A Review on Different Datamining Schemas with their Design Methodologies

V. Sumalatha
Assistant Professor, Malla Reddy Engineering College,
MRITS, Computer Science Engineering Department, Hyderabad, Telangana State , India.

Abstract: *Database design for data warehouses is based on the notion of the snowflake schema and its important special case, the star schema. The snowflake schema represents a dimensional model which is composed of a central fact table and a set of constituent dimension tables which can be further broken up into sub dimension tables. We formalise the concept of a snowflake schema in terms of an acyclic database schema whose join tree satisfies certain structural properties. We then define a normal form for snowflake schemas which captures its intuitive meaning with respect to a set of functional and inclusion dependencies. We show that snowflake schemas in this normal form are independent as well as separable when the relation schemas are pairing wise incomparable. This implies that relations in the data warehouse can be updated independently of each other as long as referential integrity is maintained. In addition, we show that a data warehouse in snowflake normal form can be queried by joining the relation over the fact table with the relations over its dimension and sub dimension tables. We also examine an information-theoretic interpretation of the snowflake schema and show that the redundancy of the primary key of the fact table is zero.*

Date of Submission: 10-10-2017 Date of acceptance: 27-10-2017

I. Introduction

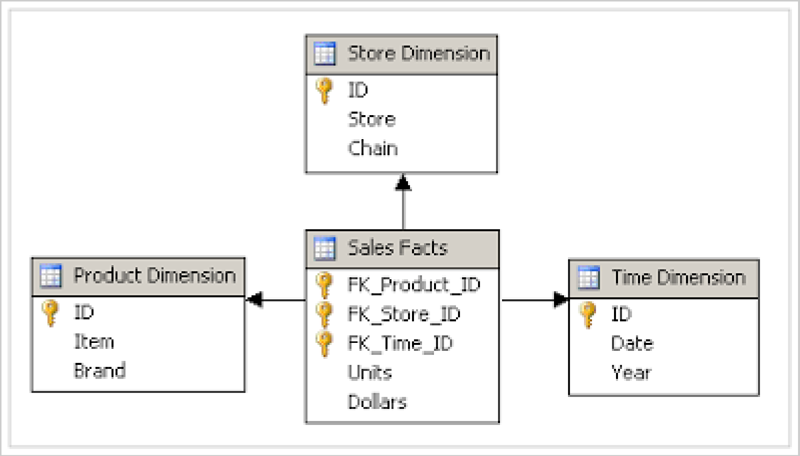
The main approaches to data warehouse design are the data-driven and requirement driven methodologies. Each of them presents advantages and weaknesses [1]. The data driven approach analyzes the data source and remodels it in order to obtain a multidimensional schema. In this way, the feasibility of the data warehouse is guaranteed, but the user needs are not taken into account, going towards a possible failure. On the other hand, the requirement-driven approach considers the business goals to start with, and then produces a multidimensional schema[3]. So, that schema is adherent to user needs but it may be not supported by the effective presence of data in the source. To overcome the limits, several efforts are currently spent to define a design methodology that integrates the advantages of both these approaches. This research issue has led to the definition of hybrid methodologies for data warehouse design [2]. Hybrid methodologies are getting increasing attention because they allow the designer to obtain multidimensional schemata able to satisfy user requirements on the basis of data effectively available in data sources[4]. As a counterpart, hybrid methodologies require a more complex design process due to the reconciliation of different approaches. Indeed, some methodologies have to consider simultaneously the data sources and the user requirements [3,4], while other methodologies have to integrate the data driven approach and the requirement-driven one [5–10]. However, the advantages of adopting hybrid methodologies justify the higher efforts to be spent in the multidimensional modeling. For these reasons, the current research is devoted to introduce automatisms in order to reduce the design efforts and to support the designer in the multidimensional modeling. Automatic methodologies provide algorithms for supporting the designer in (part of) the multidimensional modeling, as to identify facts in data sources [11] and to construct multidimensional views of data [12,13]
Types of schemas

There are four types of schemas are available in data warehouse. Out of which the star schema is mostly used in the data warehouse designs. The second mostly used data warehouse schema is snow flake schema.

Different types of Schemas are as listed below:
1. Star schema
2. Snow flake schema
3. Galaxy schema and
4. Fact constellation schema

DOI: 10.9790/0661-1905050508

Star Schema:
A star schema is the one in which a central fact table is surrounded by demoralized dimensional tables. A star schema can be simple or complex[6]. A simple star schema consists of one fact table where as a complex star schema have more than one fact table.



Snow Flake Schema:
A snow flake schema is an enhancement of star schema by adding additional dimensions. Snow flake schema is useful when there are low cardinality attributes in the dimensions[7].

Galaxy Schema:
Galaxy schema contains many fact tables with some common dimensions (conformed dimensions). This schema is a combination of many data marts.

Fact Constellation Schema: The dimensions in this schema are segregated into independent dimensions based on the levels of hierarchy. For example, if geography has five levels of hierarchy like tertiary, region, country, state and city[8]. Constellation schema would have five dimensions instead of one. Sometimes, both the galaxy schema and fact constellation schema are considered as one and the same. The difference between them in practical applications galaxy schema is collection of fact constellation schemas.

Cost-benefit analysis of data warehouse design methodologies

Methodologies for data warehouse design are increasing more and more in last year’s, and each of them proposes a different point of view. Among all the methodologies present in literature, the promising ones are the hybrid methodologies—because they represent the only way to ensure a multidimensional schema to be both consistent with data sources and adherent to user business goals—and those able to support the designer by providing some kind of automation[9]. However, the results obtainable by the methodologies can differ substantially in terms of schema quality and required efforts. In this paper, we provide metrics for evaluating the quality of multidimensional schemata in reference to the effort spent in the design process and the automation degree of the methodology. As a case study, we apply our evaluation to the major emerging hybrid methodologies for data warehouse schema design.

Building a secure star schema in data warehouses by an extension of the relational package from CWM

Data Warehouses (DWs) are widely accepted as the core of current decision support systems. Therefore, it is vital to incorporate security requirements from the early stages of the DWs projects and enforce them in the further design phases. Very few approaches specify security and audit measures in the conceptual modeling of DWs. Furthermore, these security measures are specified in the final implementation on top of commercial systems as there is not a standard relational representation of security measures for DWs (i.e. the well-known star schema does not allow us to specify security and audit measures on its multidimensional representation of data; instead, they must be specified on top of the implemented relational tables). On the other hand, the Common Warehouse Metamodel (CWM) has been accepted as the standard for the exchange and the interoperability of metadata. Nevertheless, it does not allow us to specify security measures for DWs. In this paper, we make use of the own extension mechanisms provided by the CWM to extend the relational package in order to build a star schema that represents the security and audit rules captured during the conceptual modeling phase of DWs. Finally, in order to show the benefits of our extension, we apply it to a case study related to the management of the pharmacy consortium business.

An empirical study on comparing the understandability of alternative data warehouse schemas

An easily understood data warehouse model enables users to better identify and retrieve its data. It also makes it easier for users to suggest changes to its structure and content. Through an exploratory, empirical study, we compared the understandability of the star and traditional relational schemas. The results of our experiment contradict previous findings and show schema type did not lead to significant performance differences for a content identification task. Further, the relational schema actually led to slightly better results for a schema augmentation task. We discuss the implications of these findings for data warehouse design and future research. An experiment compared the understandability of the star schema to the relational schema. The findings contradict the conventional wisdom that the star schema is easier to understand. Subjects identified content in a relational schema as effectively as those given a star schema. Subjects given the relational schema performed slightly better on a schema augmentation task.

II. Conclusion

With the continuous growth in the amount of data, data storage systems have come a long way from flat files systems to RDBMS, Data Warehousing (DW) and Distributed Data Warehousing systems. This paper proposes a new distributed data warehouse model. The model is built on a novel approach, for the intelligent distribution of data warehouse. Overall the model is named as Intelligent and Distributed Data Warehouse (IDDW). The proposed model has N-levels and is based on top-down hierarchical design approach of building distributed data warehouse. The building process of IDDW starts with the identification of various locations where DW may be built. Initially, a single location is considered at top-most level of IDDW where DW is built. Thereafter, DW at any other location of any level may be built. A method, to transfer concerned data from any upper level DW to concerned lower level DW, is also presented in the paper. The paper also presents IDDW modeling, its architecture based on modeling, the internal organization of IDDW via which all the operations within IDDW are performed.

References

[1] O. Romero, A. Abelló, A survey of multidimensional modeling methodologies, Int. J. Data Warehous. Min. 5 (2009) 1–23.

[2] F. Di Tria, E. Lefons, F. Tangorra, Hybrid methodology for data warehouse conceptual design by UML schemas, Inform. Softw. Technol. 54 (4) (2012) 360–379.

[3] O. Romero, A. Abelló, Automatic validation of requirements to support multidimensional design, Data Knowl. Eng. 69 (2010) 917–942.

[4] F. Di Tria, E. Lefons, F. Tangorra, GrHyMM: a graph-oriented hybrid multidimensional model, in: O. De Troyer, C.B. Medeiros, R. Billen, P. Hallot, A. Simitsis, H. Van Mingroot (Eds.), Proceedings of the ER Workshops, Lecture Notes in Computer Science, Vol. 6999, 2011, Springer-Verlag Berlin, Heidelberg, Germany, pp. 86–97.

[5] J.N. Mazón, J. Trujillo, A hybrid model driven development framework for the multidimensional modeling of data warehouses, SIGMOD Record 38 (2009) 12–17.

[6] C. Phipps, K.C. Davis, Automating data warehouse conceptual schema design and evaluation, in: Laks V.S. Lakshmanan (Ed.), Proceedings of the DMDW: CEUR Workshop on Design and Management of Data Warehouses, 2002, CEUR-WS.org, pp.23–32.

[7] M. Schneider, A general model for the design of data warehouses, Int. J. Prod. Econ. 112 (2008) 309–325.

[8] P. Giorgini, S. Rizzi, M. Garzetti, GRAnD: a goal-oriented approach to requirement analysis in data warehouses, Decis. Support Syst. 45 (2008) 4–21.

[9] A. Bonifati, F. Cattaneo, S. Ceri, A. Fuggetta, S. Paraboschi, Designing data marts for data warehouses, ACM Trans. Softw. Eng. Methodol. 10 (2001) 452–483.

[10] J.N. Mazón, J. Trujillo, J. Lechtenbörger, Reconciling requirementdriven data warehouses with data sources via multidimensional normal forms, Data Knowl. Eng. 63 (2007) 725–751.

[11] A. Carmè, J.N. Mazón, S. Rizzi, A model-driven heuristic approach for detecting multidimensional facts in relational data sources, in: T. Bach Pedersen, M.K. Mohania, A. Min Tjoa (Eds.), DaWaK, Lecture Notes in Computer Science, Vol. 6263, 2010, Springer-Verlag Berlin, Heidelberg, Germany, pp. 13–24. [12] M. Golfarelli, D. Maio, S. Rizzi, The dimensional fact model: a conceptual model for data warehouses, Int. J. Cooperative Inform. Syst. 7 (1998) 215–247.

[12] C. dell'Aquila, F. Di Tria, E. Lefons, F. Tangorra, Dimensional fact model extension via predicate calculus, in: Proceedings of the IEEE 24th International Symposium on Computer and Information Sciences, 2009, IEEE Press, North Cyprus, Turkey, pp. 211–217.

[13] M. Serrano, C. Calero, M. Piattini, Metrics for data warehouse quality, Effect. Databases Text Doc. Manag. (2003) 156–173.

[14] M. Pighin, L. Ieronutti, Quantitative analysis of data warehouse design quality, Int. J. Intell. Def. Support Syst. 3 (1/2) (2010) 52–65.

[15] M. Serrano, J. Trujillo, C. Calero, M. Piattini, Metrics for data warehouse conceptual models understandability, Inform. Softw. Technol. 49 (8) (2007) 851–870.

[16] H. Zuse, A Framework of Software Measurement, Walter de Gruyter, Berlin, Germany, 1998.

[17] M. Serrano, C. Calero, M. Piattini, Validating metrics for data warehouses, IEE Proc. – Softw. 149 (5) (2002) 161–166.

[18] S. Kesh, Evaluating the quality of entity relationship models, Inform. Softw. Technol. 37 (12) (1995) 681–689.

[19] D.L. Moody, Metrics for evaluating the quality of entity relationship models, in: T.W. Ling, S. Ram, M.L. Lee (Eds.), Proceedings of the 17th Table 10 Design effort metrics.

[20] B. Dhanalaxmi, G. Apparao Naidu, and K. Anuradha, "Defect Classification using Relational Association Rule Mining based on Fuzzy Classifier along with Modified Articial Bee Colony Algorithm," *Indian Journal of Applied Engineering Research*, Vol. 12, Number 11,June 2017, pp 2879-2886.

[21] B. Dhanalaxmi, G. Apparao Naidu, and K. Anuradha, "A Rule Based Prediction Method for Defect Detection in Software System," *Journal of Theoretical and Applied Information Technology*, Vol. 95, Number 14, 31st July 2017, pp 3403-3412.

[22] B. Dhanalaxmi, G. Apparao Naidu, and K. Anuradha, “A Fault Prediction Approach based on the Probabilistic Model for Improvising Software Inspection,” *Indian Journal of Science and Technology*, Vol. 9, Issue 45, December 2016.

[23] B. Dhanalaxmi, G. Apparao Naidu, and K. Anuradha, “Practical Guidelines to Improve Defect Prediction Model – A Review”, *International Journal of Engineering Science Invention*, Vol. 5, Issue 9, pp. 57-61, September 2016.

[24] Dhanalaxmi, G. Apparao Naidu, and K. Anuradha, “A Review on Software Fault Detection and Prevention Mechanism in Software Development Activities,” *Journal of Computer Engineering*, Vol. 17, Issue 6, pp. 25 - 30, Nov – Dec. 2015.

[25] B. Dhanalaxmi, G. Apparao Naidu, and K. Anuradha, “A Survey on Design and Analysis of Robust IOT Architectute”, *International Conference on Innovative Mechanisms for Industry Applications*, 13th July 2017, pp 375-378, IEEE

[26] B. Dhanalaxmi, G. Apparao Naidu, and K. Anuradha, “Adaptive PSO based Association Rule Mining Technique for Software Defect Classification using ANN”, Procedia Computer Science Vol. 46,2015, pp 432-442.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

V. Sumalatha. “A Review on Different Datamining Schemas with their Design Methodologies.” IOSR Journal of Computer Engineering (IOSR-JCE) , vol. 19, no. 5, 2017, pp. 05–08.