International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT21752

Hybrid Data Warehouse Development Method

Pooja D. Kavishwar*, Dr. S. R. Pande
Department of Computer Science, Shivaji Science College, Congress Nagar, Nagpur, Maharashtra, India

Article Info
Volume 7, Issue 5
Page Number : 09-15

Publication Issue :
September-October-2021

Article History
Accepted : 07 Sep 2021
Published : 15 Sep 2021

ABSTRACT

Building a data warehouse is a new discipline and has no concrete strategy for its development process. Currently there are three development approaches for building a data warehouse: Data driven. Goal-driven and User-driven. These development approaches are compared on the basis of certain parameters and by this comparison a new Hybrid multidimensional development methodology has been evolved. This Hybrid multi-dimensional Data model is a combination of Data driven methodology with Business driven which is Goal- driven methodology. We have stated in this paper that this model starts by collecting Business requirements and deriving Fact and Dimension tables along with its multiple constraints which defines their relations. After which a logical structure of the model can be built. Which in turn could be developed into a physical model and can be populated by data for Mining and Analysing. This new multidimensional model can be compared on the same parameters which were used to compare the stated three methodologies and thus we can come up with enhanced features.

Keywords : Data Warehouse, Data Warehouse Methodology, Hybrid multi-dimensional Data model.

I. INTRODUCTION

A Data Warehouse (DW or DWH) is a process used for reporting and performing data analysis, it is considered a core component of business intelligence. DWs are central repositories of integrated data from one or more external and varied sources.

According to Barry Devlin, IBM Consultant, "a DW is simply a single, complete and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context". According to W.H. Inmon, "a DW is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process" [1, 2].

The data stored in the warehouse is uploaded from the system which processes day to day transactions of an organisation. To ensure the data quality before it can be used for reporting and analysing in the DW it might pass through an OLTP and for additional operations it may require data cleansing.

According to Watson and Haley, the significant likely benefits of the data warehouse occur when it is used

to redesign business processes and to support strategic business objectives.[3].

Building a data warehouse may be a very challenging issue because compared to software engineering it's quite a new discipline and doesn't yet offer well-established strategies and techniques for the development process. plenty of projects fail because of the complexity of the development process. up to now there's no common strategy for the development of data warehouses. Current data warehouse development methods can be categorized into three basic groups: data-driven, goal-driven and user-driven.[4]

The database community is devoting increasing attention to the research themes concerning data warehouses (DWs); nevertheless, the crucial issues related to DW design have not been deeply investigated yet [5].

Designing a DW requires techniques completely different from those adopted for OLTP systems. While most scientific literature on the design of DWs focuses on specific issues such as multidimensional data models, materialization of views and index selection [6], no significant effort has been made so far to develop a complete and consistent design methodology [7]. The different phases in DW design are described informally, but no impromptu conceptual model to support them is devised.

We evaluate these development methodologies by the means of application areas, targeting organisational level, extent of end user involvement, duration of development and completion, complexity of data model, amount of source systems and longevity of data model.

This paper presents a hybrid-driven data warehouse modelling method. We use ontology to describe the source of the data in order to achieve clean data. This hybrid approach combines the data-driven method and goal-driven(Business-driven) method to make the resulting data warehouse good features in practical application and describing the requirements.

II. DATA WAREHOUSE DEVELOPMENT METHODOLOGY

A. Data-Driven Methodologies

Bill Inmon, the founder of data warehousing argues that data warehouse environments are data driven, in comparison to classical systems, which have a requirement driven development life cycle [8].He states that initially data warehouses have to be populated with data and results of queries have to be analysed by users and then requirements are considered in the decision support development life cycle. The data warehouse development procedure is based on the analysis of the corporate data model and relevant transactions. The approach passes over the needs of data warehouse users and also does not reflect company goals and user requirements at all. User needs are integrated in the second cycle. Golfarelli, Maio and Rizzi propose a semi-automated methodology to build a dimensional data warehouse model from the pre-existing E/R schemes that represent operational databases[9].

B. Goal-Driven Methodologies

Böhnlein and Ulbrich-vom Ende present an approach that is based on the SOM (Semantic Object Model) process modelling technique in order to derive the initial data warehouse structure [10]. The first stage of the derivation process focuses on goals and services the company provides to its customers. Then the business process is examined by applying the SOM interaction schema that features the customers and their transactions with the process under study. In a third step order of transactions are transformed into order of existing dependencies that refer to

information systems. The last step recognizes measures and dimensions: One has to find most executed(information request) transactions for measures and get dimensions from existing dependencies. In our opinion this highly complex technique works well when business processes are designed throughout the company and are combined with business goals. Kimball proposes a four-step approach where he starts to choose a business process, takes the grain of the process, and chooses dimensions and facts [6]. He defines a business process as a major OLTP process in the organisation that is supported by some kind of heritage system (or systems).

C. User-Driven Methodologies

Westerman describes an approach that was developed at Wal-Mart and has its main focus on implementing business strategy [11]. The methodology assumes that the company goal is the same for everyone and the entire company will therefore be chasing the same direction. It is proposed to set up a first model based on the needs of the business. Business people define goals and gather, prioritise as well as define business questions reinforcing these goals. Afterwards the business questions are prioritised and the most important business questions are described in terms of data elements, including the definition of hierarchies. Although the Wal-Mart approach focuses on business needs and business goals which are defined by the organisation are not taken into consideration at all. Poe proposes a catalogue for conducting user interviews in order to collect end user requirements [12]. She recommends interviewing different user groups in order to get an in-depth understanding of the business. The questions cover a very broad field and also include topics like job responsibilities.

II. COMPARISON TABLE

Table 1 : Comparison of Three Data Warehouse development methodologies

Methodology Criteria	Data-Driven	User-Driven	Goal-Driven
Basic Approach	Bottom-up	Bottom-up	Top-Down
Project Support	None	Department	Top Management
Application Area / Requirement Domain	Data Exploration and Data Mining	Raise the Acceptance of a System	Foundation for Decision Support
Targeting Organizational Level	Operational Partly Tactical	Depends on the Group of Inter- view Partners	Strategic Tactical Operational
Focus	Short-Term Focus	Short-Term Focus	Long-Term Focus
Extent of End User Involvement	None	High	Moderate
Project Duration	Low	Very High	High
Number of Measures	Many	Many	Few

Type of Measures	Non-Financial and Quantitative Time-Based and Frequency-Based	Non-Financial and Quantitative Time-Based and Frequency-Based	Balanced: Financial and Non-Financial as well as Qualitative and Quantitative
Level of Granularity	Low	Low	High
Number of Dimensions	Few	Many	Few
Type of Dimensions	Represents the Basic Structure of the Application	Represents the Basic Structure of the Application and external Sources	Represents the Strategic Building Blocks of the Organisation
Number of Source Systems	Low	Moderate	High
Longevity / Stability of Data Model	Long	Short	Long
Cost	Low	High	High

A. Analysis of Table

We have evaluated different data warehouse development methodologies. This section compares all these methodologies and ends up establishing a link between them and requirement domains. A comparison of the same has been placed in Table 1.
Basically, a monopolisation of this user-driven development methodology is high risk and must be avoided, as it gives rise to performance information that reflects the organisational level of the people involved. Therefore, selected measures, dimensions, the level of granularity and the targeting level of the organisational hierarchy are very unsteady. The methodology has a bottom-up tendency, because most employees do not see the organisation from a broader angle. The project duration may be lengthy and expensive, as project participants request long discussions on a lot of unnecessary measures and dimensions. Hence, inspecting the criteria of the user-driven methodology does not make sense, because results change with the people involved. This development methodology may well raise acceptance of a system, but must be amalgamated with the data-driven or goal-driven development methodology in order to improve the longevity of the system. The more a system suffers refusals, the more user involvement is required beside a focus on organisational strategies or the corporate data model.

The goal-driven development methodology supports modern management methods and is a base for decision support at all organisational levels. The level of granularity is higher compared to that of the data-driven approach. The development duration of the project tends to be lengthy and costly, as a lot of highly qualified professionals and managers take part in numerous workshops and derive performance indicators from strategy. End-users are involved when operational detail matters. As the model is situated with the corporate strategy, it is very stable. Measures and dimensions are balanced: financial, non- financial, qualitative and quantitative aspects are reviewed.

The data-driven development methodology is recommended for data mining and data exploration purposes. The bottom-up approach makes extensive use of the database. The data-driven development methodology is particularly convenient for production workflows and thus create a high business value, have a high degree of repetition, are customer focused, often time critical and therefore require compact and close monitoring. All development methodologies have been applied and measure the process cycle time. The goal-driven development methodology measures solely the process cycle time. Working time and waiting time are differentiated by the user-driven development methodology, while the data-driven development methodology evaluates three states: ready, suspended and running.

The working time in a process cycle is equal to the running state. The ready and the suspended states narrates the waiting time in more detail. The ready state narrates the duration a work item is assigned to a user and has not been accessed before. The suspended state narrates the duration a work item is moved off the work list because it cannot be processed because information required is not accessible. The disintegration of the waiting time into different states enables the detection of work overload, missing resources or lazy employees. The longevity of the data model is directly related to the robustness of the structure of the underlying system. As no end users and no other source systems were involved, the project duration was very short-term which contrasted the goal-driven approach. Due to the restriction of the audit trail, measures and dimensions are time- based. Their main target is the operative level of the organisation.

The data-driven and goal-driven development methodology do not stress inconsistency of data. As they follow different purposes they may exist in parallel. As the drill-down path of the goal-driven development methodology, the data-driven development methodology can even be seen as a lower level of detail. These methodologies are interdependent and when used in parallel, the benefit is even higher.

III. PROPOSED MODEL

In this modelling framework, ontology is used to differentiate fact and dimension tables; business requirements are used to decide business process and granularity.

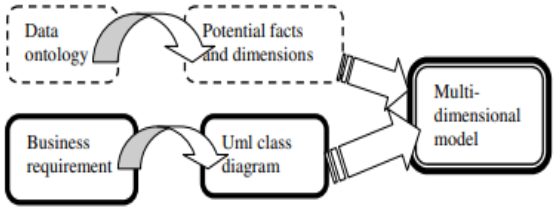

Fig : 1. The Framework of Hybrid-Driven data warehouse modelling method.

In Fig : 1, the main framework of this method is divided into two parts. First the constraints of ontology concepts are analysed to find the potential facts and dimensions in these ontology concepts; and then depending on the business needs, UML class diagrams and business process diagrams are established.

According to the definition in The Data Warehouse Toolkit, the fact table is the main table to store measures of business; the Dimensions table reflects the different business analysis perspective. These Fact and Dimension tables are defined from Domain knowledge but in this Hybrid Driven Data Warehouse Model these facts and dimensions are determined using business concepts. The multiple constraints of the fact table should be analysed to clear the multiple relationships between fact and dimension, so as the potential fact and dimension could be identified by ontology constraints. Business

requirements should be obtained and then a dual group called requirement set is defined to represent the business requirements. Depending on the various conditions in the business a logical model is determined using UML class diagram by either creating a single UML class diagram or a combined UML class diagram with business process. Classes in a class diagram are mapped to Dimensions tables and Fact tables. The class properties in a class diagram are mapped to columns of dimension tables or fact tables. The associations between classes in a class diagram are associated with relationships between the dimension tables and the fact tables. After deriving the Logical model which has potential facts and Dimensions sets depending on Business requirement and Business process. Then this multi dimensional model should be populated with data and mined for various needs of the business.

IV.CONCLUSION

This new Hybrid multi dimensional Model which is a blend of data driven and goal(Business) driven methodology has combined benefits. Data driven methodology and Goal driven methodologies are complementary to each other and when used in parallel gives higher benefits. This multidimensional model is supported by Top Management employees and thus lays the foundation for decision support and covers Data Mining and Data Exploration application areas. This model targets both operational and strategic tactical operations. End user is moderately involved in this model. Types of measures in this model are Financial, Non Financial, Qualitative, Quantitative and Time and frequency based. Thus,this model represents both basic structure of application as well as strategic building blocks of organisation. And thus we can say that this hybrid project holds a longer stability of the data model.

V. REFERENCES

[1]. Franconi E., Introduction to Data Warehousing, Lecture Notes, http://www.inf.unibz.it/~franconi/teaching/2002/cs636/2 ,2002

[2]. W. H. Inmon, “Building the Data Warehouse, 3th Edition”, John Wiley, 2002

[3]. Watson, H., Haley, B.: Managerial Considerations. In Communications of the ACM, Vol. 41, No. 9 (1998)

[4]. List, Beate, et al. "A comparison of data warehouse development methodologies case study of the process warehouse." International Conference on Database and Expert Systems Applications. Springer, Berlin, Heidelberg, 2002.

[5]. Widom, J. Research Problems in Data Warehousing, in Proc. 4th Int. Conf. ‘on Information and Knowledge Management, 1995.

[6]. Kimball, R. The data warehouse toolkit. John Wiley & Sons, 1996.

[7]. McGuff, F. Data modeling for data warehouses. http://members.aol.com/fmcgufYdwmodel/dwmodel.htm, 1996

[8]. Inmon, W. H.: Building the Data Warehouse. Wiley & Sons (1996)

[9]. Golfarelli, M., Maio, D., Rizzi, S.: Conceptual Design of Data Warehouses from E/R Schemes. In: Proceedings of the 31st HICSS, IEEE Press (1998)

[10]. Boehnlein, M., Ulbrich vom Ende, A.: Business Process Oriented Development of Data Warehouse Structures. In: Proceedings of Data Warehousing 2000, Physica Verlag (2000)

[11]. Westerman, P.: Data Warehousing using the Wal-Mart Model, Morgan Kaufmann (2001)

[12]. Poe, V.: Building a Data Warehouse for Decision Support. Prentice Hall (1996)

Cite this article as :

Pooja D. Kavishwar, Dr. S. R. Pande, "Hybrid Data Warehouse Development Method", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 5, pp. 09-15, September-October 2021. Available at doi : https://doi.org/10.32628/CSEIT21752
Journal URL : https://ijsrcseit.com/CSEIT21752