Data Warehouse using Parallel Processing on a Distributed Environment

WALDEMAR RUGGIERO JUNIOR, LIRIA MATSUMOTO SATO
LAHPC
EPUSP
Av. Professor Luciano Gualberto,Trav 3 -158- ZIP-05508-900
BRAZIL

Abstract: - This paper proposes architecture for implementation of a Data Warehouse in a distributed environment, using parallel programming. With the increase of volume of data stored in Data Warehouse, the traditional architectures need high performance in terms of processors and input and output systems. This kind of problem is well characterized when it is performed a high complex queries (Ad hoc). Using distributed environment, together with parallel programming is a good choice to increase the performance and to reduce cost. It's presented a proposal here for architecture of distributed Data Warehouse, integrated with the use of parallel programming. his is a sample of the format of your full paper.

Key-Words: - Data Bases, Data Warehouse, Parallel Programming.

1 Introduction

Many commercial and scientific applications manipulate voluminous data sets [1, 2, 3] and in many cases the input and output systems meet problems the same occurs in the processors. This question is best characterized when complex operations on a data base are executed.

Solutions that involve great amounts of data in centered systems make possible the implementation of centered Data Warehouse. In these cases they necessarily need processors and systems of storage of data with high performance. However, even with such systems the execution of complex searches can be very slow.

Distributed Data Warehouse architecture reduces the systems needs especially high performance. This approach is still more advantageous if part of the data is remotely used and information is segregated according to some criteria, especially the regional criteria [4].

The proposed system, in this article, uses distributed processing data base concepts allowing the information processing and access in remote places through nets of high or low speed. It includes and can be applied to other necessities. Either systems OLTP "On Line Transaction Process" as OLAP "On Line Analytical Processing" can be adapted to use this solution.

Parallel programming using MPI, Message Passaging Interface, are used to standardize the codification and to minimize the time of execution of the searches. The execution of these activities in distributed environment will prevent difficulties of performance of the complex searches. Such searches will become more efficient reducing databases or increasing numbers of processors.

2 Operational and Data Warehouse Data

The operational data systems are characterized by processing information daily. They are used as storage for the commercial transactions such as inventory control, payment control and bank automation.

Data Warehouse is used in systems planning, management systems and systems forecast. This knowledge is obtained from the operational data, transformed into information and later into knowledge.

Table 1 makes a comparison between Data Warehouse and Operational Data.

Data Warehouse	**Operational Data**
Subject Oriented	Application Oriented
Integrated Data	Limited Integrated Data
Non Volatile	Permanent Update
Permanent Data	Current Data
Ad Hoc Searches	Predicable Searches

Table 1. – Data Warehouse and Operational Data

2.1 OLTP and OLAP

OLTP (On Line Transaction Process) Systems are used to support business operations of processes of an organization day by day. OLAP (On Line Analytical Processing) Systems are used on decisions support. They provide administrators of an organization a multidimensional vision used to analyze the existing business-oriented profiles or to create conditions of analysis of existing standards of behavior. The following characteristics differentiate systems OLTP of systems OLAP:

- In OLTP systems the data are updated and shown in details, in OLAP systems the data are historical, summarized and consolidated from some operational bases, enclosing long periods.
- The size of the data stored in OLTP systems are ordered by sets of ten of Gigabytes, in systems OLAP hundreds of Terabytes.
- In systems OLTP the searches are simple like inserting, retrieving, updating. In OLAP the searches are complex, but using a standard model.
- The data model of data in OLTP is normalized, unlike OLAP.
- OLAP represents a set of projected technologies to support analysis and search ad hoc [6].

2.2 Data Mining and KDD Process

The data mining is the most important phase in the process of transformation of the operational data in knowledge. This transformation is known as discovered of knowledge in data bases (KDD - Knowledge Discovery in Databases). The objective of the KDD process is to facilitate comprehension of standards to the people through interpretation of the existent data. What searching knowledge needs in the data bases is a consequence of the growth of the storage of the historical data. To use the advantage contained in this data set, they are organized in a easy form to identify standards that can help in the prediction of future actions. With the aid of statistical techniques a mechanism for such accomplishments is created. [7].

It is possible to use some tools and techniques for the mining, being the most used in those databases based in consultation such as language SQL.

A data mining is the most important part of a process involving the discovery of the knowledge. The process of Knowledge Discovery in Databases (KDD) involves other stages:

- Data Warehousing
- Pre-Processing
- Cleanness, Selection and Codification
- Enrichment
- Data Mining
- Post-Processing

3 Data Warehouse Architecture

The Data Warehouses are especially dedicated and voluminous data bases containing integrated data from various independent sources, supporting customers whom desire is to analyze the data and to verify trends and anomalies. The analysis process is usually executed by operations like adding, filtering and grouping the data in a variety of ways [8].

The data, stored in the Data Warehouses, are proceeding from systems that produce operational data. The operational data have a lot of information and are used in the daily operations of the informational systems.

Using the processes of extraction, transformation and loading (ETL - Extract, Transformation and Load) from the operational data, it gets the data for fulfilling the Warehouse.

The Data Marts are subset of the Data Warehouses that group part of the information, generally applied to one determined purpose. Generally, they are data referring to a subject in particular, for example, selling, engineering, controlling or different levels of summarizing, such as, annual selling, monthly selling, five years selling, focusing on one or more specific areas. Data marts extract portions of Data Warehouses to the specific requirements of sectors or departments of enterprises. There are some approaches for the architecture of one Data Warehouse, among them are the following:

- Centralized Data Warehouse
- Virtual Data Warehouse Virtual
- Distributed Data Warehouse

3.1 Centralized Data Warehouse

Centralized Data Warehouse is usually used when organizations or companies have a clear definition of the users needs access. The centralized data propitiate greater quality and integrity.

3.2 Virtual Data Warehouse

Virtual Data Warehouse provides access to the operational data for the final user to do searches of behavior knowledge or profile direct into the operational base. This solution restricts the types of search implementation but the bases are also used by the transactional systems. To facilitate this access layers of software are created with the prescribed objective of the access to the bases shared with two purposes operational and informational. One advantage of this solution is the low cost; therefore it does not have duplicity in the storage of data. A disadvantage is that the complex searches are carried through in the operational base, being able to diminish the availability of this base, many of the data can not be in the necessary form to be used by a final user, i.e., the development were for systems OLTP and not OLAP [9].

3.3 Data Marts

The Data Marts supplies the data of interest of a user, a sector or a department. This allows having more control of the part of the user in terms of requirements of data and manipulations. The data in the Data Marts are finer or of bigger interest for the user while data in the Data Warehouses are detailed. This characteristic makes possible for the Data Marts to be minor, increasing the performance of the search [9]. The disadvantage of this technique is that a global vision of the data does not exist.

3.4 Data Warehouse and Data Marts

This data architecture is a combination of the Centralized Data Warehouse and the Data Marts. It gets advantage of the two solutions: complete integration of the Data Warehouse and the optimized access of the Data Marts. They supply the data of interest on a user, sector or department allowing bigger control of the user in terms of requirements of data and manipulation.

3.5 Distributed Data Warehouse

The proposal of Inmon, figure 1, for implementation of the Distributed Data Warehouses [10] mentions the existence of central and local Data Warehouse, where the data are mutually exclusive. The extraction and loading in this architecture are also being distributed.

The proposal of White [11], figure 2, known as "two to tier Warehouse Date", combines Centralized Data Warehouse and Decentralized Data Marts. The Data Marts contains data not normalized and reduced, reflecting aspects of some users or groups and users.

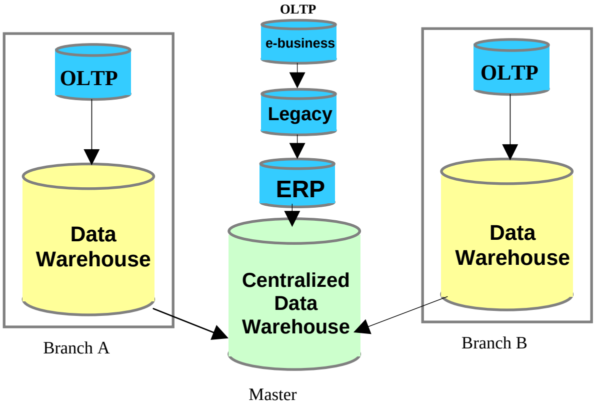

Figure 1 – Distributed Data Warehouse – Inmom

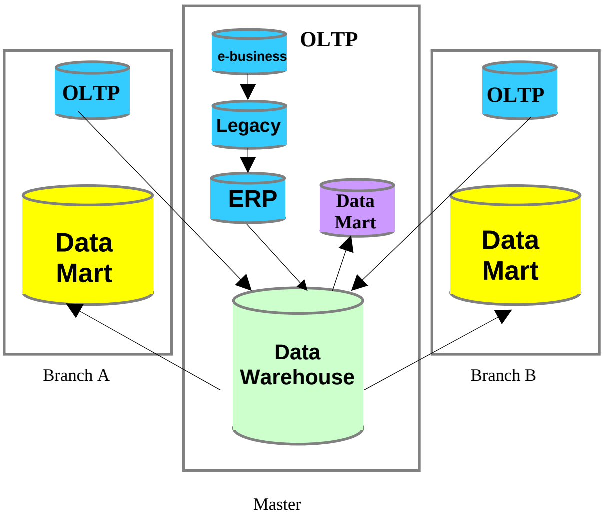

Figure 2 – Distributed Data Warehouse – White

4 Completely Distributed Data Warehouse Architecture Proposal

The architecture proposed claims to distribute the Data Warehouses in distant points receiving the load from the data of local business systems. From a central point programs are distributed to be executed in the distant points. The result of this processing is returned to the central point and is consolidated, being stored in a posterior database for consultation. The considered system, for example an application to a banking environment, allows the creation of behavior reports of the customers. The information explored in the Distributed Data Warehouses is sent to the central system for consolidation and to provide a global vision of results. The use of Data Warehouse allows extracting information of the logs registers of the transactional systems, OLTP, for posterior verification. These analyses are important for the knowledge of the behavior of the customers in several aspects.

It is possible to forecast the type of product to be offered to one determined customer, being based on information as sex, age, income, geographic profile and economic profile. Moreover, there is information of the relationship of the customer with the institution whose canals of contact are used, what frequency, which products the customer already possesses, which type of claims the customer has already effected, etc.

Based on these information it is possible to forecast the propensity use, of purchase and the behavior proper to the customers. It is also possible to identify customers with similar profiles, where the results of the actions will get success with minor effort. Proposed technique architecture presents similarities with the proposal of Inmom [10]. The figure 3 shows the proposed architecture. The main difference is that the proposal is not centralized. The operational data are converted and stored into local Data Warehouse through techniques of extraction, transformation and loading. In the central point the requests to the Data Warehouses are really distributed by the agencies. These requests are dealt with through execution of the searches in the processing of the branches, being the results returned to the central point, to be consolidated and stored. With the objective to diminish the time of replying to similar requests that are carried out by other users, a data base in the central point for storage of the result of the searches is created, or the same of the gotten reports. These results will be consolidated for the use in the central point, being able to be consulted by the remote points, through a portal.

The central point data base is used to store the reports of the searches temporarily but not being organized as one Data Warehouse. The period of storage can be very small, for example, one month, therefore an old report will be requested and a new search in the Warehouses Distributed data is affected in the way that then the stored volume is low. At the central point, optionally, it could have one Data Warehouse, not for centralization of the data purpose, but only for local use of the branches.

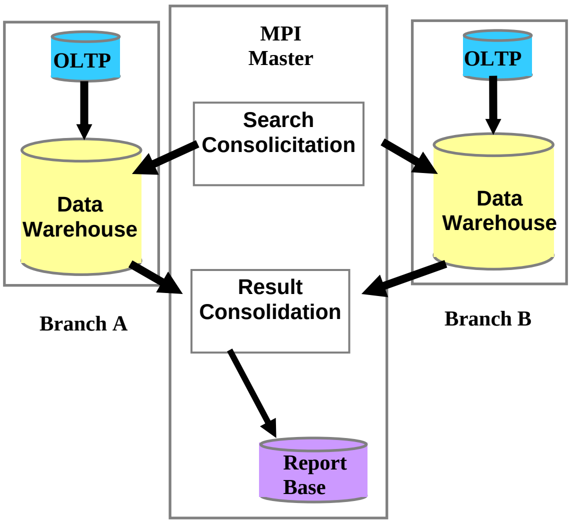

Figure 3 – Completely Distributed Architecture proposal.

5 Meessage Passing Communication System Solution

One method of parallel programming computation is the use of a messages ticket library. This library transfers data between instances of programs being processed using multiple processors. With this method it is possible to have available great space of memory and a bigger number of processing central units. Therefore it is possible to solve problems of high complexity that are not solved normally with traditional methods.

In the proposed solution the implementation of parallelism of tasks is done through the paradigm master slave, as shown in following figure 4.

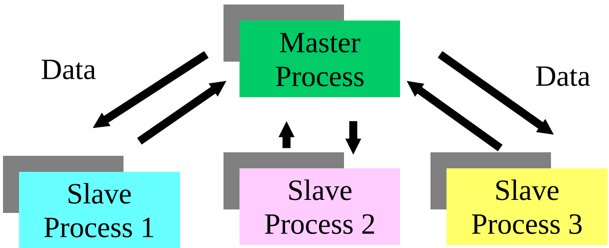

Figure 4 - Process Master Slave

The programs, using the messages passeging library, are composed of programs with multiple instances that can be communicated through the calls of the library. These calls are divided in four classes:

- Initiation, Management and End of communication.
- Communication between pair of processes
- Communication operations among several processes.
- Data types creation.

The communication is made by using message passeging. Pair of processes communicates by using commands send and receive. Generically, the message is composed of an envelope indicating the source and the destination is a body with data to be transmitted. The master process is executed in the central point. It distributes tasks for the enslaved processes sending programs to be processed. Each slave executes its task returning the result. The master, then, makes the consolidation mounting the reports. The programs are written in language C using library MPI (Message Passage Interface) [12] for the ticket of the messages and distribution of the code between the slaves. It also used on a library SQL for the execution of the necessary commands for access to the data banks. Either OLTP or OLAP systems can be adapted to use the solution that combines the use of distributed data bases with the ticket of messages programming.

6 An Application Based on the Proposed System

As a case study it was chosen a presentation of a bank application, having involved concepts of CRM (Customer Relationship Management). This application involves the Data Warehouses located in the branches. At headquarters there is the point of distribution of the programs to be executed in all the branches. The information is extracted of the transactional systems and is loaded in the Data Warehouse, one in each branch, obeying a model of data that allows through exploration of data the generation of knowledge for boarding to the customers. With this commercial law action they could be carried through in the scope of a branch, of a group, or the same, involving all the branches of the bank.

The model of data, in the applications of Data Warehouse, is one of the most important points for the success of the implementation of the project. In systems with the technique vision only in its implementation the technical project can be a success under the performance, availability and security aspects, but not in the business-oriented aspects where it finishes being a failure and with little usability for the users. Therefore it is important to think about two aspects: technical and business one.

Each branch has its Data Warehouse that is loaded with the data proceeding from the transactional systems obeying a data model. The model is identical for all the branches. The historical information is stored in the Data Warehouse, once stored it is possible use tools to explore information in local searches.

The global searches using all branches Data Warehouses are made from the master unit through the use of message interface to send commands SQL for all the agencies. After the accomplishment of the requested operations for the master the results are returned for consolidation having generated reports by itself that are stored in the data base specified in the proposed architecture.

This frequent research is requested by the users. In this in case, the results are available in the master, being able to be consulted through a vestibule, where it is not necessary to use any tool of exploration.

In this application the model of data, as showed in figure 5, possesses information of: name, address, city, state, country, telephone, agency, account, CPF, general register, products that the customer possesses, canals what the customer uses, claims of a customer and etc.

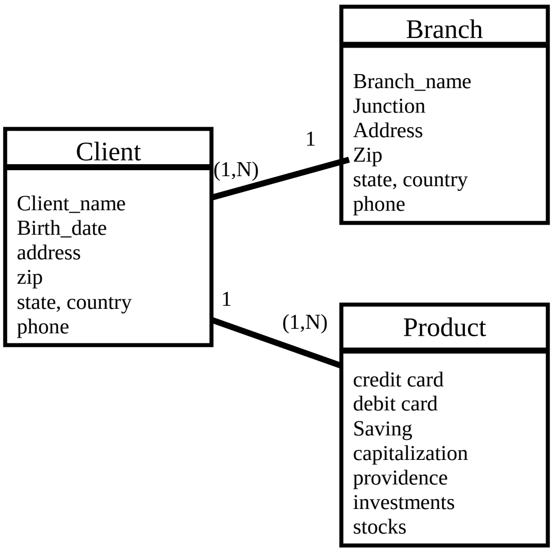

Figure 5 - Conceptual Data Model

A figure 6 shows the integration of the architecture proposal to the conceptual model of data, showing the main elements of the solution.

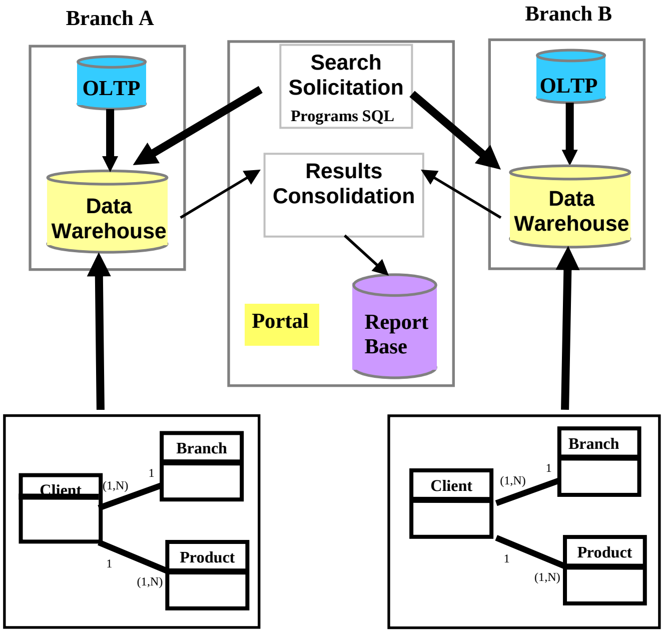

Figure – Integrated Architecture and Data Model

7 Conclusion

The proposed architecture uses distributed data bases driven from a central point and it is generic to be able to apply it in many environments such as information and Data Warehouse. The parallel programming standardizes the codification and minimizes the time of execution for determined searches. The system makes possible consolidated vision of the information possessing distributed databases that is done through the publication of the main searches in portal. As future work, it can be studied the aspects of performance under the point of view of the net of communication between the master and the branches, as well as, implementation of a tolerant system to the imperfection.

References:

[1] Oldfield,R.;Kotz, D. Applications of Parallel I/O. *Technical Report PCS-TR98-337 ,PCS-TR96-297,* Computer Science, Dartmouth College, 1998.

[2] Kotz, D.; Ellis, C. S. Practical Prefetching Techniques for Parallel File Systems. in Proceedings of the *1st International Conference on Parallel and Distributed Information Systems,* 1991,pp.182-189.

[3] Poole, J. T. Preliminary Survey of I/O Intensive Applications. Technical Report CCSF-38, *Scalable I/O Initiative, Caltech Concurrent Supercomputing Facilities, Caltech*, 1994.

[4] Costa Neto, José Craveiro *Considerações sobre a integração de um Banco de Dados e um Data Warehouse sobre um Sistema de Arquivos Paralelos*: São Paulo, 2001.

[5] Poess, Meikel e Othayoth, Raghunath K. Large Scale Data Warehouses on Grid: Oracle 10g and HP Proliant Servers – *Proceedings of the 31th VLDB Conference, Trondhein ,Norway* 2005

[6] Kimball, R. *The Data Warehouse Toolkit : Practical Techniques for Building Dimensional Data Warehouses*. 1996.

[7] Michael Goebel e Le Gruenwald. A Survey of Data Mining and Knowledge Discovery Software Tools *ACM SIGRID*. Volume 1 issue 1. 20 33. June 1999.

[8] P. O'Neil, D. Quass. Improved Query Performance with Variant Indexes. In Proc. of the *ACM SIGMOD Conference, Tuscon, Arizona*, May, 1997.

[9] Noaman A.Y. *Distributed Data Warehouse Architecture and Design*: Manitoba, Canada, 2000.

[10] Inmon W.H. J.E. *Building the Data Warehouse*: ,1996

[11] White C. A Technical Architecture for Data Warehousing. *InfoDB Journal.9 (1):5 11.* February 1995.

[12] Quinn, Michael J. *Parallel Programming in C with MPI and OpenMP*. June 2003.