GRIDS IN DATA WAREHOUSING

By Madhu Zode

Oct 2008

ABSTRACT

The main characteristic of any data warehouse is its ability to hold huge volume of data while still offering the good query performance. It is designed to facilitate the analysis of the organization's data and generate the reports based on the analysis. To provide the better performance to analytical queries of business users, companies keeps on adding the hardware resources thereby making the data warehouse system very expensive. Sometimes, the power of these added resources is underutilized. In today's economic climate, the IT organization needs to find ways to maximize the resource utilization they already own which will reduce the cost and improve performance.

To address this issue, the IT organizations have started implementing the cost-effective, reliable and scalable technology in the data warehousing components like ETL tools, databases and reporting tools. One such technology is known as 'Grid computing' which utilizes the computing power of underutilized resources and make the data processing faster.

This paper discusses how IT organizations face challenges when data volume explodes and how grid technology has proven to be an effective solution to handle this data volume and make the data processing faster. Also, this paper will throw light on data warehousing products of some of the vendors who have implemented the grid technology in their databases and ETL tools to make them respond faster.

INTRODUCTION

When we think of a data warehouse as a solution, the first thing which comes to our mind is the technology which will centralize all the data in a single system and at the same time offer good response to the user queries enabling them to take analytical decisions. In today's competitive market, IT organizations are under pressure to increase operational agility, to establish and meet IT service levels and to control costs. They need to think of the cost-effective solution which will make them endure in the market. Whenever designing a new data warehouse, the data architects has to think about the technology which will respond with reliable and secure performance their application need. They need to understand the current business key indicators, query historical information and perform trend analysis to predict future consequences. As a volume of data is made available to large number of business users and analysts making the tactical decisions, scalability and high availability are of paramount importance.

Considering all these needs of data warehousing, most of the IT organizations have started implementing the newly emerging technology which is known as 'Grid Computing' in their data warehouse products. With the introduction of grid-enabled technology in data warehousing, customers can now build the feasible architecture which provides the speed and performance for their application.

WHAT IS GRID?

A grid is typically a collection of low-cost servers connected over a high-speed network in which IT resources such as computer power, storage and network capacity are pooled and shared into a single set of shared services which can be distributed on demand. This leads to maximum utilization of resources already available.

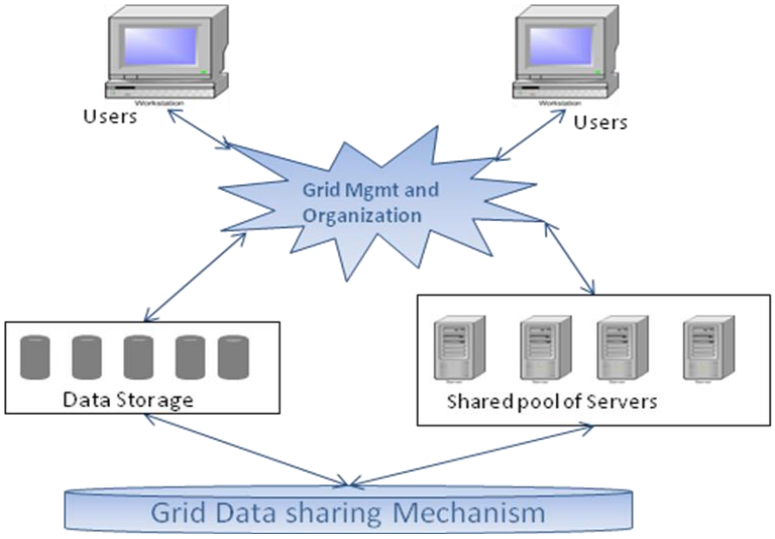

Figure: A simple Grid Implementation

WHY GRID?

Data warehouse is about loading of data from heterogeneous sources like operational system, mainframes, files, etc. which is queried by business users to make analytical decisions.

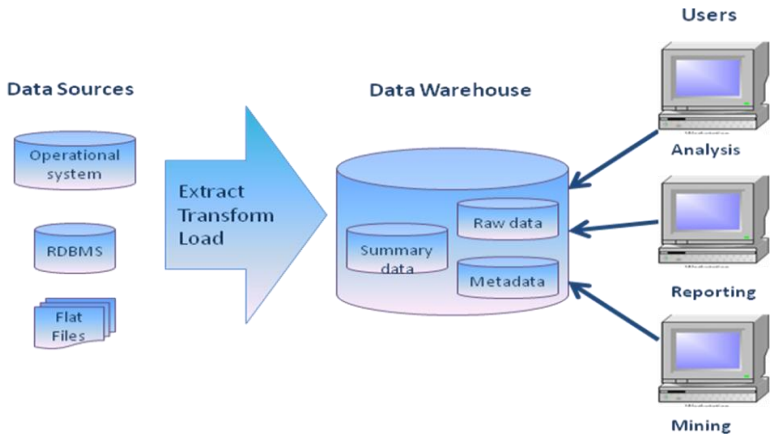

Figure: Data warehouse Architecture

As shown is figure above, the data from various sources is extracted, transformed and loaded in warehouse. This process is called as ETL. In ETL process, loading becomes inefficient as the data volume grows thereby increasing the loading window for data warehouse loading. The data warehouse is loaded during the night time and queried during day time. The ETL process should be so efficient that the data loading should be completed within the given load window or service level agreement (SLA) irrespective of the volume of the data. To overcome the bottleneck of delayed SLA, Organizations increases the hardware resources to make the system effective but making the overall system expensive. IT organizations face the challenges to increase the computational power of already available resources and make the cost-effective, scalable and highly available system.

To handle this data explosion and IT challenges, the grid computing is an innovative solution which provides:

- **Scalability**: By distributing the task over a shared pool of resources, the scalability and performance is improved.
- **Reliability**: In Grid, if any of the server fails then the other server will be used for further processing without failing the job thus proving a reliable structure.
- **Cost Saving**: By utilizing the computing power of unused resources, organizations can optimize their return on investment and lower cost of ownership.
- **Throughput**: Number of users can access the shared pool of resources in order to obtain the best possible response time by maximizing the utilization of all resources available in pool.

With grid computing, groups of independent, hardware and software components can be pooled and shared on demand to meet the changing needs of businesses. Instead of being dedicated to specific applications, grid allows computing resources to be shared, while also making systems highly

scalable and available. The accelerating adoption of grid technology is in direct response to the challenges that IT organizations face with today's rapidly changing and unpredictable business needs.

HOW GRID IS IMPLEMENTED IN DATA WAREHOUSE?

Keeping in mind the pros of Grid technology, most of the organizations have started implementing grids in their ETL tools and databases. By combining grid technology with data warehouses, organizations can reduce processing timelines while lowering the costs. The ETL tools like Informatica, SAS and database like Oracle have implemented the Grid in their products.

Below we will discuss in brief how the grids are implemented in data warehouse and how they are benefited from this technology.

1. **Informatica Corporation**

   Informatica PowerCenter 8 is the latest release of Informatica which harnesses the power of grid computing for greater data integration performance and scalability. The enterprise Grid option delivers the load balancing, dynamic partitioning, parallel processing and high availability to ensure optimal scalability, performance and reliability. The grid technology implemented in Informatica PowerCenter 8 distributes the workload across the available resources doing the proper load balancing thereby increasing the scalability and performance. The High Availability option reduces the system failure chances and provides uninterrupted availability of computer resources.

2. **Oracle Corporation**

   Oracle has implemented the Grid Technology in their Oracle 10g version ('g' for Grid). Oracle has incorporated the fundamentals of grid computing and implemented them in Oracle database, application server and Enterprise manager. In Oracle 10g, the database can balance the workload across a new node with new processing capacity as it gets re-provisioned from one database to another and can abandon the machine when no longer needed – which is on demand sharing of resources.

   Oracle Database 11*g* delivers the benefits of grid computing with more self-management and automation, making it easier to partition and compress tables to store more data and run queries faster and to protect and audit data.

3. **SAS Institute**

   SAS Grid Computing delivers enterprise-class grid computing capabilities that enable SAS applications to automatically leverage grid computing, run faster and takes optimal advantage of computing resources. SAS Grid Manager helps automate the management of SAS Computing Grids with dynamic load balancing, resource assignment and monitoring, and job priority and termination management. Customers can process high-volume SAS programs faster, improve hardware utilization and future proof computing infrastructures while increasing the resilience of SAS applications. Computing resources can be scaled out to cost-effectively add new users and meet fluctuating processing demands.

WHO BENEFITS FROM GRID IN DATA WAREHOUSING?

Grid reduces the bottlenecks occurred during the data loading in data warehouse making the data processing faster. Speeding up the data processing means the data is available to business user for analysis so the business decisions are not delayed and taken in timely manner. Also in grid, as already available resources are being utilized in a proper way, the cost remains the same.

Due to these advantages of grid implementation, the business personnel at every level are getting benefited:

IT Managers and Directors: As grid harnesses the power of underutilized resources providing more computational power; there is no need of extra hardware required which makes it cost-effective.

Data warehouse architects and specialist: Due to availability of grid option in databases and ETL tools, data warehouse architects do not have to worry about the loading time windows even if the data explosion happens in future.

Business Analysts and Decision makers: Due to grid, the data is available to business analysts and decision makers in timely manner and they can take the analytic decisions quickly.

CONCLUSION

As Grid computing harnesses the power of underutilize resources it will have a major impact on productivity and cost improvements at enterprise level. The benefits of grid such as flexibility to manage the resource utilization, high availability, scalability and greater performance at lower cost are attracting most of the IT organizations to make better use of them in their tools and technologies.

While the development and implementation of grid computing is still continuously emerging, it will continue to increase rapidly over next several years. However, the data warehousing will be forerunner of utilizing grid and will benefit from using grid.

REFERENCES

1. SAS Institute, Grid Computing and SAS, by *Merry Rabb* and *Cheryl Doninger*
2. Oracle Corporation, Oracle Grid Computing, May 2008
3. Informatica corporation, Informatica PowerCenter today and in the future, Nov 2006
4. United Devices, 'Grid-enabled Data Management and ETL' , Jun 2007

AUTHOR's BIOGRAPHY

Madhu Zode is data warehousing consultant and worked extensively in ETL architecture, design integration and implementation. Previously she has written white papers on 'Grid Computing' (Published in DMReview Newsletter) and 'ETL Evolution' (Published in ITToolBox). She can be contacted at madhuzode@gmail.com .