ISSN:0975-9646

Shaweta/ (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1692-1695

A Review on Designing of Distributed Data Warehouse and New Trends in Distributed Data Warehousing

Shaweta
Assistant Professor, UIET, Panjab University,
Chandigarh, India

Abstract- A distributed data warehouse is a conglomeration of separate components that are connected via a network. The goal is to have these separate components appear as a single global data warehouse image. A distributed DW, the nucleus of all enterprise data, sends relevant data to individual data marts from which users can access information for order management, customer billing, sales analysis, and other reporting and analytic functions. This paper particularly discusses distributed data warehouse systems frame work for distributed data warehouse systems and recent developments of data base designs in distributed data warehouse architectures.

Keywords-- *Distributed Data warehouse, Database, Framework, Optimization, schema and data mart.*

1. INTRODUCTION:

In modern era of economic changes, market strategies of information dimensions changes globally and locally. Each and Every company has its information of business and other things on internet and by this facility its decision power become easy and on right time. Each company stored their information in a database so they required a data warehouse. All organization spending a larger sum of money in this technology so that they can gain first in this competitive environment of productivity. Data warehouse is a set of materialized views over data sources [12], [13], [14]. Ralph Kimball et al defined "A data warehouse is a copy of transaction data specially structured for query and analysis" [14].

But due to competitiveness' of market all enterprises has thinks on a larger platform and then acted on it in their enterprises. For this need of changes in data warehouse required and distributed data warehouse fulfill these requirements perfectly.

The rest of this paper is as follows: Section 2 illustrates the Distributed Data warehouses. Section 2.1 illustrates the features that are required in a distributed data warehouse. Section 3 describes framework of distributed data warehouse. A brief about design approach of distributed data warehouse is shown in section 4.1 & 4.2. Section 5 explains the recent developments and trends in distributed data warehouses. I conclude in section 6 and give a brief description of future work.

2. DISTRIBUTED DATA WAREHOUSES:

When it comes to make a data warehouse either its simple data warehouse or distributed data warehouse. Some features that are primary feature of a data warehouse are required as follows:

2.1. Distributed Data warehouses Features:

- The data copied into a data warehouse does not change (except to correct errors). The data warehouse is a historical record of the state of an organization. The frequent changes of the source OLTP systems are reflected in the data warehouse by adding new data, not by changing existing data.
- Data warehouses are subject oriented, that is, they focus on measuring entities, such as sales, inventory, and quality. OLTP systems, by contrast, are function oriented and focus on operations such as order fulfillment.
- In data warehouses, data from distinct function-oriented systems is integrated to provide a single view of an operational entity.
- Data warehouses are designed for business users, not database programmers, so they are easy to understand and query.

However, there are considerable disadvantages involved in moving data from multiple, often highly disparate, data sources to one data warehouse that translate into long implementation time, high cost, lack of flexibility, dated information and limited capabilities:

- Major data schema transforms from each of the data sources to one schema in the data warehouse, which can represent more than 50% of the total data warehouse effort
- Data owners lose control over their data, raising ownership (responsibility and accountability), security and privacy issues
- Long initial implementation time and associated high cost
- Adding new data sources takes time and associated high cost
- Limited flexibility of use and types of users - requires multiple separate data marts for multiple uses and types of users
- Typically, data is static and dated
- Typically, no data drill-down capabilities
- Difficult to accommodate changes in data types and ranges, data source schema, indexes and queries
- Typically, cannot actively monitor changes in data

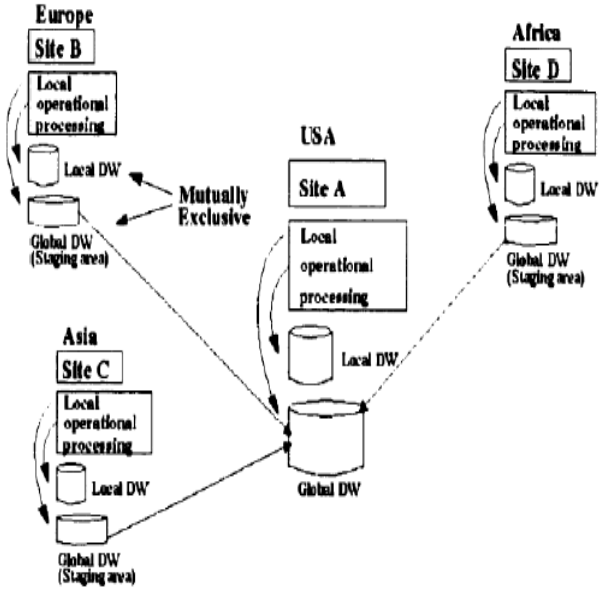
Therefore, Lot of enterprises decided small, pliable data marts that are specific to specific business areas. To get the cross-functional analysis, there are two possibilities [8, 9, and 10]. The first one is to create again a centralized data warehouse for only cross-functional summary data. The other one is to integrate the data marts into a common conceptual schema and therefore create a distributed data warehouse. "Information Data warehouse" a strategic approach helps mainly to understand the successful execution of enterprise initiatives [15]. For building an information data warehouse, build a well-organized and successful tool for the whole organization where different individuals feel agreeable and easy going to solve their problem or tasks. This paper is regarding a review of different architecture and design methods in organization for a data warehouse.

3. FRAMEWORK FOR DISTRIBUTED DATA WAREHOUSES:

Two approaches to build the distributed data warehouses are available which are described as follows:
(1)Inmon's Approach, (2) White's Approach.

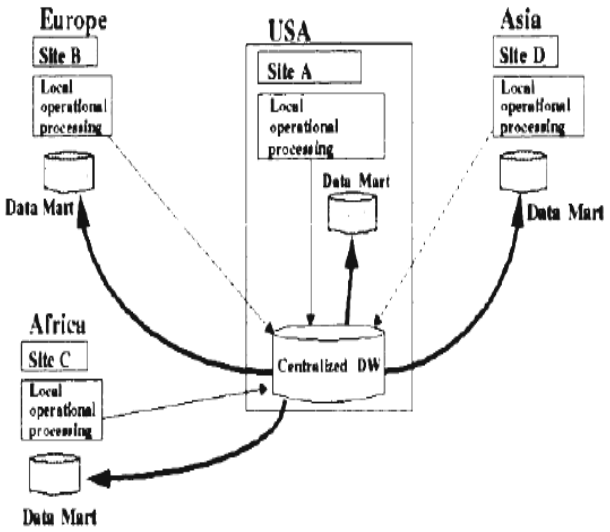
3.1. Inmon's Approach:

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 data 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.



3.2. 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 data warehouse data and may contain denormalized detailed data as well as summarized data.



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.

4. DISTRIBUTED DATABASE DESIGNS:

In the literature of distributed data environment, where two approaches for distributed data base 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.
(1) The top-down approach and
(2) The bottom-up approach.

4.1. Top-Down Design [1, 2, and 6]:

In the top-down design approach the, data warehouse is built first. The data marts are then created from the data warehouse. This design is a process of creating data models that contain high-level entities and relationships, to which successive refinements are applied, in order to identify the corresponding low-level entities, relationships and attributes. The top-down approach is illustrated by using the concepts of the entity-relationship model.

- Analyzing requirements;
- View integration and conceptual design;
- Data distribution design;
- Local physical schema design.

4.2. Bottom up Approach [3, 4, and 5]:

Bottom-up approach 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. One of the most important aspects of design strategy is to determine how to integrate multiple database system together. Implementation alternatives are classified according to the autonomy, distribution, and heterogeneity of the local systems.

- 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.

5. RECENT DEVELOPMENTS IN DISTRIBUTED DATA WAREHOUSES ENVIRONMENT:

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 [16].

Before the iPhone and Xbox, prior to the first Tweet or Facebook "Like," and well in advance of tablets and the cloud, there was the data warehouse. For 30 years, businesses have centrally stored data for analysis and data-driven decision making.

For all of that time, the data warehouse has been the business-insights workhorse of enterprise computing. The big trend in the mid 1990's was the emergence of data warehouses that were a terabyte in size, which at the time was considered a huge amount of data. Today's leading-edge systems are a thousand times larger—measured in petabytes.

Data warehouses have had staying power because the concept of a central data repository—fed by dozens or hundreds of databases, applications, and other source systems—continues to be the best, most efficient way for companies to get an enterprise-wide view of their customers, supply chains, sales, and operations.

ASMs support refinement method in developing a data warehouse and OLAP systems [8]. of One strategy Abstract State Machines (ASMs) also be used to design a distributed data warehouses also Abstract State Machines provide a meticulous mathematically tricks for high-level system design, validation and verification at earliest stage of system development [9].

Wehrle et al (2007) also deal with a distributed, grid-aware environment. They apply the Globus Toolkit together with a set of specialized services for grid based data warehouses. Fact table data is partitioned and distributed across participant nodes.

Grid computing has emerged as a new technology, whose main challenge is the complete integration of heterogeneous computing systems and data resources with the aim of providing a global computing space. A new Data Mining Grid Architecture, named DMGA, which for their composition in a real scenario [18].

Dimension tables data is replicated. 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.

Bindia et. al. (2012)A multi agent based system based query cycling process in distributed DWH and also remove the disadvantage of multi agent approach by placing an buffer, so that there will be no need for client to connect at all the time to get the results[10].

Distributed systems provides support where optimization methods to data design and OLAP queries in the data 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].

Another work on grid-aware data warehouses is presented in (Lawrence Rau-Chaplin 2006).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. 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 [17].

Below are some new trends and opportunities in data warehousing.

- The "datafication" of the enterprise requires more capable data warehouses.
- Physical and logical consolidation help reduce costs.
- Hadoop optimizes data warehouse environments.
- Customer experience (CX) strategies use real-time analytics to improve marketing campaigns.
- Engineered systems are becoming a preferred approach for large scale information management.
- On-demand analytics environments meet the growing demand for rapid prototyping and information discovery
- Data compression enables higher-volume, higher-value analytics
- In-database analytics simplify analysis

6. CONCLUSIONS AND FUTURE TRENDS:

Data warehousing is not a new phenomenon. All large organizations already have data warehouses, but they are just not managing them. Over the next few years, the growth of data warehousing is going to be enormous with new products and technologies coming out frequently. In order to get the most out of this period, it is going to be important that data warehouse planners and developers

have a clear idea of what they are looking for and then choose strategies and methods that will provide them with performance today and flexibility for tomorrow. Future work should attempt to various allocation techniques of distributed database.

REFERENCES:

[1] Ileana ŞTEFAN and Maricel POPA "Distributed Database Design – Top-Down Design '' , Volume 48, Number 1, 2007.

[2] Ruby Bhati et.al. Distributed Database System: The Current Features And Problems. In International Journal of Computer Science and Management Research, Vol. 2 Issue 3 March 2013 ISSN 2278-733X.

[3] Shailesh R. Thakare, C.A. Dhawale, Ajay B.Gadicha "Design Distributed Database Strategies for SQMD Architecture '', (IJEAT)ISSN: 2249 – 8958, Volume-1, Issue-2, December 2011.

[4] Ceri and Pelagatti " Distributed Databases: Principles and Systems", McGraw-Hill, 1984.

[5] Hsiang-Jui Kung,LeeAnn Kung ,Adrian Gardiner "Comparing Top-down with Bottom-up Approaches '', in 2012 Proceedings of the Information Systems Educators Conference ISSN: 2167-1435,New Orleans Louisiana, USA v29 n1910.

[6] Ozsu and Valduriez " Principles of Distributed Database Systems", Prantice Hall, 1991.

[7] Xiao et al, 2007-"Evolving a secure grid-enabled, Distributed Data warehouse: A standards –Based Perspective".

[8] Refinements in Typed Abstract State Machines Sebastian Link, Klaus-Dieter Schewe, and Jane Zhao, PSI 2006, LNCS 4378, pp. 310–321, 2007. Springer-Verlag Berlin Heidelberg 2007

[9] Zhao and Dieter, 2004- "Using Abstract State Machines for Distributed Data warehouse Design", ACS, APCCM2004, Dunedin, New Zealand.

[10] Bindia, Jaspreet Kaur Sahiwal "Agent Based Architecture in Distributed Data warehousing by '', International Journal of Scientific and Research Publications, Volume 2, Issue 5, May 2012 ISSN 2250-3153.

[11] Rogério Luís de Carvalho Costa, Pedro Furtado "Data warehouses in Grids with High QoS(2006)'' ,Online ISBN 978-3-540-37737-5 Springer Berlin Heidelberg, pp 207-217.

[12] Z. Bellahsene, Schema, "Evolution in Data warehouses", Knowledge and Information Systems, Springer-Verlag, pp 283-304, 2002.

[13] 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.

[14] 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

[15] Sagar Yeruva and Dr.P.V.Kumar, "Development of Information Data warehouse- A Strategic Approach"- International Journal of Computing and Applications, Vol. 5, No. 2, July-December-2010, pp. 153-158.

[16] Nouman Maqbool Rao, Muheet Ahmed Butt, Majid Zaman, Waseem Jeelani Bakshi "Distributed Data warehouse Architecture: An Efficient Priority Allocation Mechanism for Query Formulation '', ISSN: 2277-3754 , ISO 9001:2008 Certified International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 9, March 2013,pp-157-159.

[17] Akinde, M. O., Bhlen, M. H., Johnson, T., Lakshmanan, L. V. S. and Srivastava, D. (2003) "Efficient OLAP query processing in distributed data warehouses", Information Systems 28, pp.111-135, Elsevier, 2003.

[18] Marıa S. Perez, Alberto Sanchez , Vıctor Robles, Pilar Herrero, Jose M. Pena "Design and implementation of a data mining grid-aware architecture'', Science Direct(Elseveir)Future Generation Computer Systems 23 (2007) 42–47.