An Efficient Big Data Analytics in Grid Framework Using Ant Colony Optimization (ACO) Algorithm

[1]Dr. N. Kamalraj, [2]Dr. S. Poongodi
[1]Assistant Professor, Department of Computer Science, Dr.SNS Rajalakshmi College of Arts and Science (Autonomous), Coimbatore, India.
[2]Assistant Professor, Department of Information Technology, Dr.SNS Rajalakshmi College of Arts and Science (Autonomous), Coimbatore, India.

ABSTRACT: - Big Data is about the volume, variety and speed of knowledge being generated today and the potential that arises from the effective use of data for insight and competitive advantage. Big data represents a new generation of technologies and architectures designed to economically extract value from these very broad and complex data volumes by allowing high-speed collection, discovery and/or analysis. In big data, the word is significant for three key characteristics. The architecture for the processing of big data consists of a variety of software resources which will be discussed in this study and briefly outlined. This is the field where the use of grid technology can help. Grid computing refers to a special form of distributed computing. The main goal of this article is to present a way of processing big data using grid technologies. In order to do this, the structure for handling big data will be presented along with the way it will be applied around the grid architecture.

But there are so many challenges in dealing with big data, such as storing, transferring, handling and manipulating big data. Many techniques are needed to explore hidden trends within big data that have limitations in hardware and software implementation. Big data has been applied in the proposed work using grid framework developments and the Ant Colony Optimization (ACO) algorithm. The performance of the big data and grid framework system integration is evaluated using the big data analytics parameters and has shown that the grid framework is effectively supported by the big data process.

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

1. INTRODUCTION

Modern computing continues to see technical advances in raw computing power, storage space and connectivity. Big data refers to datasets whose complexity is beyond the capacity of traditional database software tools to record, store, handle, and analyse. This definition is deliberately arbitrary and includes a shifting definition of how large a dataset has to be to be considered big data. Many concepts of big data concentrate on the scale of the storage data. Size matters, but there are other essential attributes of big data, namely data variety and communication bandwidth. Te three Vs of big data (volume, variety, and velocity) are a thorough concept, and they smash the misconception that big data is just about data volume. With the emergence of today's technology, data size has increased dramatically in many industries such as manufacturing, industry, and research, and web applications. Some data are organized, semi-structured, while others are unstructured and mix with various types of data, such as documents, records, images and videos [19-20]. Big Data analytics is where sophisticated analytical methods are applied to big data.

A solution to most of these emerging problems can be seen in grid computing. Grid computing is a field motivated by the pervasiveness, ease of use and reliability of the electrical grid [1]. Taking advantage of big data also requires the advancement of cultural and technological changes in your business, from pursuing new business opportunities to widening your area of inquiry to taking advantage of new insights as you combine conventional and big data analytics.

The journey also starts with conventional business data and resources that provide information from revenue estimates to inventory levels. Data usually resides in a data warehouse and is analysed using SQL-based Business Intelligence (BI) software. Most of the data in the warehouse comes from business transactions that were initially registered in the OLTP database. Although reports and dashboards account for the majority of BI use, more and more organisations are conducting "what-if" analysis on multi-dimensional databases, particularly in the context of financial planning and forecasting. These planning and forecasting applications will benefit from big data, but organisations need advanced analytics to make this aim a reality. For more sophisticated data processing, such as statistical analysis, data mining, predictive analysis, and text mining, businesses have historically transferred data to dedicated analysis servers. Exporting data from the data warehouse, making copies of data from external analytical servers, and drawing up insights and forecasts is time consuming. Duplicate data storage environments and advanced data analysis skills [18] are also needed. Once you have successfully developed a predictive model, using the production data model requires either a complicated rewrite of the model or an additional movement of massive data volumes from a data warehouse to an external data analysis server.

New data and new data sources make it possible to learn new skills. Often the current skill set will decide where the research can and should be conducted. Where the necessary skills are missing, a combination of preparation, recruiting and new resources will solve the issue. Since most organisations have more people who are able to analyse data using SQL than using MapReduce, it is important to be able to support both types of processing. Data security is necessary for a range of enterprise applications [7-8]. Data warehouse users are accustomed not only to carefully specified measurements and dimensions and characteristics, but also to a reliable collection of management policies and security controls. These robust methods also neglect unstructured data sources and open source research tools. Pay attention to the security and data governance criteria of each research project and ensure that the tools you use will fulfil those requirements.

2. GRID COMPUTING FRAMEWORK

The aim of this section is to explain some of the concepts and software components relevant to grid computing [4].

2.1. Resources

A grid is a set of machines commonly referred to as "nodes," "resources," "clients," "hosts" and similar words. These collections apply to some mix of services in the grid as a whole and may have related user-based access and use restrictions [5].

2.2. Scheduling, Reservation and Scavenging

Scheduling refers to the method of choosing machines that are suitable for the execution of a specific job. In a simple grid, the scheduling process can involve users manually selecting machines that are suitable for running tasks and executing commands that send these jobs to machines. A more sophisticated grid will likely include a work scheduler capable of executing these tasks automatically on behalf of the user.

In order to resolve these two concerns, grids usually provide dedicated machines that cannot be pre-empted until the machine is assigned to a job. This enables schedulers to measure the estimated completion period for a group of jobs whose run-time characteristics are known[6]. In addition, the use of the resource reserve may be used to support time-limits and to guarantee performance specifications (i.e. Quality of Service (QoS)) for grid-based jobs. Resource booking is a lot like booking an appointment. It includes reserving the use of a resource for a fixed period of time at a given date and time.

2.3. Architectural Models

Several grid architectural models exist to solve various types of problems. Some of these models use additional processing resources, while others are structured to facilitate collaboration between different organizations [9]. Most models fall into one of two categories: data grids or computer grids.

A data grid focuses on safe access to distributed, heterogeneous data pools. The goal is to harness storage, data and network resources located in different administrative jurisdictions, plan resources efficiently and provide high-speed and secure access to data while respecting local and global policies regulating how data can be used. The computing grid aggregates the processing power of a distributed array of heterogeneous systems [10]. These grids are used in situations where an entity needs more computing power than is currently available and are willing to change their applications in order to take advantage of parallelization. Typical applications include the calculation of mathematical equations, derivatives, portfolio valuation, and simulation.

2.4. Characteristics

The grid can be identified by four key attributes [2-3]:

- **Heterogeneous**:-The grid aggregates a number of different resource types, whether hardware or software, encompassing a wide range of technologies.
- **Scalable**:-The amount of resources exchanged on the grid could be increased from a few hundred to a few thousand to a few million. The grid must be able to accommodate this shift in number in an effective manner without causing any major decrease in overall efficiency.
- **Dynamic nature**:-The number of resources exchanged on the grid also fluctuates as new resources are introduced and old ones removed. Also, due to the intricacies of the grid and the high number of resources exchanged, the likelihood of a resource failing is high.
- **Encompasses multiple administrative domains**: - The services in the grid may be geographically distributed and are owned and managed by a number of individuals and organisations.

3. GRID ARCHITECTURE

Building a grid infrastructure involves the design and implementation of protocols and services that address security, resource aggregation, resource discovery, resource selection, job scheduling, job execution, and more. The basic template for the architectural layers of the grid infrastructure is shown in Figure 3.1.,

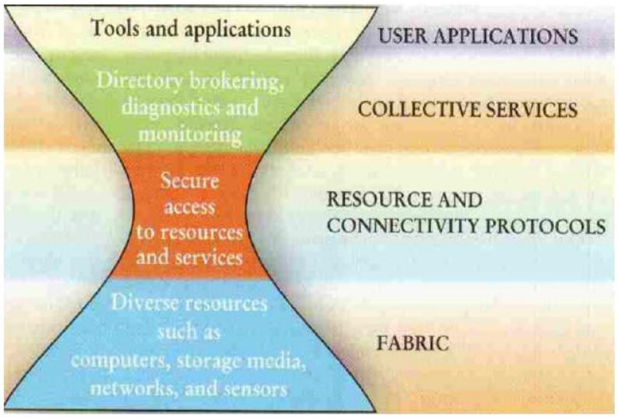
Tools and applications
USER APPLICATIONS
Directory brokering, diagnostics and monitoring
COLLECTIVE SERVICES
Secure access to resources and services
RESOURCE AND CONNECTIVITY PROTOCOLS
Diverse resources such as computers, storage media, networks, and sensors
FABRIC

Fig 3.1: - Layered Grid Architecture

The key components involved in the architecture are:

- **Fabric:** It consists of all the services distributed, owned by various individuals and organisations, which are shared on the grid. This includes workstations, resource management systems, storage systems, specialized equipment, etc..
- **Resource and Connectivity Protocols:** It includes core communication and authentication protocols that provide security mechanisms for verifying the identity of users and resources and allow data to be exchanged between resources.
- **Collective Services:** It includes Application Programming Interfaces (APIs) and services that incorporate interactions through resource sets. This includes directory and brokering services for resource allocation and discovery, monitoring and diagnostic services, application preparation and execution, and more.
- **User Applications:** It includes programming tools and user applications that rely on the resources and services of the grid during their execution.

3.1. Grid Topologies

The grid topology refers to the implementation and organisation of resources within the network infrastructure. There are three basic grid computing topologies: intra-grids, extra-grids and inter-grids. These topologies can be found in Figure 3.2.,

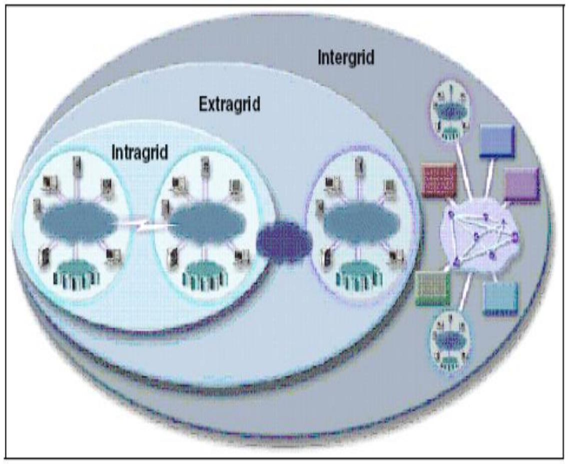

Fig 3.2: - Grid Topologies

Intra-grids are the most straightforward of the three grid topologies. The intra-grid consists of a simple collection of resources within a single organization [11]. The intra-grid is defined as having a single protection provider and a single environment within a single network. Data and services in the intra-grid setting are limited to a single entity.

Extra-grids are further complicated by the fact that they require the consolidation of various intra-grids. The extra-grid is characterized by scattered defence, multiple organisation and remote/wide area network (WAN) connectivity [12]. Security is a growing issue because data goes beyond organizational boundaries. Resources are more heterogeneous (due to the presence of many organisations), more competitive in nature (organisations may not have influence over each other's resources) and usually need policies to control resource usage.

Inter-grids are the most complex of the three topologies. Inter-grids have the same features as extra-grids, except that environmental data and services are national and open to the public. Regardless of the topology used the consumer still has the same view of the structure.

4. PROBLEM STATEMENT & OVERVIEW OF THE PROPOSED WORK

Big data has the potential to radically change the manner in which institutions and organisations use their data. Transforming vast quantities of data into information would improve the efficiency of organisations. Scientific and business organisations will benefit from the use of big data. However there are many difficulties in dealing with big data, such as storing, transferring, handling and manipulating big data. Many techniques are needed to explore hidden trends within big data that have limitations in hardware and software implementation. This paper provides a framework for large data clustering that uses grid technology and ant-based algorithms.

Ant Colonies have a means of formulating some strong nature-inspired heuristics to solve clustering problems. Several clustering approaches based on behaviour have been proposed in the literature. The main goal of the proposed work is to present a way of processing Big Data using Grid Technologies [16]. In order to do this, the structure for handling Big Data will be presented along with the way it will be applied around the grid architecture. This paper implements how to monitor volume, speed and data storage, the main advantages provided by Grid computing are storage capabilities and processing power, and the main advantages of using Hadoop, especially HDFS, are reliability, the ability of the scheduler to collect jobs and the high throughput data for the jobs processed on the grid.

Nature-inspired methods such as ant-based clustering techniques have been successful in solving clustering problems. In recent years, they have received special attention from the research community. It is because these methods are especially suited for exploratory data analysis, and also because there is still a lot of research to be done in this field – research is currently based on improving efficiency, stability, convergence, speed, robustness and other key features that would enable us to apply these methods in real applications.

6. ACO ALGORITHM

Ant Colony Optimization (ACO) is a swarm intelligence-modeled algorithm that constitutes certain Meta heuristic optimizations. The algorithm was originally proposed by Marco Dorigo in 1992.

The ACO algorithm is a probabilistic technique used to solve computational problems, which can be reduced to finding good paths through graphs. Ant algorithms were inspired by the observation of the real and the colony. Ants are social insects, that is, insects that live in colonies and whose action is oriented more to the survival of the colony as a whole than to the survival of the colony as a whole. When walking from food sources to the nest and vice versa, the ants deposit a substance called pheromone on the ground, thereby creating a pheromone trail. Ants can detect pheromones and in their choice of pathways, prefer to select probabilities marked by high pheromone concentrations.

Ants travel straight from the nest to the food source (Figure 1(a)). Assume that there is an obstacle at the next level (Figure 1(b)). In this case, to escape an obstacle, each ant chooses to be left or right at random Figure 1(c)). Let us presume that the ants travel at the same pace as the pheromone in the trail. However the ants who by chance, choose to turn left will soon hit the food source, while the ants that turn right will take a longer path. The strength of the pheromone on the shorter path is more than the other path. We will therefore be gradually directed by ants to travel along the shorter path (Figure 1(d)). The strength of the pheromone deposited is one of the most important factors for ants to find the shortest path.

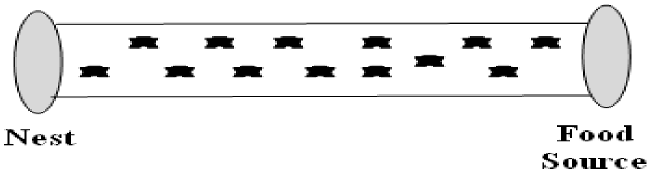

Fig 6.1: - A path between their nest and foods source

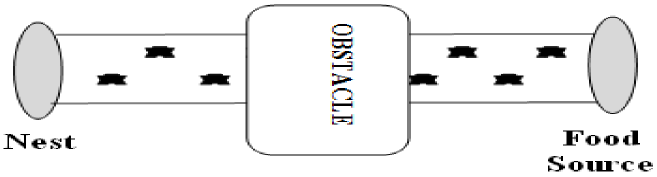

Fig 6.2: - Encountering obstacles of ants

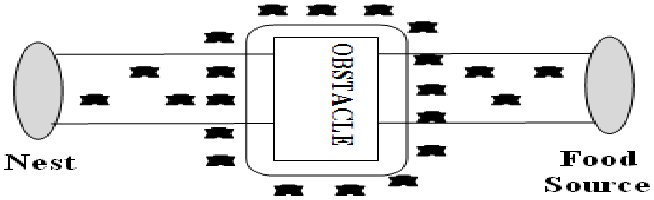

Fig 6.3: - Selecting of Ants

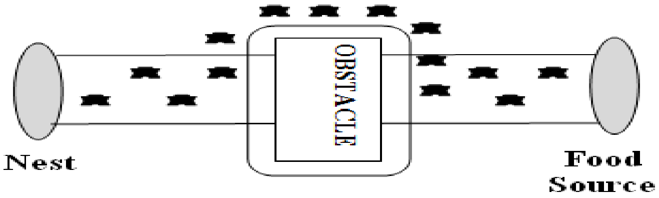

Fig 6.4: - Finding shortest path of ants

The proposed ACO algorithm is based on an eco-intelligent machine, autonomous and cooperative ants. In this proposed algorithm, ants will procreate and commit suicide, depending on the current situation. In order to increase the efficiency of the mechanism, a level load balancing is proposed. Ants are generated on demand over their lifetime adaptively to achieve grid load balancing. Ants can bear offspring when they detect that the system is radically unbalanced and commit suicide when they detect environmental equilibrium. The ants will take care of any node they visit during their steps and record the node requirements for future decision-making. Theoretical and simulation findings suggest that this new algorithm beats its predecessor. However, pheromone values have not been modified in this proposed algorithm, which allows jobs to be allocated to the same resource.

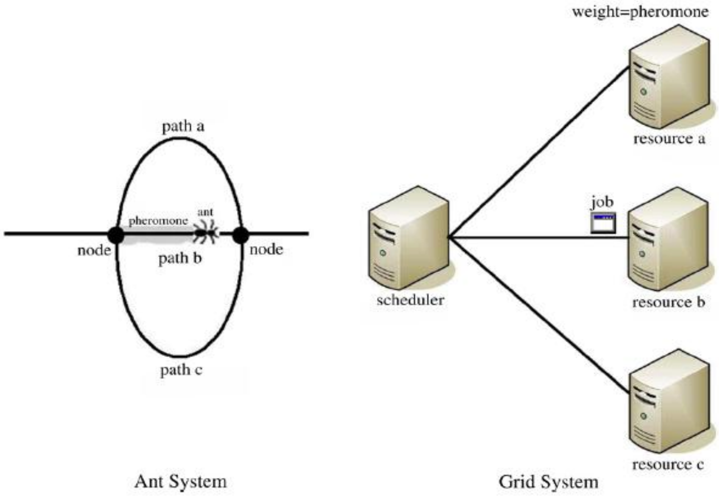

Fig 6.5: - Mapping between the Ant System and the Grid System

This has inspired the discovery of the ACO algorithm. This algorithm uses a colony of artificial ants that function as cooperative agents in a mathematical space where pathways (solutions) can be searched and reinforced to find the optimal ones. This population-based approach has been successfully applied to several NP-hard optimization problems.

Algorithm: - Ant Colony Optimization (ACO)

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

The underlying concept of the ACO algorithm emphasizes on agents where agents describe ants that travel randomly around in their world, which is a square grid with periodic boundary conditions. When ants roam around in their surroundings, they pick up a data object that is either isolated or surrounded by dissimilar ones. The item selected will be transported and dropped by ants to form a group of related neighbourhood objects based on the similarity and density of data items [15]. The probability of choosing an element increases with a low density and decreases with the similarity of the element. The concept behind this form of pheromone aggregation is the attraction among data items and artificial ants. Small clusters of data items expand by attracting ants to store more items.

7. PERFORMANCE EVALUATION

High dimensional data are data characterized by a few hundred or several thousands of dimensions. And any dataset that is capable under a relational model is chosen as a high dimensional dataset (shown in table 5.1). It is worth remembering that the following six separate datasets have been used: 20NG, Sports, Wellness, Culture, and Local News.

Table 5.1: - High Dimensional Dataset Details

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

The data sets used have different characteristics in terms of the size of the vocabulary and the distribution of the group. The attributes or measurements of each document 1000 are interrelated and are classified using the Bigdata Analytics Techniques.

7.1. Performance Evaluation Parameters

The following output metrics are widely used in the assessment of privacy security techniques. The current method is contrasted with the proposed scheme using these evaluation parameters [13-14]. The efficiency of the TC process can be calculated using one or more of the following methods:

7.1.1. Recall and Precision

They are two well-known indicators of effectiveness in text mining. Although Recall is a measure of the system's correctly predicted documents among the positive documents, Precision is a measure of the system's correctly predicted documents among all the predicted documents. The method is tested in terms of accuracy, recall and F-measure.

Recall is described as the amount of relevant documents obtained by a search divided by the number of current 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 the search.

$$\textbf{Precision} = \frac{\text{Number of correct results}}{\text{Number of all returned results}}$$

$$\textbf{Recall} = \frac{\text{Number of correct results}}{\text{Total number of actual results}}$$

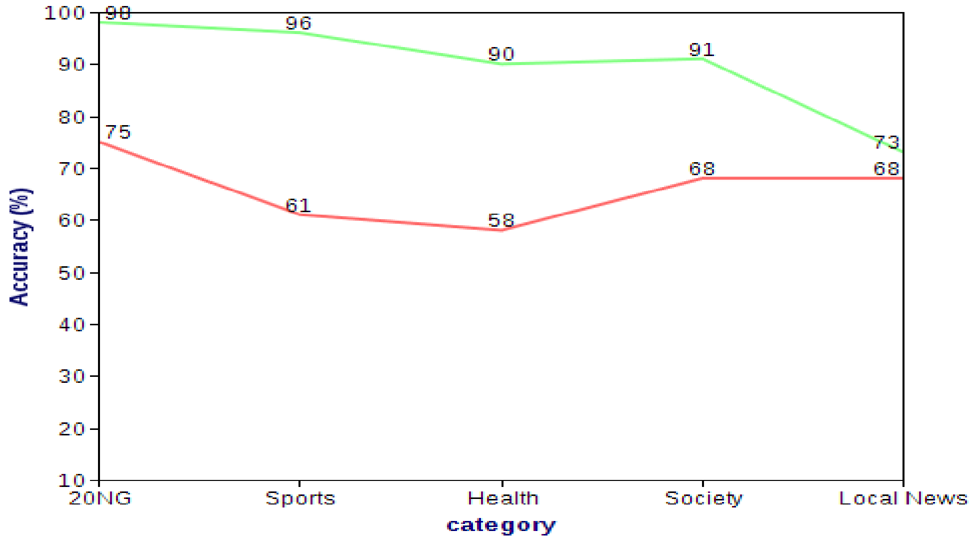

Fig 7.1: - Evaluation of Precision

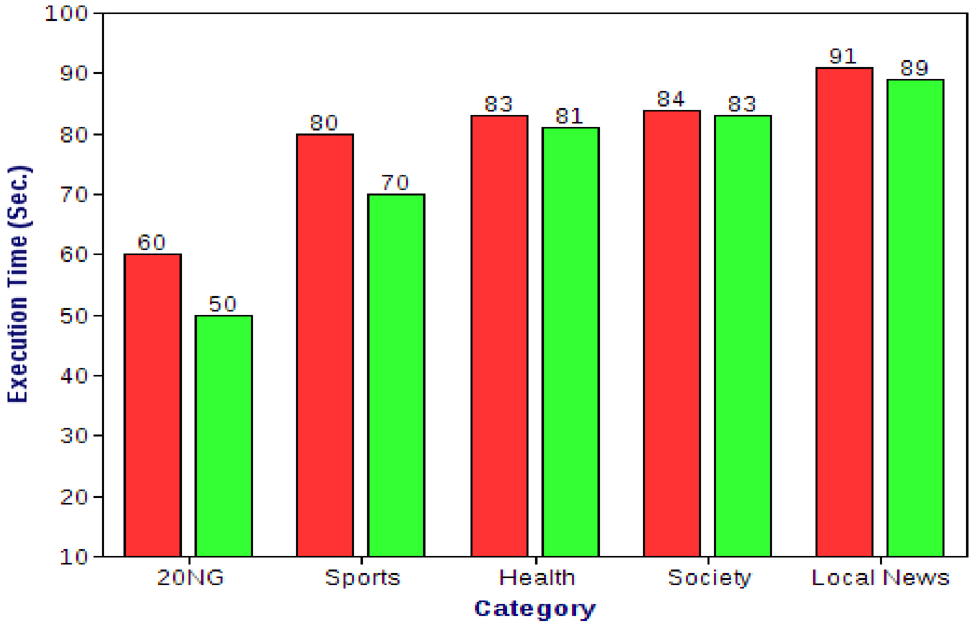

Fig 7.2: - Evaluation of Recall

7.1.2. F-Measure

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

$$\textbf{F-measure} = \frac{\text{Precision x recall}}{\text{Precision + recall}}$$

Several experiments were performed with different query documents and the accuracy, recall and F-Measurement of output was measured. This higher increase in the precision value can be undermined by a very small percentage reduction in the recall value. In addition, the F-Measure, which incorporates accuracy and recall, is much better than the current system.

7.1.3. Distortion

It is calculated with the assistance of the analysis between the original dataset and the updated dataset. Every tuple Xi in the unique dataset [10], of the m columns, and each one of the columns has n characteristics, which is changed to Yi in the updated dataset, is used to register contortion in that tuple by defining the difference between them by the Euclidean separation by the equation.

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

They hope that this partnership would provide them with control over whatever remains of their competitors who have not taken part in the concerted effort. All things considered, there might be little willingness on the part of working together organisations to unveil such touchy principles stored in their respective information sets.

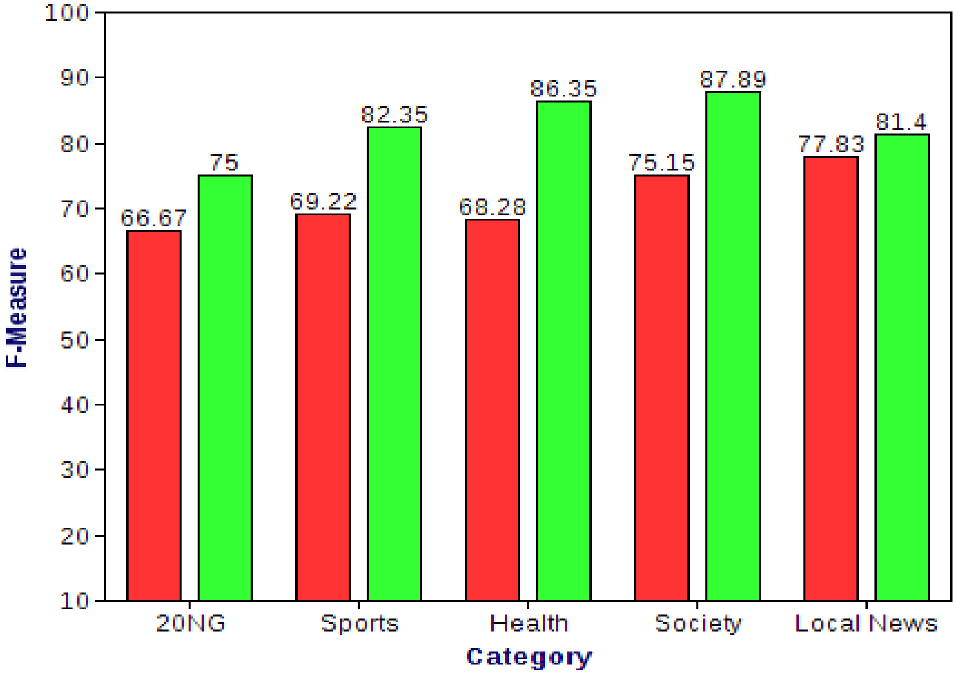

Fig 7.3: - Evaluation of F-Measure

8. CONCLUSION

Big data has the potential to radically change the manner in which institutions and organisations use their data. Transforming vast quantities of data into information would improve the efficiency of organisations. Scientific and business organisations will benefit from the use of big data. However there are many difficulties in dealing with big data, such as storing, transferring, handling and manipulating big data. Many techniques are needed to explore hidden trends within big data that have limitations in hardware and software implementation. The implementation of large data solutions needed an infrastructure that supports the scalability, delivery and management of data. Thus this study proposes grid technology to address hardware limitations in terms of storage space, computing power and memory speed. An ant-based clustering algorithm is proposed for algorithm scalability. A method for clustering large data using grid computing and ant colony algorithm has been suggested. The grid principle is designed to allow the storing of data in distributed databases across a large geographical region, while the ant-based algorithm is for the clustering of big data.

ACO algorithm has many advantages to be used in large data mining because it has the potential to scale with the size of the data set, no prior knowledge of the number of expected clusters is required and easy to incorporate with the cluster ensemble model. Big data analysis opens the door to many fields of study and data protection is one of the most important areas. In the future, we will review whether there are any other conditions that we do not take into account in our pheromone indicator concepts or pheromone update functions.

REFERENCES

[1] Garlasu, D., Sandulescu, V., Halcu, I., Neculoiu, G., Grigoriu, O., Marinescu, M., & Marinescu, V. (2013, January). A big data implementation based on Grid computing. In *2013 11th RoEduNet International Conference* (pp. 1-4). IEEE.

[2] Das, T. K., & Kumar, P. M. (2013). Big data analytics: A framework for unstructured data analysis. *International Journal of Engineering Science & Technology*, *5*(1), 153.

[3] Tu, C., He, X., Shuai, Z., & Jiang, F. (2017). Big data issues in smart grid–A review. *Renewable and Sustainable Energy Reviews*, *79*, 1099-1107.

[4] Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, *5*(12), 2032-2033.

[5] Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.

[6] Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293-314.

[7] Daki, H., El Hannani, A., Aqqal, A., Haidine, A., & Dahbi, A. (2017). Big Data management in smart grid: concepts, requirements and implementation. *Journal of Big Data*, *4*(1), 1-19.

[8] Stimmel, C. L. (2014). *Big data analytics strategies for the smart grid*. CRC Press.

[9] Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, *30*(4), 431-448.

[10] Lee, Y. H., Leu, S., & Chang, R. S. (2011). Improving job scheduling algorithms in a grid environment. *Future generation computer systems*, *27*(8), 991-998.

[11] Wang, J., Huang, A. Q., Sung, W., Liu, Y., & Baliga, B. J. (2009). Smart grid technologies. *IEEE Industrial Electronics Magazine, 3*(2), 16-23.

[12] Gungor, V. C., Sahin, D., Kocak, T., Ergut, S., Buccella, C., Cecati, C., & Hancke, G. P. (2011). Smart grid technologies: Communication technologies and standards. *IEEE transactions on Industrial informatics*, *7*(4), 529-539.

[13] Veiga, J., Expósito, R. R., Pardo, X. C., Taboada, G. L., & Tourifio, J. (2016, December). Performance evaluation of big data frameworks for large-scale data analytics. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 424-431). IEEE.

[14] Colak, I., Sagiroglu, S., Fulli, G., Yesilbudak, M., & Covrig, C. F. (2016). A survey on the critical issues in smart grid technologies. *Renewable and Sustainable Energy Reviews*, *54*, 396-405.

[15] Hong, T., Chen, C., Huang, J., Lu, N., Xie, L., & Zareipour, H. (2016). Guest editorial big data analytics for grid modernization. *IEEE Transactions on Smart Grid*, *7*(5), 2395-2396.

[16] Tu, C., He, X., Shuai, Z., & Jiang, F. (2017). Big data issues in smart grid–A review. *Renewable and Sustainable Energy Reviews*, *79*, 1099-1107.

[17] Lai, C. S., & Lai, L. L. (2015, October). Application of big data in smart grid. In *2015 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 665-670). IEEE.

[18] Munshi, A. A., & Yasser, A. R. M. (2017). Big data framework for analytics in smart grids. *Electric Power Systems Research*, *151*, 369-380.

[19] Russom, P. (2011). Big data analytics. *TDWI best practices report*, *fourth quarter*, *19*(4), 1-34.

[20] Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of parallel and distributed computing*, *74*(7), 2561-2573.