GRID COMPUTING – A SOLUTION TO DATA MINING AND BIG DATA CHALLENGES

[1]Dr.C.Jeyabharathi
[1]Assistant Professor & Head
[1]PG Department of Computer Science
[1]Arulmigu Palaniandavar Arts College for Women, Palani

Abstract : Grid computing system is a virtual resource provider in a high computing environment, which is fully capable to utilize or handle the less availability of resources situation with effective load balancing techniques and enhance the proficiency of the interacting users. In the last few years there has been a rapid exponential increase in computer processing power, data storage and communication. But still there are many complex and computation intensive problem, which cannot be solved by super computers. These problems can only be met with a vast variety of heterogeneous resources. The main advantages offered by Grid computing are the storage capabilities and the processing power. Data mining and Big data management are the booming research areas and they face lot of challenges in processing huge amount of unstructured data. This paper discusses the intelligent ideas of combining Grid computing technologies with data mining and big data concepts and deal the challenges of Data mining and Big data technologies. At the end it is presented that how the Grid is going to take an active role in the development of future internet.

IndexTerms **– Grid Computing, Data mining, Big data, Future internet.**

I. INTRODUCTION

Grid computing [5] is a service for sharing computer resources and data storage over the internet. As resource requirements of recent applications increased greatly, Grid systems have gained importance in the last decade. Complex and large-scale problems in science, engineering and business need more powerful computing machines because these problems are more computing-intensive and data-intensive. Grids [6] are geographically distributed platforms with heterogeneous resources in which users can access via a single interface. Grid provides a common resource-access technology and operational services across widely distributed and dynamic virtual organizations. The nature of Grid infrastructure is to integrate large computational and storage resources, data, services and applications from different disciplines. The ability to manage Grid systems depends on the accuracy and availability of the resources like computational power, storage and networking.

As Grid techniques are growing rapidly in recent years, large-scale Grid systems appear to provide flexible, secure, coordinated resource sharing, and problem solving among dynamic virtual organizations [7]. These systems consequently are required to manage a large amount of related resources. In particular, the shared Grid resources can vary from plain desktop systems to clusters and from storage devices to large datasets, even Grid service could be seen as an extension of Grid resource. Therefore, Grid Resource Discovery [11] plays a crucial role in the whole system, and its discovery Models and strategies have a vital influence on the performance of Grid system.

In most organizations, there are large amounts of underutilized computing resources. Most desktop machines are used less than 5 percent of the time. In some organizations, even the server machines can often be relatively idle. Grid computing provides a framework for exploiting these underutilized resources and thus has the possibility of substantially increasing the efficiency of resource usage. The processing resources are not the only ones that may be underutilized. Often, machines may have enormous unused disk drive capacity. Grid computing, more specifically a “data Grid”, can be used to aggregate this unused storage into a much larger virtual data store, possibly configured to achieve improved performance and reliability over that of any single machine.

Locations of resources in Grid environment is the primary necessity which permits the discovery of services across multiple administrative domains. Resource discovery is the process of locating proper resource candidates which are suitable for executing jobs within a reasonable time. The characteristics of the Grid systems make the resource discovery a time consuming process which can decrease the performance of the whole system. Resource discovery in Grid is a challenging issue because characteristics of its resources are heterogeneous, dynamic, various and autonomous.

Data mining [10] is a widely used approach for the transformation of data to useful patterns, aiding the comprehensive knowledge of the concrete domain information. Grid-based data mining applications are likely to use federated and existing Grid technologies which hide the complexity of multiple ownership, domains, and users. However, the algorithmic approaches used at higher levels are very important for scalability and optimization of the distributed processing cost. Well-adapted algorithmic approaches are then of prime importance in the design of data mining applications and frameworks for the Grid.

Massive, fast and diverse data moving quickly everywhere creating what is known as Big Data [1] era. This data becomes very important source for valuable insights and ultimately helping to make more informed decision. However this data with very special attributes can't be managed and processed by the current traditional software systems, which became a real problem.

Two of the main problems that occur when studying Big Data are the storage capacity and the processing power [3]. The increased use and popularity of the internet and availability of high-speed networks have gradually changed the way we do

computing .That is the area where using Grid Technologies can provide help. Grid Computing refers to a special kind of distributed computing which gives solution to the Big data problems. This paper discusses the idea of combining Grid computing technologies with data mining and big data concepts and deals the challenges of Data mining and Big data technologies. This paper also discusses that how the Grid is going to be the Future Internet. The future Internet will play an increasingly important role in the whole society, and it is therefore vital that the new designs take into account the needs of the applications which will be run on the future Internet. Grid computing, thanks to its highly heterogeneous applications, is in an ideal position for providing valuable information about such application requirements.

II. GRID AND DATA MINING

Data mining is a well-established field of computer science concerned with the automated search of large volumes of data for patterns that can be considered knowledge about the data. Data mining has been developed to address the information needs in modern knowledge sectors. To address the main dimensions of complexity the data mining process is in need of reformulation. This leads to the concept of distributed data mining, and in particular to Grid-based data mining. Applying data mining to grand challenge problems brings its own computational challenges.

Grid computing is a good way to address computational challenges. Grid computing is a model of distributed computing that uses geographically and administratively disparate resources. In Grid computing, individual users can access computers and data transparently, without considering location, operating system, account administration, and other details. In Grid computing, the details are abstracted, and the resources are virtualized. Resources in Grid system are heterogeneous, geographically distributed, belong to different administrative domains and apply different management policies**.** The main goal of Grid computing is to enable collaborative and secure resource sharing over multiple organizations, which are geographically distributed.

While Grid technology has the potential to address some of the issues of modern data mining applications, the complexity of the Grid computing environments themselves gives rise to various issues that need to be tackled. Amongst other things, the heterogeneous and geographically distributed nature of Grid resources and the involvement of multiple administrative domains with their local policies make coordinated resource sharing difficult. Ironically, data mining technology could offer possible solutions to some of the problems encountered in complex Grid computing environments. The idea of distributed data mining is that the operational data that is generated in Grid computing environments could be mined to help improve the overall performance and reliability of the Grid. Hence, the integration of Grid computing and Data mining paradigms could look forward to a future of fruitful and meaningful outcomes.

2.1 Complex Data Mining Problems

The complexity of modern data mining problems [4] is challenging the researchers and developers. The sheer scale of these problems requires new computing architectures, as the existing systems can no longer cope. Typical large-scale data mining applications are found in areas such as molecular biology, molecular design, process optimization, weather forecast, climate change prediction, medical environment, astronomy, fluid dynamics, physics, earth science and so on.

Another current complex data mining application is found in weather modeling. Here, the task is to discover a model that accurately describes the weather behaviour according to several parameters. The intrinsic geographic distribution of the data is the great challenge of data mining whereas availability of computing power or massive memory are tackle problems. The mining of medical databases is such an application scenario. The challenge in these applications is to mine data located in distributed, heterogeneous databases while adhering to varying security and privacy constraints imposed on the local data sources.

Other examples of complex data mining challenges include large-scale data mining problems in the life sciences, including disease modeling, pathway and gene expression analysis, literature mining, biodiversity analysis and so on.

2.2 A Solution to Data Mining

Grid technologies, combined with distributed data mining techniques, obtain best results over heterogeneous data and jointly these approaches called as Grid based distributed data mining. Grid based distributed data miners are the software architectures for geographically distributed high-performance knowledge discovery applications such as Knowledge Grid, Data Mining Grid etc. These miners are designed on crest of different Grid environments such as Globus [8] and GridBus [12]. Grid computing also handle the data mining problems by establishing communication among heterogeneous data sources, well-designed resource sharing facilities and provide large storage for data repositories.

III. GRID AND BIG DATA

Big Data is a term defining data that has three main characteristics. First, it involves a great volume of data. Second, the data cannot be structured into regular database tables and third, the data is produced with great velocity and must be captured and processed rapidly. Big Data is a relatively new term that came from the need of big companies like Yahoo, Google, Facebook to analyze big amounts of unstructured data, but this need could be identified in a number of other big enterprises as well in the research and development field.

Big data storage management is one of the most challenging issues for Grid computing environments, since large amount of data intensive applications frequently involve a high degree of data access locality [2]. Grid applications typically deal with large amounts of data. In traditional approaches high-performance computing consists dedicated servers that are used to data storage and data replication. New mechanisms for distributed and big data storage and resource discovery services are suggested in Grid environment by researchers which allows not only sharing the computational cycles, but also share the storage space. The storage can be transparently accessed from any Grid machine, allowing easy data sharing among Grid users and applications. The

concept of virtual ids that, allows the creation of virtual spaces also has been introduced and used. Grid service based storage resources are adopted to stack simple modular service piece by piece as demand grows.

Eventhough Grid computing being beneficial in many ways, current Grid infrastructure cannot support big data, expert are yet to find an precise solution for the database to deal with large volumes of data. Although Grid computing provides technology to overcome the hardware limitation in term of storage space, processing power and memory capacity, Implementation of big data processing and management using Grid computing requires additional techniques for managing the huge data effectively.

Adding the ease of use, ease of maintenance and scalability combining these two technologies seems like a good choice. With the current advances of today's technology in many sectors such as manufacturing, business, science and web application, a variety of data to be processed continues to witness an exponential rise. Efficient management and processing of this data poses an interesting but significant problem. To utilize the numerous benefits of Grid computing, Big data processing and management techniques should be integrated in the current Grid environment.

3.1 A Solution to Big Data

In spite of Grid computing being advantageous in many ways, experts have not yet found an exact solution for the computer database to deal with large volumes of data. The process of introducing distributed caching into the Grid environment might help in solving the issue of big data storage, management and processing. This would also help in increasing the speed of the systems working on this server. Hence the idea of combining the concepts of distributed caching and Grid computing into a single framework will help to increase the efficiency and capability of future computing systems.

IV. Grid – The Future Internet

The internet is just a few decades old, but in that short span of time it has experienced significant changes. It grew out of a hodgepodge of independent networks into a global entity. It serves as a platform for business, communication, entertainment and education. Grid computing applications are by their nature networked and highly heterogeneous. A single Grid has several partners who are connected by the Internet, each providing some resources for other partners to use and using resources provided by other partners. Grid applications place considerable emphasis on authentication and accounting, so that the resource provider ultimately remains in control over her resources. Furthermore, the resources in a Grid application are more often than not extremely heterogeneous, not just in terms of differences in computing power and connectivity, but also in their very nature.

Resources in Grid applications can include scientific instruments and experimental devices, the handling of which is outside the realm of typical network protocols and requires solutions tailored to the device. Such resources may have very specific requirements on how they are accessed, what kind of data they produce, and in general, the notion of "satisfactory service quality" may differ greatly from the traditional network-oriented quality of service point of view. It is exactly this highly heterogeneous nature of Grid applications which makes them especially interesting and important for the development of the future Internet. Because Grid applications are inherently networked, the quality they offer is determined by how well the network is able to meet their specific requirements. The heterogeneous quality requirements of different Grid resources and applications imply that the network architectures and protocols must be extremely flexible in how they allow applications to specify their requirements as well as in how the network manages its resources in order to meet the requirements of the applications.

The future Internet should provide a basis for a multitude of applications, and the key to an efficient and flexible future Internet lies in identifying how different applications use the resources provided by the network. Identifying these application- and user-level demands is therefore a prerequisite for a successful future Internet. The highly heterogeneous nature of Grid computing applications make them ideal candidates for providing different sets of requirements. In order for the future Internet [9] to allow the development of novel applications and foster innovation, the basic architecture needs to be built such that it is flexible enough to accommodate the heterogeneous application requirements. The richer the application pool from which the initial requirements are drawn, the richer the architecture and the protocols will be. Although the goal of the future Internet research is to create a flexible network, the flexibility cannot be achieved without careful studies of what kinds of applications users want to run and what the requirements of those applications are. It is important that the researchers in the Grid community participate actively in the efforts for creating the future Internet. The highly heterogeneous nature of Grid applications makes them ideal candidates for analyzing the requirements of applications running on the future Internet.

V. Conclusion

Today many organizations, companies, and scientific research centres manage and produce large amount of complex data and information. Climate data, astrology data and transaction data are just some examples of massive amounts of digital data repositories that today must be stored and analyzed to find useful knowledge in them. Grid Computing is the suitable technology which handles large amount of heterogeneous data and unstructured patterns. Grid technologies can be combined with Data mining and also with Big data management concepts which will help to increase the efficiency and capability of future computing systems.

References

[1] Acharjya D.P., Kauser Ahmed P, "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.

[2] Alexandru Adrian TOLE, "Big Data Challenges", Database Systems Journal vol. IV, no. 3, 2013

[3] Dan Garlasu et al., "A Big Data implementation based on Grid Computing", 11th RoEduNet International Conference, Sinaia, Romania, 2013.

[4] Emanuel Weitschek et al., "Clinical Data Mining: Problems, Pitfalls and Solutions", Proceedings - International Workshop on Database and Expert Systems Applications, DEXA. 90-94, 2013.
[5] Foster.I and C. Kesselman, "The Grid: blueprint for a new computing infrastructure", Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1999.
[6] Foster.I, C. Kesselman, J. Nick and S. Tuecke, "The physiology of the Grid: an open Grid services architecture for distributed systems integration", In Open Grid Service Infrastructure WG, Global Grid Forum, 2002.
[7] Foster.I, C. Kesselman and S. Tuecke, "The anatomy of the Grid: Enabling scalable virtual organizations", International Journal of Supercomputer Applications, 2008.
[8] Foster.I, "Globus toolkit version 4: Software for service oriented systems", Journal of Computer Science and Technology, vol. 21, pp.513-20, July 2006.
[9] Gavras.A, A. Karila, S. Fdida, M. May, and M. Potts, "Future Internet research and experimentation: The FIRE initiative. Computer Communications Review", 37(3):89–92, July 2007
[10] Hemlata Sahu, Shalini Shrma, Seema Gondhalakar, "A Brief Overview on Data Mining Survey", International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 3.
[11] Mahamat Issa Hassan, Azween Abdullah, "Self-Organizing Grid Resource Discovery", 978-1-4244-2328-6/08, IEEE, 2008.
[12] Rajkumar Buyya and Srikumar Venugopal, "The Gridbus Toolkit for Service Oriented Grid and Utility Computing: An Overview and Status Report", Proceedings of the 1st IEEE International Workshop on Grid Economics and Business Models (GECON), 19-36pp, New Jersey, USA, April 2004.