A Survey on Data Warehouse Architecture

Rajiv Senapati[1], D.Anil Kumar[2]

[1]Assistant Professor, Department of IT, G.I.E.T, Gunupur, India

[2]Associate Professor, Department of CSE, G.I.E.T, Gunupur, India

ABSTRACT: This paper presents the data warehouse architecture for one-level, two-level, three-level classical, and novel three-level architecture with their critiques and the potential areas for future work. The architecture proposed so far exhibits some drawbacks when applied to work over large number of heterogeneous data sources. This paper addresses the issues of classical architecture of data warehouse as well as it presents the architecture for next generation typical data warehouse.

KEYWORDS: Data warehouse; BDW; Scheme integration

I.INTRODUCTION

Data warehouse is a repository of data and related information whose purpose is to allow the extraction, reconciliation and the re-organization of data stored in traditional operational databases. Bill Inmon [1] stated that "A source of data that is subject-oriented, integrated, non volatile and time-variant for the purpose of management's decision process is called data warehouse". A data warehouse is obtained by defining its architecture and mapping by which source data are extracted, filtered, integrated and stored into the defined structure. The review of the available literature on DW has identified three architectures i.e. one-level architecture, two-level architecture, three-level architecture [2], [3, 4], [5].

II.LITERATURE REVIEW

In one-level architecture [2] each piece of data is stored once. A middleware level operates as an interface between users and operational databases. This kind of architecture allows quick development of warehouse with reduced cost.In two-level architecture [2, 3, 4] source data has been separated from derived data where first level of this architecture stores source data and second level stores derived data. This architecture is convenient especially when operational sources are homogeneous.

Three-level architecture [5] comprises both operational and reconciled, derived data. Here derivation of data for decision support is performed in two phases: the reconciliation of operational data belonging to different sources, the derivation of decision support data from reconciled one. In this architecture the first level stores operational data, the second level stores reconciled data and the third level stores support decision data. Therefore a three level architecture is represented as: DW = < OS, BDW, BIWS, M > .Here the Reconciling activity eliminates inconsistencies and irregularities of operational data and integrates them in a rational way, once it is constructed then decision support data can be derived.

The proposed novel three-level architecture overcomes the drawbacks of classical three level architecture. This approach does not directly integrate operational schemes to construct a global flat scheme. This model first collects subsets of operational schemes into homogeneous clusters and then integrates schemes on a cluster by cluster manner. Each integrated scheme obtained is then abstracted to construct a global scheme representing the cluster. This process is an iterative process, the process will continue until one scheme is left. In this way a hierarchical structure is obtained, which is called data repository.

By adopting such a data repository as a global scheme of the reconciled level, we obtained a data warehouse architecture, which is represented in figure 1.
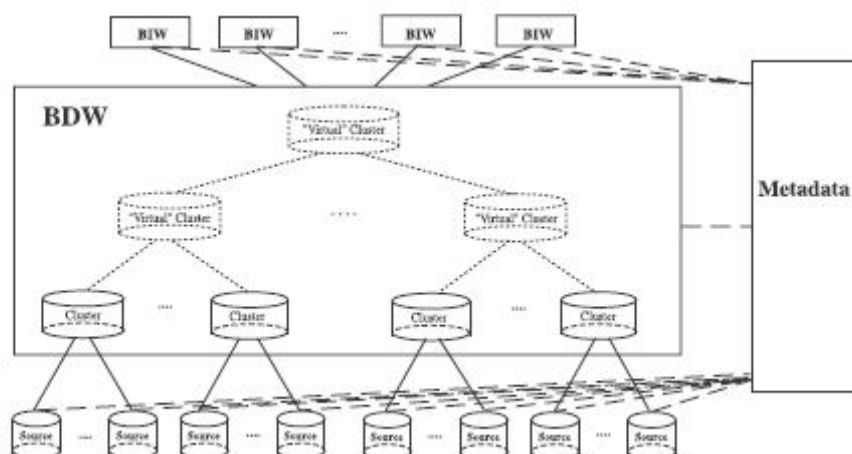

Fig. 1: Novel three-level architecture

There are some improvements in the proposed approach which overcomes the drawbacks of classical three level architecture, these are:

1. In classical three- level architecture source operational databases are all grouped into one cluster and no abstraction is carried out over the associated C-Scheme.

2. The proposed architecture constructs an incremental data warehouse where while adding, deleting or modifying a source operational database scheme causes only the scheme of the clusters involving this database to be updated rather than causing the modification of the entire reconciled global scheme.

3. It is very difficult to analyze the data in classical architecture which is designed over a large number of source databases .Whereas in proposed model database cluster would corresponds to each of these sub schemes and therefore would have been level wise associated to a precise semantics.

4. In this approach metadata are obtained during the derivation of reconciled data, which automatically maintains the consistency among reconciled data and metadata.

Inmon's and Kimball's approach

Now the companies require an enterprise wide data repository to support a variety of analytical applications, the data warehouse began to emerge and resulted in two competing architectures, i.e. EDW and data mart bus architecture. Inmon supports a top-down development approach that adopts traditional relational database tools to the development need of EDW. From this enterprise wide data store, individual departmental databases are develop to serve most decision support

needs [6].Kimball proposes a bottom-up approach that employs dimensional modeling [7,8].A comparison of essential features of Kimball's and Inmon's approach is represented in table1.

Table 1: Comparison of essential features of Inmon's and Kimball's ideology

	Kimball's Approach	Inmon's Approach
1	Everyone is allowed to fabricate their database according to their requirements and departmental structure. All these independent repositories can be integrated as and when required. This method is known as bottom up approach.	Supports top-down approach. Here no one is allowed to develop any database independently. The database for an organization should be planed and designed centrally. Every department within the organization will follow the central design schema to fabricate their database.
2	This structure is easier to build.	The structure proposed is very typical.
3	It is a nimble approach.	Rigorous analysis and design is required.
4	Problematic to maintain as an enterprise resource.	Easier to maintain as an enterprise resource.
5	Data is often redundant.	Redundancy is regulated to a great extent.
6	Very difficult to integrate independent data marts with varying structure.	Integration of data mart is comparatively easier.
7	This approach is flexible.	This approach is comparatively rigid.

Typical Data Warehouse Architecture

[9] Present the typical data warehouse architecture, shows the components needed in a data warehouse system, while [10] also propose a similar architecture. Typical data warehouse architecture is shown in figure.2.
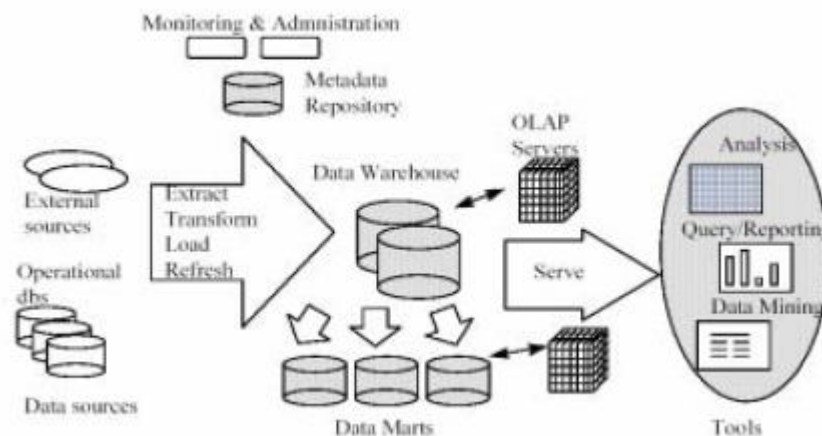

Fig. 2: Typical Data Warehouse Architecture

The typical data warehouse architecture includes:

➢ software tools for extracting data from multiple operational databases and external data sources; for cleaning, transforming and integrating this data; for loading data into the data warehouses; and periodically refreshing the data warehouse to reflect the updates at the source.

➢ Here the main data warehouse consists of several data marts.

➢ OLAP server present multidimensional views of data to a variety of front end tools like query tool, report writer, analysis tools and data mining tools.

➢ Repository for storing and managing metadata.

➢ Tools for monitoring and administrating the warehouse system.

III.CRITIQUE OF THE LITERATURE

The one-level architecture induces the planning activities such as data source identification, data integration, transformation etc. to be carried out for each query which leads to the unpredictable access time for the end user.

In two-level architecture each decision support application has its own derived data; there is no possibility of storing a single copy of information for all decision support application, which indicates data duplication.

In three-level architecture the reconciled activity eliminates inconsistencies and irregularities of operational data but again data replication found in this level. It has some limitations with complex heterogeneous operational source systems. In this architecture metadata are derived and updated separately from operational, reconciled and derived data, therefore they are not well related to corresponding schemes.

The proposed novel there level architecture eliminates all the limitations presented in the classical architecture. This architecture will produce a semi automatic data warehouse. This architecture may not sufficient to deal with unstructured data streams.

IV.CONCLUSION AND FUTURE WORK

The novel three-level architecture proposed in this paper eliminates all the drawbacks of the classical one which can handle structured data. In future the architecture can be extended to deal with big data which can able to process structured as well as unstructured data. In such situation a distributed file system (like Hadoop) sits between source data systems and the data warehouse. It collects, aggregates, and processes huge volumes of unstructured data, and stages it for loading into the data warehouse. Structured and unstructured data from back end systems can be brought into the data warehouse in real- and near-real time.

REFERENCES

1.   W.H. Inmon, DW 2.0 Architecture for the Next Generation of Data Warehousing, DM Review, Vol. 16 Issue 4, 8-25, Apr 2006.
2.   W.H. Inmon, What is a data warehouse? , Prism Tech.Topic 1 (1) ,1997.
3.   K. Orr, Understanding data warehousing, Am. Program. Vol.8 (5), 2–7, 1995.

4.   J. Widom, Research problems in data warehousing, in:Proceedings of International Conference on Information and Knowledge Management (CIKM'95), ACM Press,Baltimore, Maryland, USA, pp. 25–30,1995.
5.   IBM, Information Warehouse Architecture I, IBM Corporation, 1993.
6.   M. Breslin, Data warehousing battle of giants: comparing the basics of the Kimball and Inmon models, Business Intelligence Journal 9 (4) , 2004.
7.   Mawilmada, Pubudika Kumari, Impact of a data warehouse model for improved decision-making process in healthcare. Masters by Research thesis, Queensland University of Technology, October 2011.
8.   W.H. Inmon, Building the Data Warehouse, Third Edition, New York: John Wiley & Sons, 2002.