A Horizontal Fragmentation Algorithm for the Fact Relation in a Distributed Data Warehouse

Amin Y. Noaman
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada R3T 2N2
+1 204 474 8691
noaman@cs.umanitoba.ca

Ken Barker
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada R3T 2N2
+1 204 474 8669
barker@cs.umanitoba.ca

ABSTRACT

Data warehousing is one of the major research topics of applied-side database investigators. Most of the work to date has focused on building large centralized systems that are integrated repositories founded on pre-existing systems upon which all corporate-wide data are based. Unfortunately, this approach is very expensive and tends to ignore the advantages realized during the past decade in the area of distribution and support for data localization in a geographically dispersed corporate structure. This research investigates building distributed data warehouses with particular emphasis placed on distribution design for the data warehouse environment. The article provides an architectural model for a distributed data warehouse, the formal definition of the relational data model for data warehouse and a methodology for distributed data warehouse design along with a "horizontal" fragmentation algorithm for the fact relation.

Keywords

distributed data warehouse architecture, distributed data warehouse design, horizontal fragmentation.

1 INTRODUCTION

Decision Support Systems (DSSs) and Executive Information Systems (EISs) can only be effective tools if the data used are readily available and represent the integration of all pertinent corporate-wide data. Data warehouses provide this integrated environment by extracting, filtering, and integrating relevant information from all available data sources. Further, as new or additional relevant information becomes available, or the underlying source data are modified by the operational systems, the new data are extracted from its autonomous, distributed and heterogeneous sources into a common model that is integrated with existing warehouse data. Once information is available at the warehouse, queries can be answered and data analysis (DSS and EIS) can be performed.

Most of the work to date has focused on building large centralized systems that are integrated repositories founded on pre-existing systems upon which all corporate-wide data is based. The centralized data warehouse is very expensive and tends to ignore the advantages realized during the past decade in the areas of distribution and support for data localization in a geographically dispersed corporate structure. Further, it would be unwise to enforce a centralized data warehouse when the operational systems exist over a widely distributed geographical area.

The distributed data warehouse supports the decision makers by providing a single view of data even though that data are physically distributed across multiple data warehouses in multiple systems at different branches. Currently, the field of distributed data warehouse in terms of architecture and design is considered an important research problem that needs investigation.

This research contributes to the problem of distributed data warehouse architecture and design by:

1. Extending the preliminary architecture model that has been presented in [8] by proposing a distributed data warehouse system architecture and describing the functionality of its components.
2. Proposing the formal definition of the relational data model for data warehouse where the relational data model represents the underlying model for the different level of schemas of the proposed system architecture.
3. Proposing a methodology for the distributed data warehouse design and a horizontal fragmentation algorithm that partitions the huge fact relation into a set of fragments.

To the best of our knowledge, this is one of the first works to propose a methodology and a horizontal fragmentation algorithm for the distributed data warehouse design.

The reminder of the paper is organized as follows. Section 2 presents our proposal for the distributed data warehouse system architecture and illustrates how the information flows in the distributed data warehouse. Section 3 provides the data model for data warehouse. Section 4 addresses our proposal for the distributed data warehouse design and presents the horizontal fragmentation algorithm for the fact relation. Finally, Section 5 draws conclusions.

2 DISTRIBUTED DATA WAREHOUSE ARCHITECTURE

This section extends the preliminary architecture model that has been presented in [8]. It proposes distributed data warehouse system architecture, and describes the functionality of its components