ISSN 0976 – 1411

Available online at www.internationaleJournals.com

International eJournals

INTERNATIONAL eJOURNAL OF MATHEMATICS AND ENGINEERING

www.InternationaleJournals.com

International eJournal of Mathematics and Engineering 120 (2011) 1106 - 1112

NETWORK RELATED ISSUES IN DATA WAREHOUSING

T. V. SURYA NARAYANA, V. MURALIDHAR, K. S. VIJAYA SIMHA
Tenali Engineering College, Anumarllapudi (v), Guntur
Don Bosco Engineering college, Guntur
Email: vmdha.research.@gmail..com, tvenkata.surya@gmail.com, s.kollam@.yahoo.com

ABSTRACT:

The Date warehouse delivery method is a joint application development of the Date warehouse concepts and Network concepts. Data warehouse may be the first large scale Client Server solution being implemented within organization, and will require new skills experience and hardware in both data management and network management. The computer networks especially in this study are considered as a vital part in development and implementation of Data warehouse.

IMPLEMENTATION:

In order to produce a good quality Data warehouse we have to understand two main points in data warehouse delivery process.

- Business Requirements.
- Technical Blueprint.

In general we allow only 20% of time to analyze business requirements and remaining 80% of time for technical analysis and evolution. Here the technical blue print phase has an overall architecture that satisfies the long-term requirements and it must identify by the following steps.

- The overall system architecture.
- The server and data mart architecture for both data and application.
- The essential components of the database design.
- The data retention strategy.
- The backup and recovery strategy.
- The capacity plan for hardware and infrastructure (for eg: LAN and WAN)

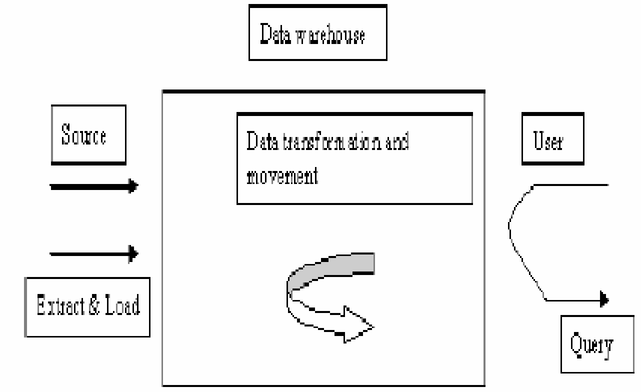
The next phase in Data warehouse delivery process is building vision which is the first production deliverable is produced. At this stage we will probably build the major infrastructure components for extracting and loading date. Here the major Infrastructure component is the Networks, which extract the data from various heterogeneous databases located at different areas to centralized Data warehouse. Another phase in Data warehouse delivery process is ad hoc query. This is an End-user access tool, which is capable of automatically generating the database query that answers any question posed by the user. The user may pose their questions from any one of the client systems on to a central Data warehouse. Here also networks play a vital role. The computer networks may influence almost all phases in delivery process and also various issues in data warehouses.

EXTRACT AND LOAD PROCESS:

Data warehouses are built to support large data volumes (above 100 GB of database) cost effectively. The data warehouse architecture must have three driving factors.

- Populating the data warehouse.
- Day-to-day management of data warehouse.
- The ability to cope with requirements evolution.

The following figure shows the process flow with in a data warehouse. The process requires populating the data warehouse focus on extracting the data, cleaning up, and making it available for analysis



This is the major area where the networks affect the extracting and loading process. The data extraction takes data from various source systems and makes it available to the data warehouse. The data load takes extracted data and loads it into data warehouse. Here we going to take care of selecting a suitable network according to the capacity of the source databases and we also analyze which kind of the network infrastructure are supported by the source database systems. We also concentrate on consistency of data during the extraction process. That is the network may not change either the format or meaning of the data during the extraction process. We are also trying to minimize the transmission errors between source databases to data warehouse. Some times the data warehouse is connected to the operational databases by a single network topology or by different network topologies.

The connection is in single network topology it is simple to optimize the performance of the data extraction activity. If the connections are multiple network topologies that make heavy complex to optimize the performance of same activity. We are also choosing the network, which is flexible to implement the data management tools for checking consistency and to clean data with in data warehouse or the user adopts the code language, which is flexible and suitable for quick processing on already established network connections between operational databases and data warehouse. The current gateway technology operates too slowly to compete with the use of FTP and database load technology. This may change in future, but the gateways operate on an SQL basis, is unlikely that they will ever better the performance of the recommended mechanisms. Gateway technology tends to be in appropriate for large data volumes. This technology may be appropriate where data volumes to be loaded are small.

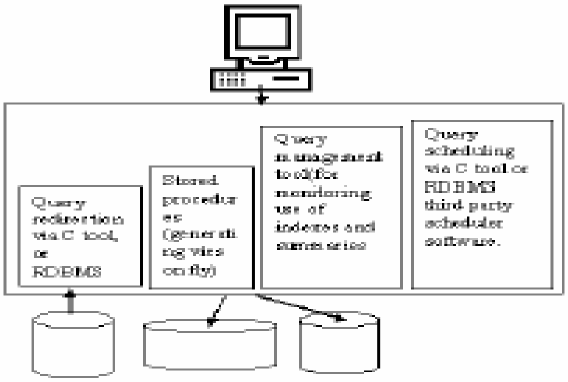
QUERY MANAGEMENT PROCESS:

The query management process is the system process the system process that manages the queries and speeds them up by directing queries to the most effective data source. This process ensures that all the system resources are used in the most effective way. This process operates at all times that the data warehouse is made available to end-users. We are take care of large queries may soak up all system resources, affecting the performance of entire system. We must ensure that no single query can affect the overall system performance.

The networks may affect this area by:

- Here we are not only considering the network topology but also consider bandwidth, channel capacity and maximum data rate of transfer. The queries posed by user may smoothly transfer to the data warehouse.
- We are also calculating the traffic on the network, because data warehouse provide OLAP operations.
- We may investigate for a new or appropriate routing algorithm to avoid the heavy traffic on network.
- Network may also be flexible to show the some data warehouse tools, which are helpful to search effective data source for the corresponding query.

The following figure illustrates the query manager architecture:


Query redirection via C tool, or RDBMS
Stored procedures (generating views on fly)
Query management tool(for monitoring use of indexes and summaries
Query scheduling via C tool or RDBMS third party scheduler software.

DESIGN OF DATA WAREHOUSE:

One of the major technical challenges with in the design of a data warehouse is to structure a solution that will be effective for a reasonable period of time. This implies that the network topology, protocols, routing algorithms etc. should not have to be restructured when a business changes or query profile changes.

Partitioning of Warehouse

One of the other components in data warehouse affected by the networks is hardware partitioning. That is we are partition the data warehouse into small chunks and put them onto different nodes to maximize the warehouse hardware performance. We are going to achieve this by two methods.

- Stripping data across nodes.
- Partitioning the tables horizontally and these partitions are distributed across nodes.

In both methods a network to exchange the information as well as to facilitate for parallel processing connects the active data disks in different nodes. One more thing is, database technology may not able to recognize on which node the table (Data warehouse) partition resides. So it is also the responsibility of network to search by it self for appropriate partition table according to the user query on the network or to allow any packaged search tool on the network for data warehouse partitions.

Effects on Data marts:

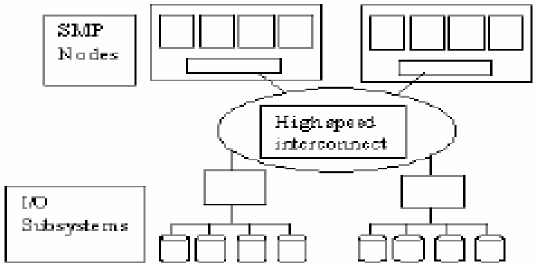
Data marts hive off a subset of the data in the data warehouse, into a separate database, possibly in a different location. Here we know that data to be loaded into an enterprise data warehouse, and then to be data mated. When we proposed for data mating we remember that hardware cost is going to offset the performance improvements. In many cases hardware costs can't be justified. Here network play key role, when connecting data warehouse and data marts. The process of loading data into each data mart will be affected by the available capacity of the physical connections between both items of hardware. We know each data mart will be in a different geographical location from the data warehouse, so we ensure that the LAN or WAN has sufficient capacity to handle the data volumes being transferred within the data mart load process.

Effects on Ware house Architecture:

Another issue of data warehouse affected by networks is warehouse architecture; it can be classified into two categories.

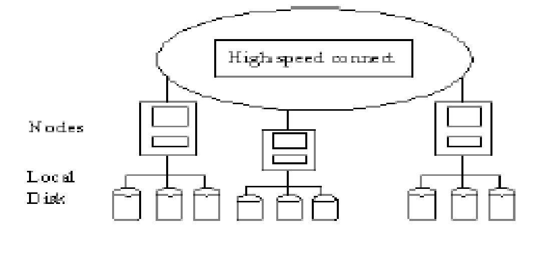
- Symmetric multiprocessing.
- Massively parallel processing.

A Symmetric multi-processing machine is a set of CPUs that share memory and disk. Here the network influence factor is, the communication bus connecting the CPUs is a natural limit to the scalability of SMP machines.



The above diagram illustrates the structure of the cluster symmetric multiprocessing system.

The massive parallel processing machine is made up of many loosely coupled nodes. These nodes will be linked together by a high-speed connection. The form of this connection varies from vendor to vendor. Each node has its own memory, and the disks are not shared each being attached to only one node. Most massive parallel processing system allows a disk to be dual connected between two nodes. The following machine shows the structure of Massive parallel processing machine.



The network although not part of the data warehouse itself, can play an important part in data warehouse's success. As long as the network has sufficient bandwidth to supply the data feed and user requirements, the architecture of the network is irrelevant. The main aspects of data warehouse design that may affect by the network architecture are,

- User access.
- Source system data transfer.
- Data extractions.

Network management requires specialist tools and lots of network experience. It is important to be able to monitor network performance. The network may play a key part in data flow through a data warehouse environment.

Data warehouse security:

When we are considering the security requirements, it is important not to overlook network security issues. In general we observe:

- Is it necessary to encrypt data before transferring it to the data warehouse machine?
- Are there restrictions on which network routes the data can take?

These restrictions have processing implications that need to be considered. The overheads of data encryption and decryption can be very high in both time and processing power. It is also important to note that the cost of encryption is bear by the source system, and that can prove to be very expensive if the system is already loaded system. Network route limitation can narrow options, and may cause restricted access on certain section of the network. Restricted routing also leads to greater dependency on the restricted routes, which may be disastrous if they fail.

Backup and Recovery:

One of the major components of data warehouse that is affected by networks is backup and recovery. Once we choose a backup strategy, the following things affect its performance.

- How the hardware is connected.
- Network bandwidth.
- Backup software used.
- Speed of I/O subsystem.

One kind of the hardware used to take the backup is standalone tape drives. Here we consider the issue of how the tape drives is connected? One way out of many ways is ‘as network available device’. Here we also consider what will be the effect of backup on the high speed interconnect. Connecting the tapes as network available devices id good, but it requires that network be up to the job of the huge transfer rates needed. If the network bandwidth is shared with other processes, we need to ensure that sufficient bandwidth is available during the time we require it. It is very difficult task unless the network is dedicated. Other king of hardware used to take the backup is Silos. Tape silos are large tape storage facilities, which can store and manage thousands of tapes. It is also common for the silo to be connected remotely over a network or a dedicated link. It is important to ensure that the bandwidth of that connection is up to the job.

Parallel technology:

We know that the heart of the computer is the Central Processing Unit (CPU). The speed of the CPU is what decades the amount of work that a computer can perform in a given time. We clearly know multiple CPUs in the same machine allow more than one job to be dealt with simultaneously. Other possibility is a single job running on more than one CPU at the same time. This is simply known as parallelism. This parallel technology is certainly effected by networks, because we have multiple CPUs, we may connect these CPUs either in LAN or WAN. So the performance issues of the networks implicitly affect the performance of data warehouse. To parallelism backups it must be possible to backup multiple parts of the database at the same time. Parallel restore requires the ability to restore multiple parts of the database at the same time. We should achieve it at the data file level. To improve the backup and recovery performance we apply parallelism.

Network storage benefits for Data Warehouse:

The network storage provides several benefits to the data warehouse and business intelligence.

- Availability and Accessibility:

  Facilities such as RAID, hot plug gable devices; cluster support and fast backup and recovery all contribute to data warehouse applications having increased availability and accessibility.

- Scalability:

  The cost-per-megabyte of network storage provides cost effective solution for large amounts of disk storage and I/O required by data warehouse applications.

- Manageability:

  Network storage applications are easy to install, can be administrated from a single web top and can handle data and disk space growth without effecting are house applications. They also make it easy to copy, replace and mirror data warehouse database for testing backup, archiving and disaster recovery.

- Security:

  Network storage method provides a secure method for data warehouse users and applications to share, copy, and backup information anywhere in the corporate network.

For data warehousing, network storage is becoming key storage management solution. For supporting scalable data warehouse applications ranging from small data marts to enterprise data warehouse. Network storage also helps solve many data warehouse operational issues in areas like the data availability and sharing. The operational applications, batch window and database backup and disaster recovery.

Conclusion:

We really need to re-examine the role of the network should be used for finding and requesting information for the repository. The movement of data itself, both in the discovery phase and in the update/refresh phase should be handled outside the operational network. This can be achieved by creating separate networks, dedicated to the data warehouses needs or by tightly coupling the many operations systems (from which the information is extracted) to the information systems where the data is consolidated and queried. This tighter integration is achieved through the data repository itself- a centralized storage unit with high-speed data access characteristics, efficient backup/restore capabilities and embedded data replication tool. That solve many of network bandwidth problems that the data warehouse can impose. Data extraction queries and informational queries can take place over the network without causing undue bottlenecking. All data movement, whether it be to move the data to the central or distributed repositories or to refresh data or to backup and restore data can all be accomplished off-network.

References:

1. Data warehouse architecture – by Ralph kimbal.
2. Data warehouse life cycle tool kit- by Ralph kimbal, Laurna Reeves, Margy Ross, Warren Mornwaile.
3. Data mining concepts & techniques.- by Jaiwei lfan. Michline kamber.
4. Computer Networks by Tanenbaum
5. J. P. Anderson, “Computer Security Threat Monitoring and Surveillance,” Technical Report, James, P. Anderson Company, Fort Washington, PA,
6. A. H. Phyo and S. M. Furnell, “A Detection- Oriented Classification of Insider IT Misuse,” Proceedings of the Third Security Conference, Las Vegas, NV,.
7. D. E. Denning, “An Intrusion Detection Model,” IEEE Trans.
8. S. Kumar, “Classification and Detection of Computer Intrusions,” Ph.D. Dissertation, PurdueUniversity, Lafayette.
9. A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur,and J. Srivastava, “A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection,” Proceedings of the Third SIAM Conference on Data Mining.
10. S. Berinato,“The Global State of Information Security “.