method/Dimensional Fact Model works depending on the ER (Entity Relationship) schema. Measures are further categorized into additive, semi-additive and non-additive. This method allows the user to determine and modify the dimension hierarchies. Also many to one mapping is done with the analyzed requirements to ensure aggregation [9]. When building dimensions the method ends with orthogonality problem, due to some failure in establishing relationship.

2.4 Manual Approach

In earlier stages, multidimensional data warehouses use manual effort to gather and create the user requirement data sources [22]. It would seem to be a complex task, usually takes more time to process and finally ends with high risk and failure. Furthermore when it performs complex mappings and aggregation it is unable to analyze and deal with the complex data, which results with finding practical solutions to reduce risk and failure rate when dealing with this method.

2.5 Automatic Approach

Automatic approach/mixed approach uses automated designing of data marts and formation of DW schema. In this work, we make use of this automatic approach via integration of both the data driven as well as the requirement driven approach. This is because our proposed work dealt with the multidimensional models using extended DFM model and reconciled UML schemas to design and manipulate the underlying data sources.

3. CONCEPTUAL DESIGN OF DATA WAREHOUSE

Our methodology integrates the features of hybrid methodologies such as data-driven and requirement-driven methods to preserve the mislaid user needs and by allowing the designers to create and execute the data modeling activity. It encloses various features that combine the process of requirement analysis and conceptual design. So that it works based on the multi-dimensional models such as UML model,

which is used to exemplify the requirement based multidimensional schema and Extended DFM (EDFM) model, to symbolize a tree style view. With this, the designer holds the access to assign many to many relationships and permits to add or remove the nodes from the attribute tree. The Overall framework of our proposed work is briefly demonstrated by the following steps:

3.1 Requirement Analysis

The process starts with deep domain analysis, performed to detect and analyze various business goals. Analyzing allows the designer to identify and formulate the user requirements. These process were carried out by defining an i* framework introduced to construct the actions of the actor, his task and his relationship with the decision makers in the data warehousing environment. This model is named as the strategic dependency model, outlines the user to reach their business goals. The user requirements are classified into strategic goals, information goals and decision goals. Let us consider an example, when a company needs to improve its fame and trend then it would increase its sales noted as strategic goal, whereas to achieve its marketing goal, a company should focus on advertisements, which is said to be decision goal and finally identifying the best client is treated as the information goal. Furthermore the detailed description about the actor will be provided by the strategic rationale model.

3.2 Multidimensional Modeling

Defining a multi-dimensional schema is done manually with the gathered user requirements and the decision goals. Here in this section a fact class has been created for the main actor, who plays a major role in the data warehouse environment. The multi-dimensional schema is modeled according to UML using both strategic dependency and rationale model defined for data warehouse.

3.3. Reconciliation

This approach uses conventional hybrid methodology (Requirement-driven approach) to develop a reconciled UML schema. When trying to provide the reconciled UML schema as input to data modeling, it gives rise to a schema translation issue, which is solved by performing mapping between two multi-dimensional models. This Mapping of reconciling multidimensional schemas on data sources will resolve inconsistency. The process of mapping is performed automatically by a QVT (Query View Transformation) facility [23, 24]. QVT will automatically perform the mapping process into the targeted platforms. It delivers a standard language format called Model Transformation Language (MTL) to perform mapping and relations.

3.4. Formation of attribute tree

Once the reconciled UML schema is created, then the process is followed by the automatic generation of attribute tree using a DFM model defined in the data-driven approach. This can be obtained via the following algorithm.

```
Root = newVertex (pk (F));    \\ Fact class F
translate (F, identifier (F))
where
translate (E, v):  \\ E be the current entity, V be the current vertex
{
  for each attribute a∈E
do
addchild (v, a);  \\ a be the child, which is added to the vertex v
   vertex v
  {
    for each attribute b∈R
do
    addchild (v, b);
    addchild (v, identifier(G));
    translate (G, identifier(G));
  }
}
```

The algorithm constructs an attribute tree for the relational schema, but it doesn't support many to many relationships. In order to resolve this issue, Extended-DFM model has been proposed. Some major functions used in this recursive algorithm are,

- base(d) will extracts the first base of the d dimension class
- descriptor(b) excerpts the descriptor of b, base class
- cardinalityRolls-upTo returns cardinality amongst the nodes u and b.

The algorithm used in extended DFM model is described below:

```
u=root (f)
 for each m in f_m                  // set of fact attributes of f
      add (m, u)
end for
for each d in f_d                   // set of dimension classes of f
       b=base(d)                    // b be the base class
       explore (b, u)
 end for
function explore (b, u)          // recursive function
k=descriptor(b)
h=descriptor(u)
n=cardinalityRolls-upTo(b, u)
if(n==1)
   add (k,h)
end if
for each v in b
     ad(v, k)
end for
```

```
for each c such that b rolls-up-to c
     explore(c, b)
end for
end function
```

3.5 Advanced Data Modeling

Advanced data modeling is also done based on the Extended DFM model. It permits manual remodeling of the attribute tree by the designer and thereby modifying the functional dependencies. Remodeling implies the intuitive operations like addition or removal of nodes from attribute tree. This process is termed as pruning and grafting of attribute tree. This can be done by the following step:

```
graft (v):                  //v be the vertex
{
  for each v"             // v" be the child of v
do
  addChild(v', v");
  drop v;
}
```

Grafting is done for shifting the whole sub-tree along with its root in v to v'. It is usually carried out when the vertex of tree holds unwanted data. Whereas the process of pruning done by dropping the sub_tree from the parent tree. This process will improve the hierarchical relationship existing among dimensional levels.

4. RESULTS AND DISCUSSIONS

In this work for a university named ABC, a data warehouse has been created using the hybrid methodologies and its design is done with respect to the conceptual UML schemas. Here for the initial step, the user requirements were gathered. The primary actor is named as rector and the second actor as data warehouse; they are supposed to get the information from numerous data resources and are provided to the decision makers. With i* framework, the decision makers are assigned with the task of identifying the goals. Here the requirements are modeled as resource dependencies. With this model, the strategic rationale prototype for each actor has been defined. The strategic dependency framework for the actors was shown in the Figure 1.

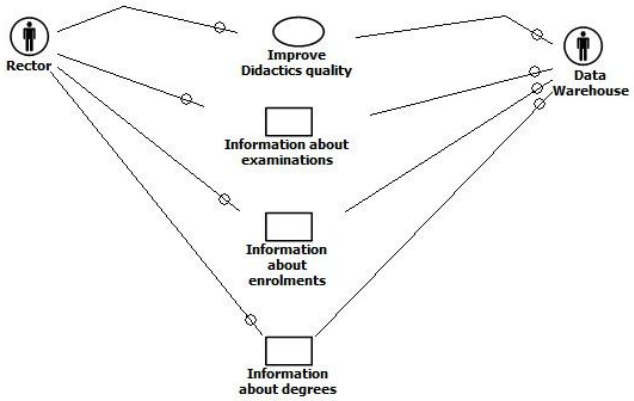

Figure 1: Strategic dependency model

The goals of a University ABC is described based on top-down approach, where decisions goals are to raise the number of graduated students, average rating of students and their enrolments. Also identifying the cities from where the students joined the university and the least populated faculties are set as the information goals. The diagram to represent the strategic rationale archetype for warehouse is created and displayed in Figure 2.

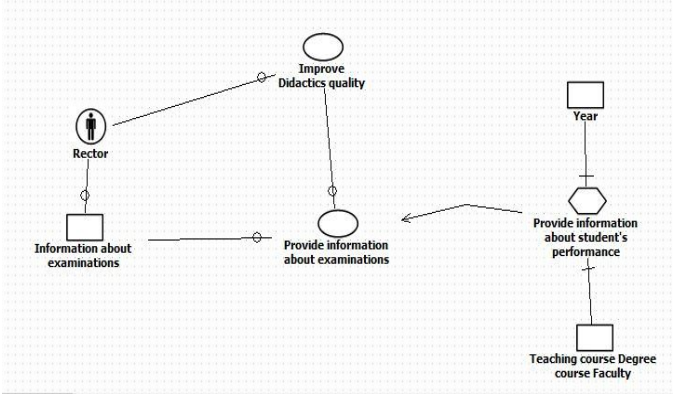

Figure 2: Strategic Rationale model for DW

With the strategic dependency as well as strategic rationale models, the multi-dimensional schema is modeled. Here for this MD schema, the fact class has been created for the actor, who plays a major role in the data warehouse environment. And then the reconciled UML schema is modeled using the conventional hybrid methodology, where mapping is done with the QVT facility. A database is then created for the university ABC with the fact class, dimension attributes and measures. The diagram to represent the relational schema for the university ABC's database is created and is shown in the Figure 3.

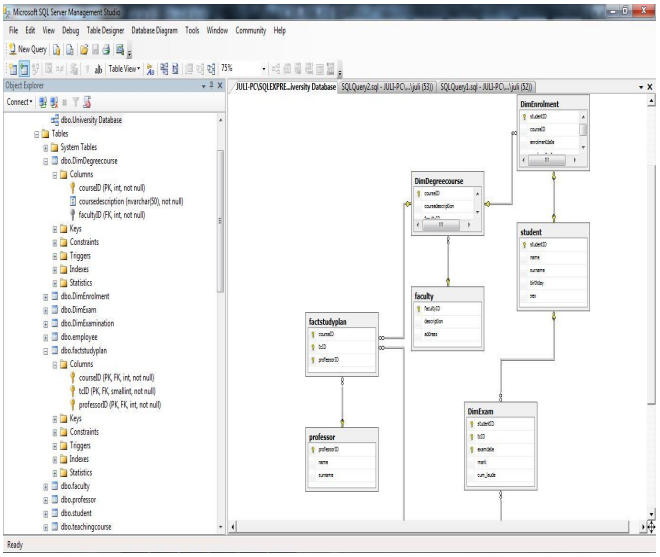

Figure 3: Database relational model created for university ABC

Here in this relational schema diagram, the fact_study_plan is considered as the fact class, whereas the dimensions are, faculty, student, teaching course etc. Also the dimension attributes are course-id, course-description, etc. The QVT mapping process will results with the reconciled UML schema formation. Following this an attribute tree has been automatically constructed. Finally in advanced data modeling the extended DFM is used to remodel the entire tree by performing pruning and grafting process. It allows proper schema alignment with essential user requirements, which result with improved hierarchical relationships between the various dimensional levels. Also an attribute tree has been generated for the relational schema using a recursive algorithm and is shown in the Figure 4.

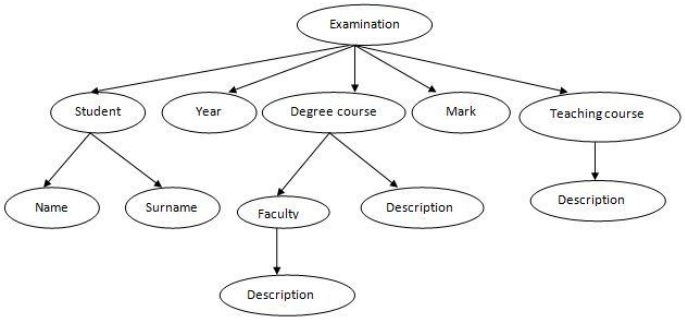

Figure 4: Attribute tree

With advanced data modeling, the attribute tree has been grafted or pruned based on extended DFM model. It allows the user to remodel the attribute tree by removing or adding with the needed nodes. Such a remodeled attribute tree is shown in Figure 5.

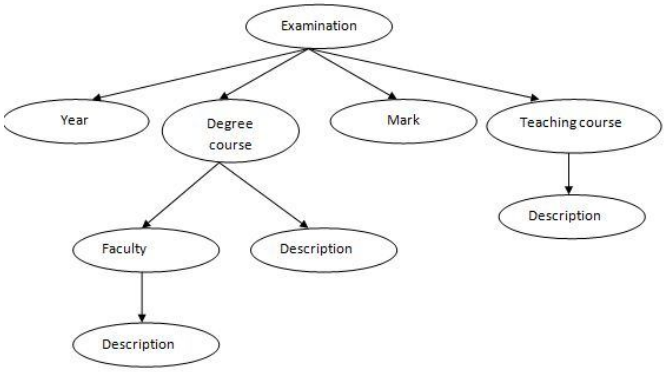

Figure 5: Remodeled attribute tree

Thus the remodeled attribute tree can be modeled into cube and is finally represented by a snowflake representation based on the relational model.

4.1 Performance Evaluation

To evaluate the quality of schema, various metrics were commonly used. Some of the metrics used in logical programming (schemas) are: RFK (Ratio of Foreign Keys), NDT (Number of Dimension Tables), NFK (Number of Foreign Keys), NADT (Number of Attributes of Dimension Tables), NA (Number of Attributes of a star), NT (Number of Tables of star), and NAFT (Number of Attributes including Number of Foreign Keys of a Fact Table). Furthermore the star flake schema metrics are: NBC (Number of Base Classes), NC (Total Number of Classes), RBC (Number of Base Classes per Dimension), NA (Total Number of Attributes), and NH (Number of Hierarchy relationships of the schema).

These metrics are used here for better understandability about various schemas. If the schema representations are easy for the user to understand, then the schema is said to be having high quality. Let consider the fact table fact_study_plan for analyzing the various metrics. While performing comparison between schemas, consider the metric NA, if the number of attributes in first schema is less than the total attributes in second schema, then it is assumed that the quality of first schema is higher than second schema. This is because the quality of metrics is affected due to the complexity of terms, where the schema with lower complexity is treated as the schema with good quality. RFK calculates the ratio of foreign keys in fact table to the total number of attributes. Similarly RSA find the ratio of attributes in the dimension table to the total number of attributes in the fact table. Some of the results to estimate the schema quality are described in Table 1 and are graphically represented in Figure 6.

Table 1: Quality Evaluation Metrics

Metrics	Automatic approach [11]	Requirement driven method [17]	Data driven method [21]	Hybrid method
NDT	03.00	03.0	04.0	04.0
NT	04.00	04.0	05.0	06.0
RSA	02.0	02.6	01.6	02.8
RFK	00.6	00.3	00.6	00.2
NA	12.0	11.0	16.0	10.0

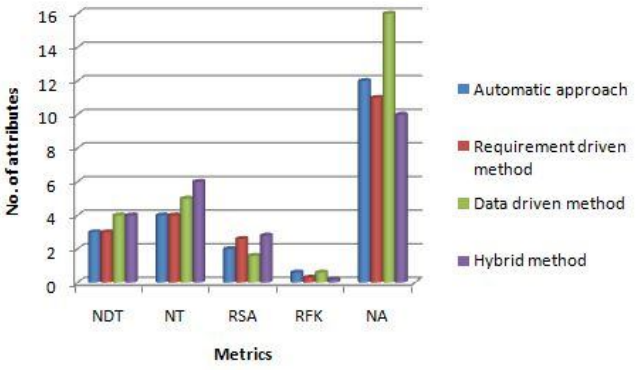

Figure 6: Graphical Representation of metrics quality evaluation

Therefore, the superiority of proposed technique is evaluated and compared with existing techniques through the preliminary outcomes. The proposed hybrid methodology is found to be high quality and this is due to the fact that, it acquires the most reliable features from data source as compared to existing techniques.

5. CONCLUSION

The requirement-based data warehouse driven technique is the only approach to acquire entire requirements of users. On the other hand, i* framework is the fundamental step implemented to reach this goal. However, an extension of i* method called as UML schema reconciliation with data resources produce a multidimensional system that captures user necessities and agrees with the protocols in data source. The novelty of the proposed work lies in this point of using a reconciled schema in DW applications. Moreover, this is achieved by integrating the requirement-driven method with data-driven to probably perform modification in the functional dependencies of reconciled UML schemas and to enhance its conceptual quality. Furthermore, it introduces an automatic design technique that works on the basis of logical programming. The proposed hybrid method is superior to similar existing techniques in terms of quality of schema, DW fully drawn based on user necessities, value the user goals and so on.

REFERENCES

1. N.M. Alotaibi, M. Abdullah, and H. Mosli. **Agent-based Big Data Mining**. *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1.1, pp. 245-252, 2019. https://doi.org/10.30534/ijatcse/2019/4481.12019
2. N.H.Z. Abai, J.H. Yahaya, and A. Deraman. **User requirement analysis in data warehouse design: a review**, *Procedia Technology*, Vol. 11, pp. 801-806, 2013.
3. R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, and B. Becker. **The Data Warehouse Lifecycle Toolkit**, 2009.
4. D.L. Moody, and M.A. Kortink. **From enterprise models to dimensional models: a methodology for data warehouse and data mart design**. *In DMDW*, p. 5, June 2000.
5. A. Thio-ac, E.J. Domingo, R.M. Reyes, N. Arago, R. Jorda Jr, and J. Velasco. **Development of a Secure and Private Electronic Procurement System based on Blockchain Implementation**. *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 5, 2020.
6. J. Trujillo, M. Palomar, J. Gomez, and I.Y. Song. **Designing data warehouses with OO conceptual models**, *Computer*, vol. 34, no. 12, pp. 66-75, Dec. 2001. https://doi.org/10.1109/2.970579
7. M. Jayakrishnan, A.K. Mohamad, and M.M. Yusof. **Information System for Integrative and Dynamic Railway Supply Chain Management**. *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, March 2020. https://doi.org/10.30534/ijatcse/2020/191922020
8. E. Malinowski, and E. Zimanyi. **Hierarchies in a multidimensional model: From conceptual modeling to logical representation**. *Data & Knowledge Engineering*, vol. 59, no. 2, pp. 348-377, Nov. 2006. https://doi.org/10.1016/j.datak.2005.08.003
9. W. Tebourski, W.B.A. Karaa, and H.B. Ghezala. **Semi-automatic Data Warehouse Design methodologies: a survey**. *Int. J. Comput. Sci. Issues IJCSI*, vol. 10, no. 5, p. 48, 2013.
10. J.N. Mazón, J. Pardillo and J. Trujillo. **A Model-Driven Goal-Oriented Requirement Engineering Approach for Data Warehouses**. *ER Workshops*, pp. 255–264, Nov. 2007. https://doi.org/10.1007/978-3-540-76292-8_31
11. O. Romero, and A. Abelló. **A framework for multidimensional design of data warehouses from ontologies**. *Data & Knowledge Engineering*, vol. 69, no. 11, pp. 1138-1157, Nov. 2010. https://doi.org/10.1016/j.datak.2010.07.007
12. O. Romero, and A. Abello. **Automatic Validation of Requirements to Support Multidimensional Design**. *Data & Knowledge Engineering*, vol. 69, no. 9, pp. 917–942, Sep. 2010. https://doi.org/10.1016/j.datak.2010.03.006

13. B. Hüsemann, J. Lechtenbörger, and G. Vossen. **Conceptual Data Warehouse Design**. *Design and Management of Data Warehouses*, Sweden, July 2000.
14. O. Romero, and A. Abelló. **Automating Multidimensional Design from Ontologies**. *DOLAP'07*, Lisboa, Portugal, Nov. 2007. https://doi.org/10.1145/1317331.1317333
15. K. R. Winter, and B. Strauch. **A method for demand-driven information requirements analysis in DW projects.** *Proc. of 36th Annual Hawaii Int. Conf. on System Sciences*, IEEE, 2003, pp. 231–239.
16. I.Y. Song, R. Khare, and B. Dai. **SAMSTAR: a semi-automated lexical method for generating STAR schemas from an ER diagram**. *Proc. of the 10th Int Workshop on Data Warehousing and OLAP*, ACM, 2007, pp. 9–16. https://doi.org/10.1145/1317331.1317334
17. J.N. Mazón, J. Trujillo, and J. Lechtenbörger. **Reconciling requirement-driven data warehouses with data sources via multidimensional normal forms**, *Data Knowl. Eng*, Vol. 63, pp. 725–75, 2007.
18. F. Di Tria, E. Lefons, and F. Tangorra. **Cost-benefit analysis of data warehouse design methodologies**. *Information Systems*, Vol. 63, pp. 47-62, 2017. https://doi.org/10.1016/j.is.2016.06.006
19. J.N. Mazon, and J. Trujillo. **A hybrid model driven development framework for the multidimensional modeling of data warehouses**. *ACM SIGMOD Record*, vol. 38, no. 2, pp. 12-17, 2009. https://doi.org/10.1145/1815918.1815920
20. D. Ibragimov, K. Hose, T.B. Pedersen, and E. Zimányi. **Towards exploratory OLAP over linked open data–a case study**. *In Enabling Real-Time Business Intelligence*, Springer, Berlin, Heidelberg, 2015, pp. 114-132. https://doi.org/10.1007/978-3-662-46839-5_8
21. M. Golfarelli, D. Maio, and S. Rizzi. **The dimensional fact model: a conceptual model for data warehouses**. *International Journal of Cooperative Information Systems*, vol. 7, no. 2, pp. 215–247, June 1998. https://doi.org/10.1142/S0218843098000118
22. R.A. Ahmed, and T.M. Ahmed. **Generating data warehouse schema**. *International Journal in Foundations of Computer Science & Technology (IJFCST)*, vol. 4, no. 1, Jan. 2014. https://doi.org/10.5121/ijfcst.2014.4101
23. S. Luján-Mora, J. Trujillo, and I.Y. Song. **A UML profile for multidimensional modeling in data warehouses**. *Data & Knowledge Engineering*, vol. 59, no. 3, pp.725-769, Dec. 2006. https://doi.org/10.1016/j.datak.2005.11.004
24. M. Thangamani, and V.R.K. Chandar. **Adverse Drug Reactions using Data Mining Technique**. *Journal of Excellence in Computer Science and Engineering*, vol. 1, no. 1, pp. 11-14, 2015. https://doi.org/10.18831/djcse.in/2015011002