ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 9, March 2013

Distributed Data Warehouse Architecture: An Efficient Priority Allocation Mechanism for Query Formulation

Nouman MaqboolRao, Muheet Ahmed Butt, Majid Zaman, Waseem Jeelani Bakshi

Abstract*—Data warehouse is the repository where data for an enterprise is stored using a centralized approach. A data warehouse environment includes an extraction, transportation, transformation and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users. The category of users which access this warehouse is differentiated with respect to priority level. This paper discusses the priority of accessing the warehouse which is assigned to various set of users on the basis of the stature in the enterprise. An efficient priority balancing mechanism has been proposed in this research so that the data retrieved from the warehouse depends on the stature of the users.

***Index Terms*—Query Management, Data Warehouse, Data Mart, OLAP**

I. INTRODUCTION

Data warehouses are used as an enterprise repository to upkeep top management in business decision making in an efficient manner [1][2]. Currently data warehousing systems are used as decision support systems to help enterprises in strategic and intelligent business decision making. A data warehouse is a logical collection of information gathered from many different operational data sources used to create business intelligence that supports business analysis activities and decision making tasks. It is used for providing the basic infrastructure for decision making by extracting, transforming, cleansing and loading huge amount of data of the enterprise. This classic definition of the data warehouse focuses on data storage. However, the means to retrieve and analyse data, to extract, transform and load data and to manage dictionary data are also considered essential components of a data warehousing system. Data warehouses support business decisions by gathering, combining and shaping of data for reporting and future analysis with tools such as online analytical processing (OLAP) and data mining. The size of data warehouse fluctuates from hundreds of gigabytes to terabytes. Different scans, joins and aggregates are performed while querying the data warehouse. The queries on data warehouse are ad hoc and multiple faced. Throughput of query determines the success of data warehousing project. The query response time is also important factor in data warehouse success. The allocation of facts and dimensions in a certain schema also effect query success. The following are the research questions.

- ✓ How to allocate data over different statures of hierarchy in an enterprise?
- ✓ How to handle numerous users based on their priority ranks in an enterprise?
- ✓ How to manage queries in an efficient manner so as to reduce the response time for data reporting and analysis?

This paper proposes and discusses distributed data warehouse architecture with Efficient Priority Allocation Mechanism Layer EPAML. In this architecture data that is coming from various operational data sources is distributed over different levels of hierarchy in an enterprise. The main focus is better management of an organizational valuable data so that queries take less time to execute. Load operation will be divided on separate individual systems for higher reliability as data is being distributed on various systems. The rest of the paper is organized into different sections. Section 2 gives an overview of data warehouse. Section 3 presents the proposed distributed data warehouse architecture for implementation of EPAML. Section 4 presents case study of University of Kashmir Data Warehouse UOKDW and section 5 provides the conclusion for the proposed research.

II. LITERATURE REVIEW

The concept of data warehousing dates back to the late 1980s when IBM researchers Barry Devlin and Paul Murphy developed the "business data warehouse". In essence the data warehousing concept was intended to provide an architectural model for the flow of data from operational systems to decision support environments. The concept attempted to address the various problems associated with this flow - mainly the high costs associated with it. In the absence of a data warehousing architecture an enormous amount of redundancy of information was required to support the multiple decision support environment that usually existed. In larger corporations it was typical for multiple decision support environments to operate independently. Each environment served different users but often required much of the same data. Data warehouse is centralized data repository maintained separately from organization’s operational databases to help organization in corporate decision making process. William Inmon has described data warehouse as “A subject-oriented, integrated, time-variant, non-volatile collection of data in support of management decisions“. Data warehouse is a set of materialized views over data sources [3], [4], [5]. Ralph Kimball *et al* defined “A data warehouse is a copy of transaction data specially structured for query and analysis" [6]. "A data warehouse combines various data sources into a single source for end user access. End user can perform ad hoc querying, reporting, analysis, data mining and visualization of warehouse information. The goal of data warehouse is to establish a data repository that makes operational data accessible in a form that is readily acceptable for decision support and other application" [7]. Basic architecture of data warehouse is discussed in [8] Connolly *et al* proposed three tier architecture of a data warehouse [1]. First tier consists of data warehouse and archive/backup data. Second tier consists of different data marts. Reporting, OLAP and data mining tools make third tier of the architecture. Hoffer *et al.* presented generic two level architecture of data warehouse [8]. Detailed elements of a data warehouse presented in [5] are shown in Figure 1. Some data warehouse architectures are also discussed in [10], [11].

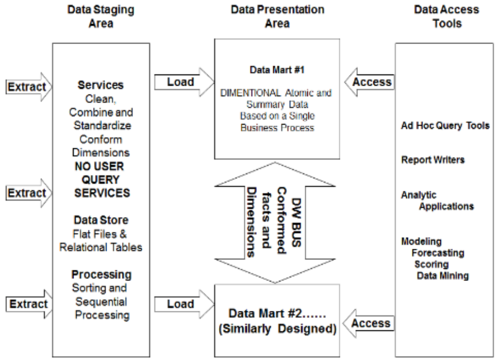

Fig 1: Data Warehouse Elements

III. PROPOSED ARCHITECTURE

The proposed architecture is shown in Figure 2. The proposed architecture uses a decentralized approach. The data warehouse is an academic one and source of the warehouse has been taken from University of Kashmir Data Warehouse. The users from multiple directions are sending set of query requests to Query Acquisition Layer (QAL). It is mentioned that the users are accessing data warehouse for reporting and analysis operations are distributed in hierarchal order. Users that are interacting with warehouse are classified according to defined distribution of classes. The top level management of the users are placed on upper level of the hierarchy and the subordinates are classified at lower level. The user interacts with the system through presentation layer. The priority level of user is already defined in the priority database. The Efficient Priority Allocation Mechanism Layer (EPAML) has three major components that are data buffer, priority allocation module and prioritized queue. The set of queries generated by users are denoted with Q. The Q is send to Priority Allocation module. Here Q is filtered on the basis of threshold value. The threshold is a predefined value which is assigned to every user and is stored in a priority database and can be optimized accordingly depending upon the change in the environment. Threshold holds the lowest possible query processing time. It is assumed that query processing time is already calculated on the basis of their complexity. The complexity might be calculated on the basis of joins, set, dimensions and access time. The queries in Q that are less than or equal to Threshold value are filtered out accordingly. Q_{Th} is the set of filtered queries. The Q_{Th} is processed first and replied back to requested user. Following is the equation to calculate set of remaining queries Qi.

$Qi = Q - Q_{Th}$

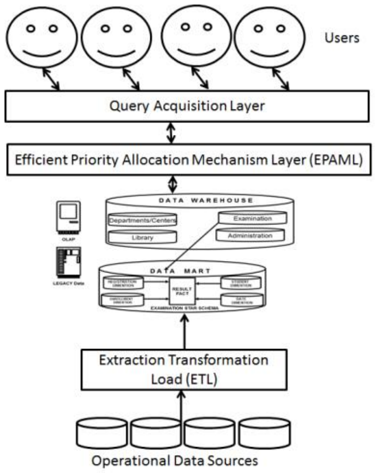

Fig 2 : Proposed Architecture of Distributed Data Warehouse

Where Q_{Th} is the set of queries less than Threshold value and Q is the set of all unordered queries. The Qi is sorted in two steps. In first step, Qi is sorted on the basis of user's priority in order to get Q_{ij} where j is priority number. In second step, sorting is applied on the set Q_{ij}individually (i.e., each set of priority is sorted separately) on the basis of processing/time. Time T_{ij} is defined as the time slot assigned to set Q_{ij} in order to interact with DWH in predefined time. The results are generated to requested users accordingly. The Algorithm of Priority Allocation Process is given below:

IV. PROPOSED ALGORITHM

Input: Q_{Th}set of unsorted queries less than threshold value, Q*i*, set of queries after subtracting $Q_{Th}$from Q and T_j, set of time slots corresponding to each priority level.
Output: Q_o, queue of queries to be executed.

Algorithm for EPAML implementation

Begin

```
        $Q_o$=0
        For each requested query in $Q_{Th}$
                $Q_o$ = Sort $Q_{\mathrm{Th}}$ on the basis of time complexity
        End For

        For each requested query in $Q_i$
                $Q_{ij}$ = Sort Qi on the basis various priority values
        End For

        For each priority level j
                $Q_{Tmp}$ = Sort $Q_{ij}$ on the basis of Time Complexity

                $QTmp = Tij$
```

```
            Qo =Qo U QTmp
        End For
End
```

Where Q be the set of unorganized queries with priority levels, Q_{Th} is the set of unsorted queries less than threshold value and Qi is the set of remaining queries.

V. CONCLUSION

Currently data warehouse is used as organizational repository to support business decision making. Mostly the data warehouse systems use centralized approach. Furthermore, the hierarchy of organization and classes of users is not considered in data warehousing systems. In this paper a distributed architecture of data warehouse with efficient priority allocation mechanism layer (EPAML) is introduced. In proposed architecture, users are assigned with various priorities based on their stature. The users with higher priority level are dealt with first so that they may not suffer long wait. Low complexity of query is also preferred based on the priority database. The set of users that belong to different hierarchal background are treated accordingly. Also in future work we will implement this work by going through case study and further enhance on different levels.

REFERENCES

[1] EMA Butt, SMK Quadri, EM Zaman, “Star Schema Implementation for Automation of Examination Records”, WORLDCOMP, 2012 FEC3568, Las Vegas USA.

[2] MA Butt, SMK Quadri, M Zaman,” Data Warehouse Implementation of Examination Databases”, International Journal of Computer Applications 44 (5), 18-23.

[3] Z. Bellahsene, Schema, “Evolution in Data Warehouses”, Knowledge and Information Systems, Springer-Verlag, pp 283-304, 2002.

[4] E.A. Rundensteiner, A. Koeller, X. Zhang, “Maintaining Data Warehouses over Changing Information Sources”, Communications of the ACM, Volume, 43, New York, NY, USA, pp 57-62, 2000.

[5] Ralph Kimball, M. Joy and T. Warren, the Data warehouse Toolkit: with SQL server and Microsoft Business Intelligence Toolset, 2nd Edition, New York: Wiley publisher. Inc., 2006.

[6] Efraim Turban, Jay E. Aronson and NarasimhaBolloju, Decision Support Systems and Intelligent Systems, 7th edition, Prentice Hall College Div, 2001 .

[7] Jeffrey A. Hoffer, Mary B. Prescott, Fred R. McFadden, Modern database management, Sixth Edition, Pearson Education Publishers, Singapore.

[8] Online Analytical Processing (OLAP) and Data Warehousing, academic2.bellevue.edu/~jwright/CIS605/Lesson10/OLAP.ppt Accessed Data: Dec 5, 2008.

[9] Daniel L. Moody, Mark A.R. Kortink, “From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design”, In Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000) June 5-6, 2000, Stockholm, Sweden .

[10] Mohammad Rifaie, Erwin J. Blas, AbdelRahman M. Muhsen, Terrance T. H. Mok, KeivanKianmehr, RedaAlhajj, Mick J. Ridley, “Data warehouse Architecture for GIS Applications”, In Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services (iiWAS '08) , November 2008, Linz, Austria.