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

Nouman Maqbool Rao, 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
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.

\[ Q_i = Q - Q_{th} \]

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\(_j\) where j is priority number. In second step, sorting is applied on the set Q\(_j\) individually (i.e., each set of priority is sorted separately) on the basis of processing/time. Time T\(_j\) is defined as the time slot assigned to set Q\(_j\) 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:

Algorithm for EPAML implementation

Begin

Q\(_o\)=0

For each requested query in Q\(_{th}\)

\[ Q_i = \text{Sort } Q_{th} \text{ on the basis of time complexity} \]

End For

For each requested query in Q\(_j\)

\[ Q_{oj} = \text{Sort } Q_i \text{ on the basis various priority values} \]

End For

For each priority level j

\[ Q_{tmp} = \text{Sort } Q_{oj} \text{ on the basis of Time Complexity} \]

\[ QTmp = T_{ij} \]

End For

End
\[ Q_{o} = Q_{o} U Q_{tmp} \]

End For

End

Where \( Q \) be the set of unorganized queries with priority levels, \( Q_{2h} \) 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 hierarchical 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


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