Egyptian Informatics Journal (2016) 17, 147–159

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

FULL-LENGTH ARTICLE

A novel approach for intelligent distribution of data warehouses

Abhay Kumar Agarwal *, **Neelendra Badal**

Kamla Nehru Institute of Technology, Sultanpur, U.P., India

Received 17 March 2015; revised 23 September 2015; accepted 19 October 2015
Available online 28 November 2015

KEYWORDS

Data warehouse;
Distributed;
Hierarchy;
Intelligent;
Meta Data

Abstract 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.

1. Introduction

The basic definition of DW was given by Inmonin [10] that DW is a subject-oriented, integrated, time-varying and non-volatile collection of data in support of the management's decision-making process. Generally, an organization starts with the centralized DW system. This centralized DW is responsible, for storing the entire data of the organization, answering all the queries and for decision making. For an organization, as the amount of data and number of queries to be answered increases to an extent, the need for the distribution of stored data generates, that in turns develops the need of Distributed DW. The need for Distributed DW is also developed when an organization grows in size with increase in number of its branches or with the increase in the size of its branches. Many approaches were proposed by many researchers for the distribution of data and construction of Distributed DW. One such approach is distribution of data in hierarchal fashion called Hierarchal Distributed DW. For building Hierarchal Distributed DW either of two design approaches, Top-Down or Bottom-Up is used. Bottom-up Approach

* Corresponding author.
E-mail addresses: abhay.knit08@gmail.com (A.K. Agarwal), n_badal@hotmail.com (N. Badal).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2015.10.002

[18,3,7]: is suitable when the objective of the design is to integrate existing database systems. The bottom-up design starts from the individual local conceptual schemas and the objective of the process is integrating local schemas into the global conceptual schema. Top-Down Design [8,17,16]: In the top-down design approach the data warehouse is built first. The data marts are then created from the data warehouse. Top-Down design approach is generally used for a very large system. An example of Top-Down design approach is DW at country level is distributed among the DW's at state level, and each DW at state level is distributed among DW's at city level and so on up to the local mart level. The advantages of doing so are local queries are answered locally while global queries are answered by the system as a whole. The IDDW proposed in this paper uses this very approach for building it.

The steps followed in designing and building IDDW are as follows: (1) Selecting a theme, (2) Framing, a theme in hierarchal structure of N-levels, (3) Identification of N-Levels in hierarchal structure, (4) Identification of locations in each levels of hierarchal structure, (5) Formation of hierarchal structure considering the identified locations, (6) Initial, location and number of DW's to start within the formed hierarchal structure, (7) Data to be stored in each DW's, (8) Incorporating any new DW build at any location, (9) Transfer of the data/hierarchal information, related to a location, in the new DW build at that location and (10) Stores the local data of a location in the DW if available at that location. The detail is presented in Section 3 of the paper.

The other sections in the paper are as follows: Section 2 presents the Related work; Section 3 presents the proposed work that includes the IDDW Modeling; the IDDW architecture; IDDW data model; the internal organization of IDDW which comprises of global MDS, local MDS etc. and with the help of which all the operations in IDDW are executed; the construction of N-level hierarchal structure, IDDW using three algorithms; in Section 4, A case study is taken for IDDW, showing an education system in India in 8 level hierarchal structure; and in Section 5, Experimental setup is developed, based on an 8-level hierarchal structure. A program was written in JAVA with SQL in backend. The real-time data is inputted using a front page in the system and is intelligently stored in the concerned DW in the system, and in Section 6, various observations and analysis are made about IDDW based on experiment performed.

2. Related works

The Distributed DW facilitates the policy and decision makers, by providing a coherent and single view of data. It does so inspite of the fact that data are physically distributed across multiple DW's in multiple systems at different branches. Many authors proposed many techniques for developing Distributed DW.

Inmon in [9] proposed an approach of building the distributed data warehouse. The approach assumes the presence of both local and global data warehouses with data stored in each are mutually exclusive. The local DW includes local data of interest, while, the global DW contains common data across the organization and data integrated from various local DW. Inmon's assumption, about the mutual exclusivity of data between the local and global data warehouses, seems to be impractical. White in [22] proposed an approach which is named as "Two Tier Data Warehouse" and is a combination of centralized data warehouse and a decentralized data marts. Noaman et. al. in [15,14] proposed an architecture for Distributed DW. It uses Top-Down design approach and presents two fundamental issues: fragmentation and allocation of the fragment to various sites. Work proposed by author in [15] is extended work in [14]. They extend it by describing the functionality of distributed DW system architecture components, by giving definition of relational data model for DW and by giving a horizontal fragmentation algorithm. Zhou et. al. in [24] proposed hierarchically distributed data warehouse (HDDW). HDDW integrates the local data marts into a hierarchy. HDDW uses bottom-up design approach for building the warehouse. This method is useless in case of having a large central data warehouse needed to be a distributed data warehouse (top-down approach) because no fragmentation schema has been used. Bernardino et al. in [2] designed a new technique called data warehouse stripping (DWS), which is a round robin data partitioning approach for relational data warehouse. The limitation of DWS technique is that it is not effective in data warehouse with big dimensions. Therefore, the researchers propose a new approach called selective loading to deal with data warehouses with big dimensions in DWS systems. The selective load technique proposed explores the fact that the subset of the fact table rows stored in each node is only related to a small part of the rows of the big dimension and not related to all of them. Thus, the idea is to store in each node only the dimension rows that are related to fact rows stored in that node, and not to replicate the entire dimension [5]. This method faces a problem because the data size will be bigger over time and the round-robin partitioning approach must be applied again to partitioning the fact and dimension table row-by-row through all nodes.

Su et al. in [19] proposed a new model of Radio Frequency Identification (RFID) distributed data warehouse, named as RFID-CHDDW. As the name suggests Concept Hierarchy Distributed Data Warehouse (CHDDW) the constructional approach for making it is hierarchical, bottom-up based on the concept hierarchy, it combines the qualities of both data marts and distributed data warehouse.

Grid based approach for distributed data warehouse is proposed in [20,6,13,21]. The other methods proposed by various authors are [23] used ASM to design distributed data warehouse, [4] proposed a new analytical model using Petri Net for distributed data management in a data warehouse and [12] presented and proposed a data warehouse decentralization strategy based on cost-based fragment allocation and replication algorithm.

The IDDW proposed in this paper uses top-down design approach. However, it has advantages over other Distributed DW using top-down design approach. The advantages that the IDDW provides over others are as follows: first, that there is no need to place a DW at each location, rather a DW can be built anytime at any location as per the need of that location; second, IDDW architecture is such that every user registered in the system (i.e. data of registered user stored in a DW within IDDW) is uniquely identified; third, operations to be performed on the system are not location specific rather it can be executed from anywhere within the system; fourth, the internal organization of the proposed system is such that it makes possible for transfer of data from one DW to another.