Distributed Warehouses: A Review on Design Methods and Recent Trends

Sagar Yeruva
Associate Professor, CSE, VNRVJIET, Technology, Bachupally, Hyderabad, India.

Dr. P. V. Kumar
Principal, Bharat Institute of Engineering & Technology, Ranga Reddy (Dt), A.P, India-501 510

Dr. P. Padmanabham
Director, Academics, Bharat Institute of Engineering & Technology, Ranga Reddy (Dt), A.P, India-501 510.

Abstract: The distributed data warehouse supports the decision makers by providing a single view of data even though that data is physically distributed across multiple data warehouses in multiple systems at different branches. This environment has changed the face of computing and offered quick and precise solutions for a variety of complex problems for different fields. This paper reviews distributed data warehouse systems in view of its appearance compared to centralized data warehouses, framework for distributed warehouse systems, data base designs and a good discussion on recent developments in distributed data warehouse architectures. It also concentrates the latest systems and various optimization methods.

Keywords: Distributed Data Warehouse, Centralized data warehouses, Optimization, schema and data mart.

I. INTRODUCTION

Once the globalization of economic scenarios came into existence, the market environment with many competitors locally and globally the information availability and processing has changed its dimensions. Each enterprise have on-line information about its general business figures as well as detailed information on specific topics to be able to make the right decisions at the right time.

To support this almost all enterprises are trying to build data warehouses. Early adaptors of data warehousing were in the “Think Globally, Act Globally” group. These organizations are willing to make huge investments in technology and take large risks in the hopes of realizing significant competitive advantages or productivity gains. These adopters skewed the perceptions of analysis and some tool companies, who thought that the entire data warehousing market, would think and act like these customers.

As the market has developed, however, it has been revealed that the vast majority of the market is in the “Think Globally, Act Locally” group. Supporting to this kind of trend data warehouses also changes its complex centralized data warehouses to distributed data warehouses which are integrated in a common conceptual schema.

Distributed Warehouses:
The centralized data warehouse or simply Data Warehouses designed for serving many different user groups has the following features:

- Provides data that is common across the organization and of interest to the entire organization on one central location.
- Provides consistent data, so that decision makers are referencing the same data when they are making decisions.
- Protects the operational systems from complex queries that could slow down the performance of those systems.

But this kind of environment which takes a long time to setup and very expensive for its maintenance and has the limitations like:

- Performance: Many users must compete to access the data, resulting in delays caused by queuing requests. In addition, if the users are geographically distributed, delays also may occur due to transmitting access requests and responses to and from the central data warehouse location.
- Expandability: Expansion is expensive in a centralized data warehouse. For example, if the volume of the data exceeds the capacity of the warehouse, or the amount of processing against the data has increased, then the
organization must replace the existing centralized data warehouse with a large one. This is costly and sometimes may not be possible within organizations constraints.

- Reliability: The centralized data warehouse is the single point of failure. When the whole system goes down, it usually takes some time to bring it back.

- Cost: Experience over the past decade in the computing environment has shown that it is cheaper to have a number computers linked together than a centralized machine.

- Vendor dependency: This is an unfortunate consequence of the centralized approach. Typically purchasing a single centralized system limits the future expansion.

Therefore, many companies today decide to start with smaller, flexible data marts dedicated to specific business areas. To get the cross-functional analysis, there are two possibilities [8, 9, 10]. The first one is to create again a centralized data warehouse for only cross-functional summary data. The other one is to integrate the data marts into a common conceptual schema and therefore create a distributed data warehouse.

"Information Warehouse" a strategic approach helps mainly to understand the successful execution of enterprise initiatives [15]. By targeting building the information warehouse we can build an efficient and effective tool for the entire organization where various stakeholders feel comfortable and user friendly to solve their tasks. To present the review of various architecture and design methods in this, the presentation is organized as follows. First begins with the frame work for distributed warehouses. Next various design methods for the construction and finally end with various data models and existing approaches were explained.

II. FRAMEWORK FOR DISTRIBUTED DATA WAREHOUSES

Two approaches to build the distributed data warehouses are available which are described as follows:

Inmon’s Approach

This approach in Fig-1 assumes the existence of both local and global data warehouses with data stored in each being mutually exclusive. The local data warehouse contains the data of interest to the local site and includes historical data in addition to local decision making functions.

The global data warehouse contains data common across the corporation and data integrated from the various local staging areas for inclusion into the central location. This is accomplished by having each local site stage warehouse data before passing it to the central global data warehouse which provides the global DSS (Decision Support System) functionality for corporate-wide queries. This approach assumes that data found in any local data warehouse are not stored in the global data warehouse and vice versa thereby guaranteeing no redundancy between them. Inmon’s assumption, about the mutual exclusivity of data between the local and global data warehouses, seems to be impractical.

![Fig-1: Inmon’s Approach to Distributed Data Warehouse](image-url)
White’s Approach

This is known as a “Two-Tier Data Warehouse”, which is the combination of both centralized data warehouses and a decentralized data mart. White’s central data warehouse contains normalized detailed data captured and cleaned from operational systems at user-defined intervals. The central data warehouse maintains data collections that consist of data derived from the detailed base data. Data collections are the user view of warehouse data and may contain denormalized detailed data as well as summarized data.

Fig-2: White’s Approach to Distributed Data Warehouses.

A data distribution service is provided by the data warehouse to distribute data collections to decentralized data marts at the various branches or sites of the corporation. The data marts are subsequently distributed to the other sites of the corporation. Data marts permit DSS processing on local systems, which improves both performance and availability.

III. DISTRIBUTED DATABASE DESIGNS

In the literature of distributed data environment, where two approaches for distributed database design were introduced: the top-down approach and the bottom-up approach. The top-down approach is used when the databases are non-existent. However, once the databases exist (for example, the multidatabase environment), the bottom-up design is the appropriate approach.

In the top-down approach, the steps of the design process are [12]:

- The requirement analysis and logical design of the global data bases. The output of these two steps is the Global Conceptual Schema (GCS) and the access pattern information.
- These two outputs represent the input for the distributed design step. The objective of the distributed design is to device the Local Conceptual Schema (LCS) by generating entries called fragments.
- These fragments are then allocated to the distributed sites.
- The physical design is the last step in the top-down design approach process. It maps the LCS with the access information to the distributed physical storage devices.

Where as in bottom-up approach, the design consists of [3]

- Selecting a common database model for describing the global schema of the existing databases.
- Translating each local schema into the common data model.
- Integrating local schemas form the existing databases into the global conceptual schema.

IV. RECENT DEVELOPMENTS IN DISTRIBUTED DATA WAREHOUSES ENVIRONMENT
Many organizations have physically distributed databases with extremely large amounts of data. Traditionally the data warehouse would be seen as a centralized repository, whereby data from all sources would be imported into that large centralized repository for analysis. Nowadays the speed and bandwidth of wide-area computer networks enables a distributed approach, whereby parts of the data may reside in different places, parts being cached and/or replicated for performance reasons, and the system functions to the outside world as a single global access-transparent repository. As the amount of data and number of sites grow, this distributed approach becomes crucial, as a single centralized data warehouse importing data from all the sources has obvious scalability limitations.

Distributed Data Warehouse is a young discipline related to distributed computing which will place its impact in the future computing especially for DSS. A methodology for distributed data warehouse design along with two approaches to horizontally fragment the huge fact relation in the data warehouse and an analytical technique that provides insight into which approach presented is most suited to a particular distributed data warehouse environment [2]. Query processing in a distributed data warehouse, consists of local data warehouses at each collection point and a coordinator site, with most of the processing being performed at the local sites through Skalla system [1]. Skalla translates OLAP queries, specified as certain algebraic expressions, into distributed evaluation plans which are shipped to individual sites, and the approach operates in a manner that reduces the amount of data that needs to be shipped among sites.

Abstract State Machines (ASMs) can also be used to design distributed data warehouses [21]. ASMs provide a strictly mathematically founded method for high-level system design, validation and verification. This method states the separation of input from operational databases from output to dialogue-based on-line analytical processing (OLAP). Data Warehouse Striping (DWS) [5] which is a round-robin data partitioning approach especially designed for distributed data warehouse environments in which fact tables are distributed by an arbitrary number of computers and the queries are executed in parallel by all the computers, guaranteeing a nearly optimal number of speeds up and scale up. This technique is combined with an Approximate Query Answering (AQA) strategy to deal with fails in one or more nodes and is tested over Oracle 9i systems [14, 19].

Grid-Dwpa, an efficient architecture to deploy large data warehouses in grids with high availability and good load balancing. This method contains efficient data allocation, partial replication strategies and scheduling solutions that maximize performance and throughput of the grid-enabled architecture for OLAP [7]. The replication strategies provide adequate guarantees that site availability problems do not impair the system and result in only small system slowdown. This system generates site and node tasks, forecasts the necessary time to execute the task at each local site, estimates total execution times, and assigns task execution to sites accordingly [4,6].

Grid technology is another useful element in distributed computing platforms, and naturally also for distributed data warehouses. The computational grid offers services for efficiently scheduling jobs on the grid, and for grid-enabled applications where data handling is a most relevant part, the data grid becomes a crucial element. It typically builds on the concept of files, sites and file transfers between sites. These use services such as GRID-ftp, plus a Replica Manager to keep track of where replicas are located. The multi-site, grid-aware data warehouse is a large distributed repository sharing a schema and data concerning scientific or business domains. Data warehouse is a single distributed schema and both localized and distributed computations must be managed over that schema. This is done with allocation and processing of data warehouses in distributed and grid environments. The OLAP-Enabled Grid considers the scenario where the data of a single organization is distributed across a number of operational databases at remote locations [11]. Each operational database has capabilities for answering OLAP queries, and access to a possible variety of other computational and storage resources which are located close by. Users who are interested in doing OLAP on these databases are distributed over the network and consider the following entities:

OLAP Server - A machine which has sole control over an operational database. It may maintain some materialized views and may also act as a computational or storage resource. The OLAP servers all have the same schema, but each maintains a partition of the total data available to the users;

Computational Resource - A machine which offers cycles for performing tasks on the behalf of other entities in the Grid;

Storage Resource - A machine which offers disk space for storing data on behalf of other entities in the Grid;

Resource Optimizer - There is exactly one resource optimizer for each site. A resource optimizer has the information necessary to perform scheduling and allocation of computational and storage resources, and to carry out queries. It may also have some cache space for storing common query results for queries generated in its site;

User - Users submit ad-hoc queries to resource optimizers and may enter and leave the network at will. Each user has an amount of cache space for caching query results.
The proposal itself is for a two-tiered grid-based data warehouse. The first tier is composed by local (cached) data. Remote database servers are in the second tier. Each submitted query is evaluated in order to verify if it can be answered with data from the local site. Then, if it cannot be entirely answered locally, the query is re-written into a set of queries. Some of those are executed locally (with the existent data) and the others, which access the data that is missing at the first tier, are executed at the second tier (database servers). The use of cached data is also considered at the database server level. A local data index service provides local information about data stored at each node. A communication service uses the local data index service from the participant grid’s nodes to enable that remote data is accessed. The first step in query execution is to search for data at the local node (using the local index service). Missing data is located by the use of the communication service and accessed remotely [17, 18]. The distributed and grid-aware data warehouse context is still an evolving one, as community data warehouses come into play in current and future systems and concerning different application scenarios. Distributed, grid-aware environment apply the Globus Toolkit together with a set of specialized services for grid based data warehouses [20]. Fact table data is partitioned and distributed across participant nodes. Dimension tables data is replicated. "Query Optimization - Distributed systems" gives support and extension to this implication into distributed data warehouse environment where optimization methods to data design and OLAP queries in the warehouse environment should be implemented with an objective of supporting the decision makers by providing a single view of data even though that data is physically distributed across multiple data warehouses in multiple systems at different branches [15].

V. CONCLUSIONS AND FUTURE TRENDS

There has been a significant amount of work during the last two decades related to distributed data warehouses environment. These works have contributed to increasing significantly our knowledge of those systems, issues and solutions, and it has also brought some maturity to the field. Still lot of research should take place for the formulation of various approaches to design methods, architectural models, data fragmentation-allocation, distribution strategies, usage of systems globally keeping in view of various parameters of distributed warehouses and finally various optimization strategies. Then we can see a role of pure distributed data warehouse systems developed by the leaders of warehouse systems globally for better and advancement in the computing. In this paper we reviewed the main concepts, architectural models and trends on distributed data warehouses architectures and systems. We explained some of the most relevant works on distributed data warehouse systems. Work in distributed data warehouses in the future is expected to advance the concepts and systems to new levels of autonomy, scalability and ubiquity. It will provide answers to the issue of how to completely automate and optimize allocation and mixes of base data, materialized views, cubes and indexes in distributed settings for optimal performance. We will also increasingly see applications of data warehouse and grid technologies to distribute and collaborative applications and problems.

REFERENCES


[20] Xiao et al, 2007-“Evolving a secure grid-enabled, Distributed Data Warehouse: A standards –Based Perspective”.