ISSN 2278-3083

International Journal of Science and Applied Information Technology (IJSAIT), Vol. 3 , No.3, Pages : 32 - 35 (2014)
Special Issue of ICCET 2014 - Held during July 07, 2014 in Hotel Sandesh The Prince, Mysore, India

Comparison Of Design Methods Of Data Warehousing

Tiruveedula GopiKrishna
Sirt University, Hoon,Libya, gktiruveedula@gmail.com

Abstract : In this paper I presented a detailed summarization of the main features of each method regarding the criteria introduced in, which provides a common framework to compare and discuss methods surveyed. Methods surveyed are distributed in these tables according to the chronological order. On the other hand, some new approaches focus on considering alternative scenarios than relational sources. I presented the most relevant methods introduced in the literature and a detailed comparison showing. All in all, I discussed the current scenario of multidimensional modeling by carrying out a survey of multidimensional design methods the main features of each approach.

Key words : Data warehousing, dimensional design methods, framework, Factual data, Dimensional Data.

INTRODUCTION

The data warehouse is a huge repository of data that does not tell us much by itself; like in the operational databases, we need auxiliary tools to query and analyze data stored. Without the appropriate exploitation tools, we will not be able to extract valuable knowledge of the organization from the data warehouse, and the whole system will fail in its aim of providing information for giving support to decision making. OLAP (On-line Analytical Processing) tools were introduced to ease information analysis and navigation all through the data warehouse in order to extract relevant knowledge of the organization. This term was coined by E.F. Codd in (Codd,1993), but it was more precisely defined by means of the FASMI test that stands for *fast analysis* of *shared* business *information* from a *multidimensional* point of view. This last feature is the most important one since OLAP tools are conceived to exploit the data warehouse for analysis tasks based on *multidimensionality*. The multidimensional conceptual view of data is distinguished by the *fact* / *dimension* dichotomy, and it is characterized by representing data as if placed in an n-dimensional space, allowing us to easily understand and analyze data in terms of facts (the subjects of analysis) and dimensions showing the different points of view where a subject can be analyzed from. One fact and several dimensions to analyze it produce what is known as *data cube*. Multidimensionality provides a friendly, easy-to understand and intuitive visualization of data for non-expert end-users. These characteristics are desirable since OLAP tools are aimed to enable analysts, managers, executives, and in general those people involved in decision making, to gain insight into data through fast queries and analytical tasks, allowing them to make better decisions [1].

Developing a data warehousing system is never an easy job, and raises up some interesting challenges. One of these challenges focus on modeling multidimensionality. Nowadays, despite we till lack a standard multidimensional model, it is widely assumed that the data warehouse design must follow the multidimensional paradigm and it must be derived from the data sources, since a data warehouse is the result of homogenizing and integrating relevant data of the organization in a single and detailed view.

UNIT- I GENERAL ASPECTS

The general criteria are summarized into nine different items:

• *Paradigm*: According to (Winter & Strauch,2003), multidimensional modeling methods may be classified as *supply-driven*, *demand-driven* or *hybrid* approaches. The reader may found a slightly different classification in (List et al., 2002). Furthermore, we distinguish between sequential and interleaved hybrid approaches (depending if their supply-driven and demand-driven approaches are performed either sequentially or simultaneously or sequentially) [3].

• *Application*: Most methods are semi-automatic. Thus, some stages of these methods must be performed manually by an expert (normally those stages aimed to identify factual data) and some others may be performed automatically (normally those aimed to identify dimensional data). In general, only a few methods fully automate the whole process. On the contrary, some others present a detailed step-by-step guide that is assumed to be manually carried out by an expert.

• *Pre-process*: Some methods demand to adapt the input data into a specific format that facilitates their work. For example, these processes may ask to enrich a conceptual model with additional semantics or perform data mining over data instances to discover hidden relationships.

• *Input abstraction level*: Most methods (mainly those automatable) work with inputs expressed at the logical level (e.g., relational schemas), whereas some others work with inputs at the conceptual level (e.g., from conceptual formalizations such as ER diagrams or from requirements in natural language).

• *Output abstraction level*: Several methods choose to directly generate a star or snowflake schema, whereas some others produce multidimensional conceptual schemas. Although many approaches argue that the data warehouse method should span the three abstraction levels, only a few of them produce the conceptual, logical and physical schema of the data warehouse.

• *Data sources:* There are three items summarizing the main features about how data sources are considered in the method.

◦ *Type of data sources*: The input abstraction item informs about the abstraction level of the input, whereas this item specifies the kind of technology of the data sources supported by the method. For example, if the method works at the conceptual level it may work from UML, ER conceptual

International Journal of Science and Applied Information Technology (IJSAIT), Vol. 3 , No.3, Pages : 32 - 35 (2014)
Special Issue of ICCET 2014 - Held during July 07, 2014 in Hotel Sandesh The Prince, Mysore, India

schemas or ontologies, and if it works at the logical level it may work from relational schemas or XML schemas.

◦ *Data sources analysis*: Most methods perform a fully supply-driven analysis of the data sources. However, some of them also perform a requirement driven analysis of the data sources. Clearly, this item is tightly related to the paradigm item. Nevertheless, note that a method may follow a hybrid approach but do not consider at all requirements when analyzing the data sources [3].

◦ *Pattern formalization*: Supply-driven stages usually define design patterns to identify the potential multidimensional role that concepts depicted in the data sources may play. Some methods present these patterns in an informal way, but most of them use some kind of structured language. For example, ad hoc algorithms are the most common representation but some other methods use description logic formulas or QVT Transformations.

• *Requirements representation:* If requirements are considered, this item summarizes how they are represented. For example, most methods use ad hoc representations (like forms, sheets, tables or matrixes), whereas some others use UML diagrams or the *i** framework. Finally, some of them lower the level of abstraction of requirements to a logical level by means of SQL queries or MDX queries.

• *Validation:* Some methods integrate a validation process to derive meaningful multidimensional schemas. For example, restricting summarization of data to those dimensions and functions that preserve data semantics or forming multidimensional spaces by means of orthogonal dimensions.

• *Implementation:* Some methods have been implemented in CASE tools or prototypes.

UNIT-II METHODS COMPARISON

In this section I presented a detailed summarization of the main features of each method regarding the criteria introduced in previous section, which provides a common framework to compare and discuss methods surveyed. Results are shown in **Table 1 and 2**. Methods surveyed are distributed in these tables according to the chronological order. A given cell contains information for a method and a specific criterion Some criteria are evaluated as *yes/no*, but most of them have alternative values. Two general values can be found for any criterion:-means that this criterion does not make sense for the method (for example, if it does not consider the data sources then, any of the criteria related to them cannot be evaluated for this method),whereas *none* means that, despite this criterion could be considered for this method, none of the alternatives are considered (i.e., it is overlooked). Therefore, *none* is the equivalent to the *no* value but for criteria having several values. Analyzing these tables we can find some interesting trends as well as assumptions that have been considered in most of the methods surveyed. First approaches tried to contextualize the multidimensional modeling task by providing tips and informal rules about how to proceed. In other words, they presented the first guidelines to support multidimensional design. Later, when main features with regard to multidimensional modeling were set up, new formal and powerful methods were developed. These new methods focused on formalizing and automating the process. Automation is an important feature along the whole data warehouse lifecycle and multidimensional design has not been an exception. Indeed, first methods were step-by-step guidelines, but in the course of time many semi-automatic and automatic approaches have been presented. This evolution also conditioned the type of inputs used, and logical schemas were considered instead of conceptual schemas. Nowadays, last methods introduced present a high degree of automation. Moreover, we may say that this trend also motivated a change of paradigm. At the beginning, most methods where demand-driven or, in case of being hybrid approaches, they gave much more weight to requirements than to data sources. However, eventually, data sources gained relevance. This makes sense because automation has been tightly related to focusing on data sources instead of requirements. Consequently, first methods introduced gave way to others largely automatable and mostly following a supply-driven framework. Nevertheless, today, it is assumed that the ideal approach to design multidimensional data warehouses must be a hybrid approach. In this line, last works introduced are mainly hybrid approaches. In these tables we can also note the evolution of how the multidimensional model has been considered. First approaches used to produce logical multidimensional schemas but later, most of them generate conceptual schemas. One reason for this situation could be that Kimball introduced multidimensional modeling at the logical level (i.e., as a specific relational implementation) [6]. With the course of time, it has been argued that it is necessary to generate schemas at a platform independent level and in fact, the multidimensional design should span the three abstraction levels (conceptual, logical and physical) like in the relational databases field. About the kind of data sources handled, most of the first approaches choose conceptual entity relationships diagrams describing the data sources. ER diagrams were the most spread way to represent operational databases (the most common type of data source to populate the data warehouse) but the necessity to automate this process and the need to provide up-to-date conceptual schemas to the data warehouse designer motivated that many methods worked over relational schemas instead **[2].**

Table 1: Summary of the comparison of multidimensional design methods

	[KRTR98]	[CT98]	[GR09]	[BvF99]	[HLV00]	[MK00]	[BCC+01]	[PD02]	[WS03]
General Aspects									
Paradigm	DD	IH	SH	SH	DD	SD	SH	SH	DD
Application	G	G	S	G	G	G	S	S	G
Pre-process	-	DCS	-	ECS	-	DCS	-	-	-
Input Abstr.	C	C	C/L	C	C	C	C/L	L	C
Output Abstr.	L	L	C/L/P	L	C	L	L	C	C
Data Sources									
↪ Type	-	ER	ER/Rel	SER	-	ER	ER	Rel	-
↪ Analysis	-	RD	Full	Full	-	Full	Full	Full	-
↪ Patterns F.	-	None	Alg	None	-	None	Alg	Alg	-
Req. Expr.	ad hoc	ad hoc	ad hoc	ad hoc	ad hoc	-	ad hoc	MDX	ad hoc
Validation	No	No	No	No	MNF	No	No	No	No
Tool	No	No	Yes	Yes	No	No	No	No	No
Factual Data									
Facts									
Factless Facts	Yes	No	No	No	No	No	No	No	No
Requirements	Expl	Expl	Expl	Expl	Expl	-	Expl	No	Expl
Data Sources									
↪ C.Num.Val.	-	No	No	No	-	Yes	Yes	Yes	-
↪ Connectivity	-	No	No	No	-	No	No	No	-
↪ Cardinality	-	No	No	No	-	No	No	Yes	-
Semantic Rels.	-	-	Ass	-	Ag	-	-	None	None
Measures									
Requirements	Impl	Expl	Expl	Impl	Expl	-	Expl	No	Expl
Data Sources	-	No	NV	No	-	NV	NV	NV	-
Dimensional Data									
Fact-centered	No	No	Yes	Yes	No	No	Yes	Yes	No
Requirements	Expl	Expl	Expl	Expl	Expl	-	Expl	No	Expl
Data Sources									
↪ FDs	-	No	Yes	Yes	-	Yes	Yes	Yes	-
↪ Bases	-	No	No	No	-	No	No	No	-
↪ Others	-	No	No	No	-	No	No	Yes	-
Related									
Interdim.	None	-	None	-	None	-	-	None	None
Intradim.	L/D	L/D	L/D	L	L/D	L/D	L/D	L	-

ISSN 2278-3083

International Journal of Science and Applied Information Technology (IJSAIT), Vol. 3 , No.3, Pages : 32 - 35 (2014)
Special Issue of ICCET 2014 - Held during July 07, 2014 in Hotel Sandesh The Prince, Mysore, India

of conceptual schemas. Almost every method either considers ER diagrams or relational schemas to describe the data sources. Lately, with the relevance gained by the semantic web area, some other works automating the process from XML schemas or OWL ontologies have been presented. About requirements, their representation have varied considerably. At the beginning, ad hoc representations such as forms, tables, sheets or matrixes were proposed but lately, many methods propose to formalize requirements representation with frameworks such as UML diagrams or *i**. Moreover, some works have also proposed to lower the level of abstraction of requirements to the logical level by means of SQL or MDX queries, which opens new possibilities for automating the process. Finally, we can also identify a trend to validate the resulting multidimensional schema as well as the importance to provide a tool supporting the method. About how to identify factual data, there are some trends that most approaches follow. Looking at the data sources, numerical concepts are likely to play a measure role, whereas concepts containing numerical attributes or those with a high table cardinality are likely to play a fact role. First methods were mainly demand-driven but later, most of them used these heuristics to identify factual concepts within supply-driven stages. However, these heuristics do not identify facts or measures but concepts likely to play that role. Thus, requirements must be considered to filter the (vast) amount of results(**shown in Table 1,2**) obtained, and in the last years requirements have gained relevance again. Capturing inter-relationships between schemas (i.e., facts) have also gained relevance lately, as they open new analysis perspectives when considering multidimensional algebras. Finally, the reader may note that although Kimball introduced the concept of factless facts from the very beginning, it has been traditionally overlooked. Lately, some methods considered them again. One of the reasons could be that it is difficult to automate the identification of facts that do not have measures. According to our study, dimensional concepts have been traditionally identified by means of functional dependencies. From the very beginning, some methods proposed to automate the identification of aggregation hierarchies. In fact, many methods use requirements to identify factual data shown in **Table. 2** and later they analyze the data sources looking for functional dependencies to identify dimensional data. May be for this reason, the use of requirements to identify dimensional concepts has not been that relevant as to identify factual data. Another clear trend with regard to dimensional concepts is that, in general, the more automatable a method is, the more fact-centered it is. About relationships among dimensional concepts, inter-dimensional relationships (like relationships between facts) open new perspectives of analysis when considering multidimensional algebras. However, in this case they have been traditionally overlooked; even more than this kind of relationships between facts. On the contrary, intra-dimensional relationships gained more and more relevance from the very beginning. Most methods agree that distinguishing among dimensions, levels and descriptors is relevant for analysis purposes [4].

Table 2: Summary of the comparison of multidimensional design methods

	[VBR03]	[JHP04]	[GRG05]	[ARTZ06]	[PACW06]	[RA10a]	[MTL07]	[SKD07]	[RA10b]	[NBPM+09]
General Aspects										
Paradigm	SH	SD	SH	DD	DD	IH	SH	SD	SH	IH
Application	S	A	S	G	G	A	S	A	S	S
Pre-process	TCS	DM	-	-	ECS	-	-	TCS	-	-
Input Abstr.	L	L	C	C	C	L	C/L	C	C	C/L
Output Abstr.	L	L	C	C	C/L/P	C	C	L	C	C/L
Data Sources										
↪ Type	XML	Inst	ER/Rel	-	-	Rel	Rel	Rel	Ont	Ont
↪ Analysis	Full	Full	RD	-	-	RD	RD	Full	Full	Full
↪ Patterns F.	None	Alg	None	-	-	Alg	QVT	None	DL	DL
Req. Expr.	ad hoc	-	*i**	ad hoc	UML	SQL	*i**	-	-	ad hoc
Validation	No	AC	No	No	AC	MC	MNF	No	MC	Res
Tool	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Factual Data										
Facts										
Factless Facts	No	No	No	No	Yes	Yes	No	Yes	No	No
Requirements	Expl	-	Expl	Expl	Expl	Impl	Expl	-	-	Expl
Data Sources										
↪ C.Num.Val.	No	Yes	No	-	-	No	No	No	Yes	No
↪ Connectivity	No	No	No	-	-	No	No	Yes	Yes	No
↪ Cardinality	No	Yes	No	-	-	No	No	No	No	No
Semantic Rels.	-	-	None	None	None	Ass/S	Ass/S	None	Ass/Ag	None
Measures										
Requirements	Expl	-	Expl	Expl	Expl	Impl	Impl	-	-	Expl
Data Sources	No	NV	No	-	-	No	No	No	NV	No
Dimensional Data										
Fact-centered	Yes	Yes	Yes	No	No	No	No	Yes	Yes	No
Requirements	No	-	Expl	Expl	Expl	Impl	Impl	-	-	Expl
Data Sources										
↪ FDs	Yes	Yes	Yes	-	-	Yes	Yes	Yes	Yes	No
↪ Bases	No	No	No	-	-	No	No	No	Yes	No
↪ Others	No	No	No	-	-	No	No	No	No	No
Related										
Interdim.	-	-	None	None	None	Ass/S	S	None	Ass	None
Intradim.	L/D	L/D	L/D	L	L/D	L/D	L/D	L	L/D	L/D

UNIT-III FACTUAL DATA

These criteria summarize how a given method identifies and handles factual data (**i.e., Table.2 shows** facts and measures). First, criteria used to identify measures are summarized as follows:

• Data sources: Up to now, looking for numerical concepts is the only heuristic introduced to identify measures from the data sources [6,7].

• Requirements: Most approaches consider requirements to identify measures. We distinguish if the method only considers explicit measures or also implicit ones. Implicit measures are those explicitly stated in the requirements but implicit in the data sources (i.e., there is not a concept in the data sources that would correspond to it, but they can be derived from an already existing concept(s) in the data sources). For example, derived measures. Therefore, some kind of reasoning over the data sources is needed. Next, we introduce criteria used to identify facts. These criteria refer to how facts are identified from the data sources or from requirements, and how they may be semantically related in the resulting schema:

• *Factless facts*: This kind of facts were introduced by Kimball in (Kimball et al.,1998). they are also known as empty facts and they are very useful to describe events and coverage, and a lot of interesting questions may be asked from them.

• *Data sources:* Most of the methods demand to explicitly identify facts by means of the requirements, but some others use heuristics to identify them from the data sources. For example, in case of relational sources, most use heuristics such as table cardinalities and the number of numerical attributes 99 that a table contains. Furthermore, some works also look for concepts with high tone connectivity (i.e., with many potential dimensional concepts).

• *Requirements*: Similar to measures, if requirements are considered, we distinguish among explicit and implicit facts. We denote by implicit facts those that have not been

explicitly stated in the requirements but can be identified from a requirement driven analysis of the sources.

• *Semantic relationships:* In case of producing a conceptual schema, some methods

are able to identify semantic relationships between facts. We distinguish among associations, aggregations (also called roll-up/drill-down relationships) and generalizations. In the multidimensional model, it means that we may perform multidimensional operators such as drill-across or drill-down over them [5].

UNIT-IV DIMENSIONAL DATA

These criteria analyze how the method identifies and handles dimensional data (i.e., dimensions, levels and descriptors). We have two main groups of items. Those referring to how dimensional data is identified (either from the data sources or from requirements), and how they are semantically related in the resulting schema **shown in Table 1,2**. The process to identify dimensions, levels and descriptors must be understood as a whole and, unlike criteria used to identify factual data, we do not distinguish among criteria to look for different dimensional concepts. Roughly speaking, most approaches start looking for concepts representing interesting perspectives of analysis and from these concepts they look for aggregation hierarchies (i.e., levels). The whole hierarchy is then identified as a dimension and level attributes are considered to play a descriptor role:

• *Fact-centered:* Most methods look for dimensional data once they have identified facts. From each fact, dimensional concepts are identified using a wide variety of techniques according to the method inputs, but always looking for functional dependencies starting from the fact.

• *Data sources:* There are several techniques to identify dimensional concepts from data sources. We classify these techniques in three main groups: discovering functional dependencies, discovering bases and others. At the conceptual level, functional dependencies are modeled as to-one relationships, and at the logical level it depends on the technology. For example, in the relational model, dimensional concepts are identified by means of foreign keys and candidate keys. Bases are used to identify dimensional concepts as well. In this case, the method looks for candidate multidimensional bases in order to identify interesting perspectives of analysis (i.e., levels).

• *Requirements:* Dimensional concepts are mostly identified from the data sources once facts and measures have been identified. However, demand-driven approaches rely on requirements to identify dimensional concepts and some hybrid approaches also enrich their supply-driven stages with requirements. Like facts, we distinguish between explicit dimensional concepts and implicit ones.

• *Intra-dimensional*: Most of the methods distinguish between descriptors and levels, but some others do not.

• *Inter-dimensional:* Some approaches are able to identify semantic relationships between dimensions. In this case, we consider associations and generalizations as potential relationships [6,7].

CONCLUSION

For each method captured its main features that were mapped onto different criteria. If a method introduced a new criterion, the rest of works were analyzed to know their assumptions with regard to this criterion. Therefore, criteria presented were defined along an iterative process during the analysis of the multidimensional design methods. I summarized these criteria in three main categories: general aspects, dimensional data and factual data. General aspects refer(shows in Table. 1,2) to those criteria regarding general assumptions made in the method and dimensional and factual data criteria refer to how dimensional data and factual data are identified and mapped onto multidimensional concepts. All in all, I have provided a comprehensive framework to better understand the current state of the area as well as its evolution.

ACKNOWLEDGEMENT

I would like to thank to all my lab assistants to provide me to test the results during evaluation of data for this work.

REFERENCES

[1] Carmè, A., Mazón, J. N., & Rizzi, S. (2010). A Model-Driven Heuristic Approach for Detecting Multidimensional Facts in Relational Data Sources. *Proceedings of 12th International Conference on Data Warehousing and Knowledge Discovery; Vol. 6263, Lecture Notes of Computer Science* (pp,13-24). Bilbao, Spain: Springer.

[2] Golfarelli, M., & Rizzi, S. (2009). *Data Warehouse Design. Modern Principles and Methodologies.* McGraw Hill.

[3] Google. (2010). Google Scholar. Retrieved October, 15th, 2010, from http://scholar.google.com/.

[4] Harzing (2010). Publish or Perish. Retrieved October, 15th, 2010, from http://www.harzing.com/pop.htm.

[5] Prat, N., Akoka, J., & Comyn-Wattiau, I. (2006). A UML-based Data Warehouse Design Method. *Decision Support Systems*, *42*(3),1449–1473.doi:10.1016/j.dss.2005.12.001.

[6] Romero, O., & Abelló, A. (2010a). Automatic Validation of Requirements to Support Multidimensional Design. *Data & Knowledge Engineering*, *69*(9), 917–942. doi:10.1016/j.datak.2010.03.006

[7] Romero, O., & Abelló, A. (2010b). A Framework for Multidimensional Design of Data Warehouses from Ontologies. *Data & Knowledge Engineering*, *69*(11), 1138–1157. doi:10.1016/j. datak.2010.07.007.