IJRSET September 2015 Volume 2, Issue 7 Pages: 1-10

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

An Efficient Big Data Analytics Using Grid Framework and ANT Based Algorithm

[1] Ms. B. NETHRA, MCA, [2] Mr. P. PERIYASAMY, MCA, M.Phil, (Ph.D)
[1] Research Scholar, [2] Assistant Professor,
[1] Department of Computer Science, [2] Department of MCA,
[1] Sree Saraswathi Thiyagaraja College, [2] Sree Saraswathi Thiyagaraja College,
[1, 2] Pollachi.

Abstract: -

Big Data is about the volume, variety and velocity of information being generated today and the opportunity that results from efficiently leveraging data for insight and competitive advantage. The Big Data describes a new generation of technologies and architectures designed to economically extract value from these very large and diverse volumes of data by enabling high-velocity capture, discovery, and/or analysis. In big data is term important of three main characteristics. The framework for processing Big Data consists of a number of software tools that will be presented in this thesis, and briefly listed. That is the area where using Grid Technologies can provide help. Grid Computing refers to a special kind of distributed computing. The main purpose of this article is to present a way of processing Big Data using Grid Technologies. For that, the framework for managing Big Data will be presented along with the way to implement it around grid architecture. However, there are many challenges in dealing with big data such as storage, transfer, management and manipulation of big data. Many techniques are required to explore the hidden pattern inside the big data which have limitations in terms of hardware and software implementation. In this thesis, the big data has been implemented using grid framework technologies and ANT based algorithm. The performance of the big data and grid framework technology implementation is evaluated using the parameters of the big data analytics and proved that the grid framework is supports effectively for big data process.

Keywords: - Big Data, Analytics, Grid Technologies, Computing, ANT Algorithm, Framework.

1. INTRODUCTION

Modern computing continues to see technological improvements in raw computing power, storage capability and communication. Despite these improvements, there exist many situations where computational resources fail to keep up with the demands placed on them. This occurs in both scientific and enterprise environments [21]. From a scientific perspective, a good example of this trend is described in [21]. Ten years ago, biologists were content with computing a single molecular structure. Today, biologists want to calculate the structures of complex assemblies of molecules and screen thousands of drug candidates. Scientific projects such as the National Fusion Collaboratory [25] can produce hundreds of megabytes of data in a matter of seconds and require quick analysis of this data. Some demands are not purely computational. The CERN Large Hadron Collider project [10] is expected to produce a few petabytes of data per year by 2006 [21]. From the enterprise perspective, companies

such as Hewlett-Packard [35] and IBM [18] envision a future of on-demand computing and application hosting, thus minimizing the need for clients to purchase and manage their own hardware. Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data. Data size has increased dramatically with the advent of today's technology in many sectors such as manufacturing, business, and science and web application. Some data are structured, semi-structured while others are unstructured and mix with different types of data such as documents, records, pictures and videos (Hall, 2013). A solution for many of theses emerging problems is seen in grid computing. Grid computing is a field inspired by the pervasiveness, ease of use and reliability of the electrical power grid. Taking advantage of big data often involves a progression of cultural and technical changes throughout your business, from exploring new business opportunities to expanding your sphere of inquiry to exploiting new insights as you merge traditional and big data analytics. The journey often begins with traditional enterprise data and tools, which yield insights about everything from sales forecasts to inventory levels. The data typically resides in a data warehouse and is analyzed with SQL-based business intelligence (BI) tools. Much of the data in the warehouse comes from business transactions originally captured in an OLTP database. While reports and dashboards account for the majority of BI use, more and more organizations are performing "what-if" analysis on multi-dimensional databases, especially within the context of financial planning and forecasting. These planning and forecasting applications can benefit from big data but organizations need advanced analytics to make this goal a reality. For more advanced data analysis such as statistical analysis, data mining, predictive analytics, and text mining, companies have traditionally moved the data to dedicated servers for analysis. Exporting the data out of the data warehouse, creating copies of it in external analytical servers, and deriving insights and predictions is time consuming. It also requires duplicate data storage environments and specialized data analysis skills. Once you've successfully built a predictive model, using that model with production data involves either complex rewriting of the model or the additional movement of large volumes of data from a data warehouse to an external data analysis server. Big data analysis involves making "sense" out of large volumes of varied data that in its raw form lacks a data model to define what each element means in the context of the others. There are several new issues you should consider as you embark on this new type of analysis:

- **Discovery –** In many cases you don't really know what you have and how different data sets relate to each other. You must figure it out through a process of exploration and discovery.
- **Iteration –** Because the actual relationships are not always known in advance, uncovering insight is often an iterative process as you find the answers that you seek. The nature of iteration is that it sometimes leads you down a path that turns out to be a dead end.
- **Flexible Capacity** – Because of the iterative nature of big data analysis, be prepared to spend more time and utilize more resources to solve problems.
- **Mining and Predicting** – Big data analysis is not black and white. You don't always know how the various data elements relate to each other. As you mine the data to discover patterns and relationships, predictive analytics can yield the insights that you seek.
- **Decision Management** – Consider the transaction volume and velocity. If you are using big data analytics to drive many operational decisions (such as personalizing a web site or prompting call center agents about

the habits and activities of consumers) then you need to consider how to automate and optimize the implementation of all those actions. With new data and new data sources comes the need to acquire new skills. Sometimes the existing skill set will determine where analysis can and should be done. When the requisite skills are lacking, a combination of training, hiring and new tools will address the problem. Since most organizations have more people who can analyze data using SQL than using MapReduce, it is important to be able to support both types of processing. Data security is essential for many corporate applications. Data warehouse users are accustomed not only to carefully defined metrics and dimensions and attributes, but also to a reliable set of administration policies and security controls. These rigorous processes are often lacking with unstructured data sources and open source analysis tools. Pay attention to the security and data governance requirements of each analysis project and make sure that the tools you are using can accommodate those requirements.

2. GRID COMPUTING FRAMEWORK

The purpose of this section is to clarify of some of the concepts and software components associated with grid computing.

2.1. Resources

A grid is a collection of machines typically referred to as “nodes”, “resources”, “clients”, “hosts”, and other similar terms [5]. These collections contribute any combination of resources to the grid as a whole and may have associated user-based access and usage restrictions [5].

2.2. Scheduling, Reservation and Scavenging

Scheduling refers to the process of selecting the machines appropriate for executing a particular job. In a simple grid, the process of scheduling might involve users manually selecting the machines suitable for running jobs and executing commands that send these jobs to the machines. A more advanced grid would likely contain a job scheduler capable of automatically performing these tasks on behalf of the user. In order to address these two problems, grids typically offer dedicated machines that cannot be pre-empted once the machine is assigned to a job. This enables schedulers to compute the approximate completion time for a set of jobs whose run-time characteristics are known [5]. In addition, the use of resource reservation can be used in order to support deadlines and guarantee performance requirements (these are referred to as Quality of Service (QoS)) of jobs running on the grid. Resource reservation is much like booking an appointment. It involves reserving the use of a resource for a specific amount of time at a certain date and time.

2.3. Architectural Models

Several grid architectural models exist for solving different types of problems. Some of these models take advantage of additional processing resources, whereas others are designed to support collaboration between various organizations [5]. Most of the models fall into one of two categories: data grids or computational grids. A data grid focuses on secure access to distributed, heterogeneous pools of data [5]. The purpose is to harness storage, data and network resources located in distinct administrative domains, schedule resources efficiently, and provide high speed and reliable access to data while respecting local and global policies governing how data can be utilized [5]. A computational grid aggregates the processing power of a distributed collection of heterogeneous systems [5]. These grids are used in situations where an organization requires more computing capacity than is currently available and they are willing to modify their applications in order to take advantage of parallelization (if they do not do so already).

Typical applications include the calculation of mathematical equations, derivatives, portfolio valuation, and simulation [5].

3.1.4. Characteristics

A grid can be characterized by four main attributes:

- Heterogeneous: A grid aggregates a variety of different resources types, whether hardware or software, that encompass a wide range of technologies [5].
- Scalable: The number of resources shared on a grid might increase from a few hundred, to a few thousand, to a few million. A grid must be able to handle this change in number in an efficient manner without causing any significant performance decrease to the overall system [5].
- Dynamic nature: The number of resources shared on a grid often fluctuates as new resources are added and old ones are removed. Also, due to the intricacies of the grid and the high number of resources being shared, the probability of a resource failing is high.
- Encompasses multiple administrative domains: Resources in a grid can be geographically dispersed and are owned and operated by a variety of different individuals and organizations.

3. GRID ARCHITECTURE

Building a grid infrastructure requires the design and development of protocols and services which address issues of security, resource aggregation, resource discovery, resource selection, job scheduling, job execution management, and more. A basic template representing the architectural layers of a grid infrastructure is shown in Figure 3.1,

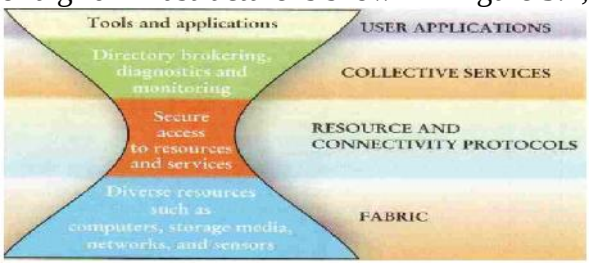

Fig 3.1: - Layered Grid Architecture

The key components involved in the architecture are:

- **Fabric:** This consists of all the distributed resources, owned by different individuals and organizations, shared on the grid. This includes workstations, resource management systems, storage systems, specialized devices, etc.
- **Resource and Connectivity Protocols:** This contains core communication and authentication protocols that provide secure mechanisms for verifying the identity of users and resources and allow data to be shared between resources [24].
- **Collective Services:** This contains Application Programming Interfaces (APIs) and services that implement interactions across collections of resources [24]. This includes directory and brokering services for resource allocation and discovery, monitoring and diagnostic services, application scheduling and execution, and more.
- **User Applications:** This contains programming tools and user applications that depend on grid resources and services during their execution.

3.2.1. Grid Topologies

A grid topology refers to the structure and organization of resources within a grid infrastructure. There are three basic topologies associated with grid computing: intra-grids, extra-grids and inter-grids. These topologies are shown in Figure 3.2,

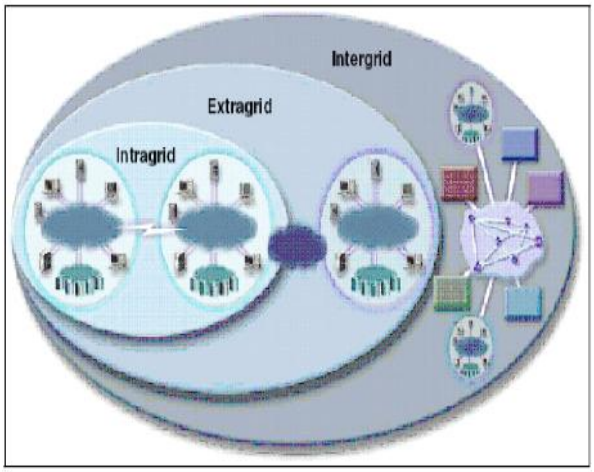

Fig 3.2: - Grid Topologies

Intra-grids are the simplest of the three grid topologies. An intra-grid is composed of a basic set of services within a single organization [5]. An intra-grid is characterized as having a single security provider and having a single environment within a single network [5]. Data and resources in an intra-grid environment are confined to a single organization. Extra-grids are more complicated in that they involve a consolidation of different intra-grids. An extra-grid is characterized by dispersed security, multiple organization and remote/wide area network (WAN) connectivity [5]. Security is an increased concern because data passes beyond organizational boundaries. Resources are more heterogeneous (due to multiple organizations being involved), more dynamic in nature (organizations do not have control over each other's resources) and typically require policies in order to control utilization of the resources. Data and resources in an extra-grid environment are confined to an organization and the partners it wishes to share with. Inter-grids are the most complicated of the three topologies. Inter-grids have the same characteristics as extra-grids, except that data and resources within the environment are global and are available to the public. Regardless of the topology being used, the user is still presented with the same view of the system.

4. PROBLEM STATEMENT

Big data has the power to dramatically change the way institutes and organizations use their data. Transforming the massive amounts of data into knowledge will leverage the organizations performance to the maximum. Scientific and business organizations would benefit from utilizing big data. However, there are many challenges in dealing with big data such as storage, transfer, management and manipulation of big data. Many techniques are required to explore the hidden pattern inside the big data which have limitations in terms of hardware and software implementation. This paper presents a framework for big data clustering which utilizes grid technology and ant-based algorithm.

5. OVERVIEW OF THE PROPOSED SYSTEM

Ant colonies provide a means to formulate some powerful nature-inspired heuristics for solving the clustering problems. Several clustering methods based on ant behavior have been proposed in the literature. The main purpose of this article is to present a way of processing Big Data using Grid Technologies. For that, the framework for managing Big Data will be presented along with the way to implement it around grid architecture. This paper we implement how to control volume, velocity and data storage, Main advantages offered by Grid computing are the storage capabilities and the processing power and the main advantages of using Hadoop, especially HDFS, are reliability (offered by replicating all data on multiple Data Nodes and other mechanism to protect from failure), the scheduler's ability to collocate the jobs and the data offering high throughput for data for the jobs processed on the grid. The nature inspired methods like ant-based clustering techniques have found success in solving clustering problems. They have received special attention from the research community over the recent years. It is because these methods are particularly suitable to perform exploratory data analysis, and also because there is still a lot of investigation to perform on this field – the research nowadays concentrates on improving performance, stability, convergence, speed, robustness and other key features that would allow us to apply these methods in real applications.

6. ANT BASED ALGORITHM

Ant Colony optimization (ACO) is an algorithm modeled on swarm intelligence, and it constitutes some Meta heuristic

optimizations. The algorithm was initially proposed by Marco Dorigo in 1992. The ACO algorithm is a probabilistic technique for solving computational problems, which can be reduced to finding good paths through graphs. Ant algorithms were inspired by the observation of real ant colonies. Ants are social insects, that is, insects that live in colonies and whose behavior is directed more to the survival of the colony as a whole than to that of a single individual component of the colony. While walking from food sources to the nest and vice versa, ants deposit on the ground a substance called pheromone, forming in this way a pheromone trail. Ants can smell pheromone and, when choosing their way, they tend to choose, in probability paths marked by strong pheromone concentrations. The ants move in a straight line from nest to food source (Figure 1 (a)). At the next stage, assume that there is an obstacle (Figure 1 (b)). In this case, to avoid the obstacle initially each ant chooses to left or right at random (Figure 1 (c)). Let us assume that ants move at the same speed depositing pheromone in the trail uniformly. However, the ants that, by chance, select to turn left will reach the food source sooner, whereas the ants that turn right will follow a longer path. The intensity of pheromone deposited on shorter path is more than the other path. So ants will we increasingly guided to move on the shorter path (Figure 1 (d)). The intensity of deposited pheromone is one of the most important factors for ants to find the shortest path.

Figure 6.1: - A path between their nest and foods source

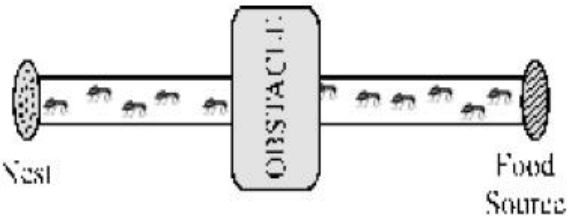

Figure 6.2: - Encountering obstacles of ants

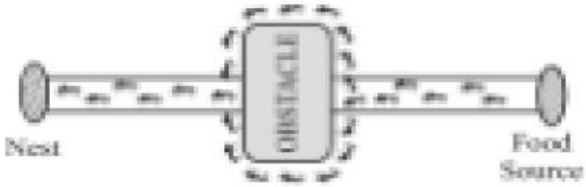

Figure 6.3: - Selecting of Ants

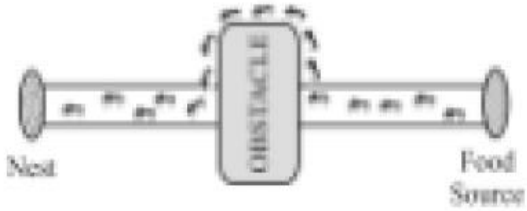

Figure 4.4: - Finding shortest path of ants

A study in [3] proposed a new algorithm that is based on an echo intelligent system, autonomous and cooperative ants. In this proposed algorithm, the ants can procreate and also can commit suicide depending on existing condition. Ant level load balancing is proposed to improve the performance of the mechanism. Ants are created on demand during their lives adaptively to achieve the grid load balancing. The ants may bear offspring when they detect the system is drastically unbalanced and commit suicide when they detect equilibrium in the environment. The ants will care for every node visited during their steps and record node specifications for future decision making. Theoretical and simulation results indicate that this new algorithm surpasses its predecessor. However, the pheromone values were not updated in this proposed algorithm which enables the assignment of jobs to the same resource.

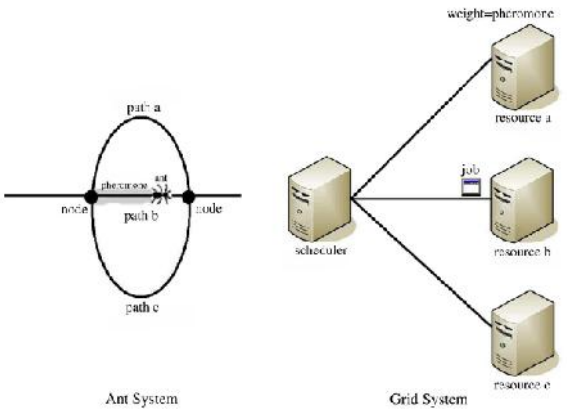

Figure 6.5: - Mapping between the Ant System and the Grid System

This inspired the discovery of ACO algorithm. This algorithm uses a colony of artificial ants that behave as cooperative agents in a mathematical space where they are allowed to search and reinforce pathways (solutions) in order to find the optimal ones. This approach which is population based has been successfully applied to many NP-hard optimization problems.

```
1: Begin
2: Initialization phase
3: Randomly scatter all data on the grid
4: While (termination condition not met) do
5:    Each ant randomly picks up one data item
6:    Each ant randomly placed on the grid
7:    For each ant (i=1, …, n) do
8:          While (ant[i] carries item)
9:                ant[i]:= move randomly on the grid
10:               if (ant[i] decide to drop item) do
11:                     ant[i]:= drop item
12:         End while
13:   End for
14: End while
15: End
```

The algorithm‘s basic principle focuses on agents where the agents represent the ants that randomly move around in their environment which is a squared grid with periodic boundary conditions. While ants wandering around in their environment, they pick up the data item that are either isolated or surrounded by dissimilar ones. The picked item will be transported and dropped by ants to form a group with a similar neighborhood items base on similarity and density of data items. The probability of picking an element increases with low density and decreases with the similarity of the element. The idea behind this type of aggregation pheromone is the attraction between data items and artificial ants. Small clusters of data items grow by attracting ants to deposit more items.

7. IMPLEMENTATION

High dimensional data are data characterized by few dozen to many thousands of dimensions. And any dataset represent able under a relational model is chosen as a High Dimensional Dataset. According to that the following six different datasets were used, it is worth noting that the 20NG, Sports, Health, Society, and Local News.

Category	No. of User Profiles
20NG	412
Sports	300
Health	669
Society	442
Local News	254

Table 5.1: - The Category of Dataset

The datasets used have different characteristics in the sense of the vocabulary size, and category distribution. Each document 1000’s of attributes or dimensions inters related and theses are categorized using the Bigdata Analytics Techniques.

7.1. Performance Evaluation Parameters

The following performance parameters are commonly used in privacy protection technique evaluation. The existing

approach is compared with proposed scheme using these evaluation parameters. The performance of the TC process can be measured by one or more of the following methods:

7.1.1. Recall and Precision

They are two well known measures of effectiveness in text mining. While Recall is a measure of correctly predicted documents by the system among the positive documents, Precision is a measure of correctly predicted documents by the system among all the predicted documents. The system is evaluated in terms of precision, recall and Fmeasure. Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

$$precision = \frac{number\ of\ correct\ results}{number\ of\ all\ returned\ results}$$

$$recall = \frac{number\ of\ correct\ results}{total\ number\ of\ actual\ results}$$

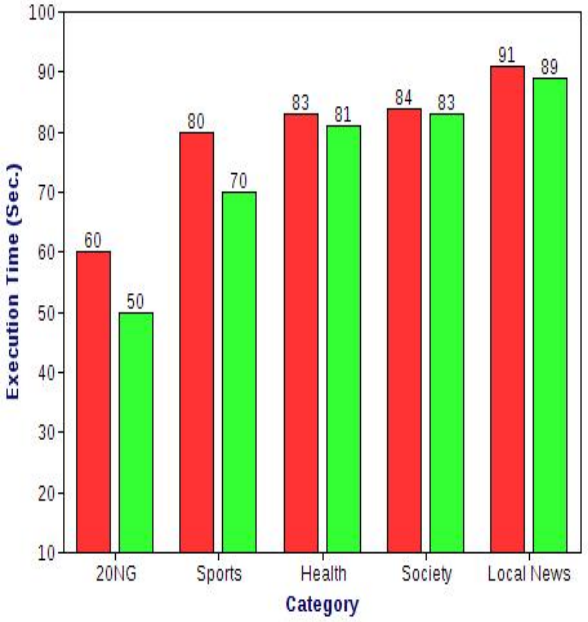

Figure 7.1: - Evaluation of Recall using Big Data ANT Algorithm

7.1.2. F-Measure

F-measure combines precision and recall and is the harmonic mean of precision and
recall.

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall}$$

Several experiments were conducted with different query documents and the precision, recall and F-measure of the output was calculated. This higher improvement in precision value can compromise for the very small percentage of drop in the recall value. Moreover, the F-measure which combines precision and recall is much improved for similarity than existing system.

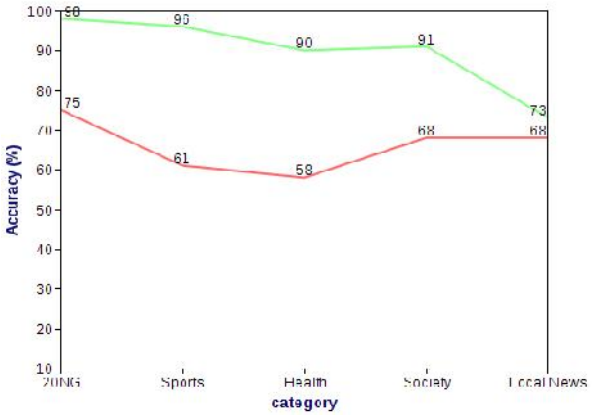

Figure 7.2: - Evaluation of Precision using Big Data ANT Algorithm

7.1.3. Distortion

It is measured with the assistance of examination between Original dataset and changed dataset. Every tuple Xi in unique dataset [10], of m columns and each one column is of n characteristics , which is changed into Yi in altered dataset is utilized to register contortion in that tuple by ascertaining disparity between them through Euclidean separation by the equation.

$$D(Xi, Yi) = \left[\sum_{k=1}^{n} |Xik - Yik|^{\frac{1}{2}}\right]^2 \qquad Distortion = \frac{1}{mn}\sum_{i=1}^{m}\left[\sum_{k=1}^{n} |Xik - Yik|^{\frac{1}{2}}\right]^2$$

They anticipate that this cooperation will provide for them leverage over whatever remains of their rivals who did not partake in

the coordinated effort. All things considered, the working together associations may not have any desire to unveil some touchy standards stowed away in their relating information sets.

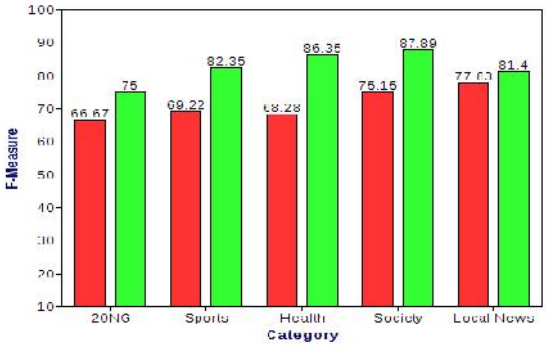

Figure 7.3: - Evaluation of F-Measure using Big Data ANT Algorithm

8. CONCLUSION

Big data has the power to dramatically change the way institutes and organizations use their data. Transforming the massive amounts of data into knowledge will leverage the organizations performance to the maximum. Scientific and business organizations would benefit from utilizing big data. However, there are many challenges in dealing with big data such as storage, transfer, management and manipulation of big data. Many techniques are required to explore the hidden pattern inside the big data which have limitations in terms of hardware and software implementation. Implementing big data solutions required an infrastructure which supports the scalability, distribution and management of data (Kim, 2012). Thus this study proposes grid technology to overcome the hardware limitation in term of storage space, processing power and memory capacity. For algorithm scalability, ant-based clustering algorithm is proposed. A framework for the clustering of big data using grid computing and ant colony algorithm has been proposed. The grid concept is to enable the storage of data in distributed databases across a wide geographical area while ant-based algorithm is for the clustering of big data. Ant-based algorithm has many advantages to be used in big data mining because it has the ability to scale with the size of the data set, prior knowledge of the number of expected clusters is not needed and easy to integrate with clusters ensemble model. Big data analysis opens the door for many research areas and one of the most important areas is the data security. In future, we will study whether there are any other situations which we do not take into account on our definitions of the pheromone indicator or the pheromone update functions.

REFERENXCES

[1] I. Foster, C. Kesselman, and S. Tuecke. “The anatomy of the Grid: Enabling Scalable Virtual Organizations”. International Journal of Super computing Applications, pp.200-222, Fall. 2001.

[2] I. Foster and C. Kesselman, “Globus: A Meta computing Infrastructure Toolkit,” Int’l J Super computer App., 1997, pp. 115–128.

[3] Abramson D, Giddy J, Kotler L. “High performance parametric modeling with Nimrod/G: Killer application for the global Grid “ Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS 2000). Cancun, Mexico, 1–5 May 2000. IEEE Computer Society Press: Los Alamitos, CA, 2000.

[4] Buyya R, Abramson D, Giddy J. “Nimrod/G: An architecture for a resource management and scheduling system in a global computational Grid”. Proceedings of the 4th International Conference and Exhibition on High Performance Computing in Asia-Pacific Region (HPC ASIA 2000), Beijing, China, May 2000. IEEE Computer Society Press: Los Alamitos, CA, 2000.

[5] Casavant TL, Kuhl JG. “A taxonomy of scheduling in general purpose distributed computing”. IEEE Transactions on Software Engineering 1988; 14(2).

[6] Feitelson DG, Rudolph L (eds.). Proceedings of the 5th IPPS/SPDP’99 Workshop “Job Scheduling Strategies for Parallel Processing” (JSSPP 1999), San Juan,

Puerto Rico, April 1999 (Lecture Notes in Computer Science, vol. 1659). Springer: Heidelberg, 1999.
[7] D. Fernandez-Baca(1989) "Allocation Modules to processors in a Distributed System", IEEE Transactions on Software Engineering. Vol.15(11): Pages 1427-1436
[8] Z. Xu, X. Hou and J. Sun, "Ant Algorithm-Based Task Scheduling in Grid Computing", Electrical and Computer Engineering, IEEE CCECE 2003, Canadian Conference, 2003.
[9] E. Lu, Z. Xu and J. Sun, "An Extendable Grid Simulation Environment Based on GridSim", Second International Workshop, GCC 2003, volume LNCS 3032, pages 205–208, 2004.
[10] H. Yan, X. Shen, X. Li and M. Wu, "An Improved Ant Algorithm for Job Scheduling in Grid Computing", In Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, 18-21 August 2005.
[11] Li Liu, Yi Yang, Lian Li and Wanbin Shi, " Using Ant Optimization for super scheduling in Computational Grid, IEEE proceedings of the 2006 IEEE Asia-pasific Conference on Services Computing (APSCC' 06)
[12] R. Armstrong, D. Hensgen, and T. Kidd, "The relative performance of various mapping algorithms is independent of sizable variances in run-time predictions," in 7th IEEE Heterogeneous Computing Workshop, pp. 79–87, Mar. 1998.
[13] Gong L., Sun X.H., Waston E.: "Performance Modeling and Prediction of Non-Dedicated Network Computing", IEEE Transaction on Computer, **51** 9 (2002) 1041–1055.
[14] Maheswaran M., Ali S., Siegel H.J., Hensgen D., Freund R.: "Dynamic Mapping of a Class of Independent Tasks onto Heterogeneous Computing Systems", 8th IEEE Heterogeneous Computing Workshop (HCW'99), San Juan, Puerto Rico, (1999) 30–44.
[15] Pinedo M.:Scheduling: "Theory, Algorithms and Systems", Prentice Hall, Englewood Clifts, NJ, (1995).
[16] Braun, T.D., Siegel, H.J., Beck, N., Boloni, L.L., Maheswaran, M., Reuther, A.I., Robertson, J.P., et al. (2001) 'A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems', J. of Parallel and Distr. Comp., Vol. 61, No. 6, pp.810–837
[17] Stefka Fidanova and Mariya Durchova," Ant Algorithm for Grid Scheduling Problem", Large Scale Computing, Lecture Notes in Computer Science No. 3743, Springer, germany, 2006, 405-412.
[18] Foster, I., Kesselman, C., M.Nick, J., Tuecke, S.: The physiology of the grid an open grid services architecture for distributed system integration. Technical report, Globus Proyect Draft Overwiev Paper (2002)