Distributed Modeling-Grid vs. Cloud Computing

[1]Darshna tanwar,[2]Jyoti Vashishtha
[1]Research Scholar,[2]Associate Professor
[1,2]Department of Computer Science and Engineering,
[1,2]Guru Jambheshwar University of Science and Technology, Hisar, India

Abstract : Getting global knowledge from decentralized local data sources is entitled as distributed data mining and data mining for a distributed environment such as Cloud and Grid. Cloud Computing is suitable for individuals and businesses where leased services fulfill sudden requirement. Grid Computing is suitable for researchers and academia's where resource sharing is for dynamic and collaborative problem. Grid Computing and Cloud computing both have their own pros and cons but whenever Grid technology is combined with data mining techniques, it provides best results over heterogeneous data. A lot of work is going on in this direction and a number of researches of data mining with these two technologies have been developed by the researchers such. This paper provides a comparative study of Grid and Cloud computing in context of distributed data mining available in research literature.

IndexTerms **- Grid Computing; Cloud Computing; Distributed Data Mining;**

I. INTRODUCTION

Millions of users are performing millions of operations based on thousands of servers. This scenario generates or demands extreme processing power and data storage, to cope with this extreme demand, distributed computing comes into existence [1]. "Distributed computing system is a network of autonomous computers that communicate with each other in order to achieve a goal as well as they are independent because they do not share memory or processor". Distributed Computing is a field related to distributed systems. Distributed system is a model in which component located on networked system where components communicates and coordinates their actions by passing messages [2]. Distributed computing uses various software's that assist to accomplish a specific goal. Most of social networking sites follows the concept of distributed computing[3]. Objectives of distributing computing achieved by a group of computers which collaborating in a network. These collaborating computers are of diverse architectures and representation format still they share data as well as system resources to accomplish the goal.

Data mining is extracting knowledge from provided datasets. Analyzing data in a distributed fashion is called as Distributed data mining[4].Distributed data mining can be applied in following three steps ,first of all, Analyzing of data at local sites , secondly, discovered knowledge from local sites is transmitted to a merger site where data is again analyzed after merging of results of local sites. Thirdly, Final results are transmitted back to local sites and these local sites get updated with the global knowledge.

Cloud computing is an environment that provides ever-present, convenient, on demand accessing to numerous resources with minimal management cost and service provider interaction [5].Cloud computing allows for reliable deliverance of software's, platforms and infrastructures in the form of service over the internet, even for remote data centers [6].Discovering of knowledge from unstructured or semi structured web data is known as data mining in cloud computing. Cloud Data Mining is helpful to a cloud provider in managing efficient, reliable and secure centralize software's and data storages[7].

Grid is a scattered communications that facilitates a virtual organization across purely detached sites for the purpose of providing computing resources for applications that requires coordination and sharing [8]. The Grid arrangements follow distributed and parallel computing paradigms together which allows heterogeneity, portability, resource cooperation and dynamic allocation of resources. Both Computational and data related tasks can be solved in Grid environments [8]. Grid based distributed data mining addresses to a structure where data mining could be applied to geographically distributed environments [9],[10],[11].The paper is divided into five sections. Section II presents brief literature review about architecture and types of cloud computing, section III discusses the architecture of grid computing and its types and section IV describes comparisons between grid computing and cloud computing on the basis parameters. Final section concludes the research and sets new research directions.

II. CLOUD COMPUTING

Cloud computing is a mixture of two computing paradigms i.e. parallel and distributed. It is not a new name, concepts of Cloud computing exists in early 1996 with the name Compaq internet document and this becomes popular firstly when amazon company introduced its elastic compute cloud [12]. Term Cloud Computing misunderstand with the terms such as client-server model, Grid computing, Fog computing ,Dew Computing , Utility computing , Peer-to-Peer Computing etc. Cloud Computing is a technology that provides or delivers computing services such as servers, storage, databases, networking devices, software's , analytics or data mining services and many more over the internet to consumers or businesses .Businesses and consumers can use these services without worrying about installation or implementation of hardware or infrastructure. This technology all depends on analogy that "if you wish milk would you get a cow." [13],[14]. Main advantages of cloud computing are Self service provisioning, physically property and pay-per-use[15].

2.1 Architecture of Cloud Computing Technique

Cloud computing may be a computing paradigm, wherever an outsized pool of systems are connected in a private network or public network, for the purpose of supplying demand based ascendible infrastructure for knowledge, application and file storage. Thanks to this computing technique, the cost of storage, product delivery, computation and application hosting is decreased considerably. Cloud architecture classified into six layers [16]. Figure 1 depicts the design of Cloud Computing.

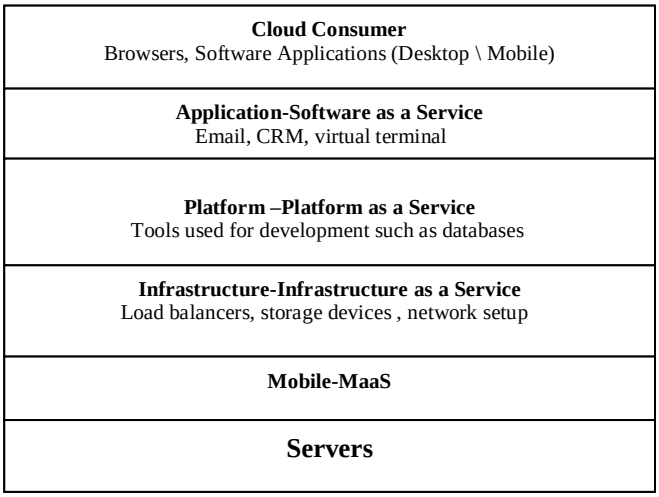


Fig. 1. Design of Cloud Computing

A Cloud consumer consists of component, computer code and business that depend on cloud computing services. Cloud client can access and use the services provided by cloud.

A Cloud Application delivers "Software as a Service" over the net. Through this cloud purchasers area unit capable to access and run applications provided by the cloud supplier while not taking care of hardware management for application. Purchasers need not to install and run the application on client's machine.

Platform services "Platform as a Service" makes available a computing platform by means of the cloud infrastructure. Using these services end user is capable enough to deploy onto the cloud infrastructure. End users or businesses can run their own application on platform provided by cloud provider or even able to create their own application.

Infrastructure services "Infrastructure as a Service" supplies the desired infrastructure as a service. The purchasers needn't to get needed network servers or knowledge centers. Shoppers ought to pay just for time period they use the services. They are able to deploy and run discretionary computer code together with package and application conjointly.

"Mobile Backend as a Service", web app as well as mobile app developers provides a way to link their applications to cloud storage.

Servers are system that means combination of embedded systems needed for the delivery of the above mentioned services.

2.2 *Deployment models of Cloud Computing Technique*

Four deployment models of cloud computing that consumer can subscribe just in few steps according to requirement [17]:

2.2.1 Private Cloud- Private cloud belongs to an individual organization therefore all computing resources also operated ,owned and governed by that individual organization. This technique is support for intra-business interactions. Wherever the computing resources can be governed, owned and operated by the identical organization.

2.2.2 Community Cloud-The cloud infrastructure is shared by many organizations and supports a particular community that has shared issues.

2.2.3 Public Cloud- This cloud infrastructure is created on the market to the final public that provided by cloud supplier. Cloud Supplier is a corporation who sells its services over web.

2.2.4 Hybrid Cloud- This sort of cloud infrastructure is may be a composition of two or additional clouds such as private or public, public or community ,community or private .This composition remain unique by bounding together two technologies.

2.3 Distributed data mining using Cloud computing technique

Cloud computing will be thought of a decent platform for data mining as a result of sometimes the input file is incredibly massive and distributed in nature. The sweetness of cloud computing is that it will offer centralized storage and process simply. Additionally we've got to buy simply what we tend to use therefore it conjointly cuts down the value [18]–[21]. Given below are some design which mixes cloud and data mining like

Hadoop is an open source framework based on distributed computing, that can process a large number of distributed data sets in the form of clusters (configured of single server to thousands of servers) as well. In recent years, most of the Internet organizations and research institutions adopt Hadoop as cloud computing platform. The Hadoop distributed file system and Map/Reduce are two parts from which hadoop cluster is made of.[22].

BC-PDM (Big Cloud Parallel Data Mining) is a collection of various data mining and analysis techniques. Features of BC-PDM are good performance, most reliable, maximum scalability with least cost [23]. For parallel data mining algorithms ETL(Extract Transform and Load) concept is used in BC-PDM and parallel mining algorithm supports enterprise BI application and accurate marketing ,provides the SaaS service mode based on Web, and reduce the IT system investment of enterprise.

PDMiner is a parallel distributed data mining platform used on Hadoop, which developed by the Institute of Computing Technology, PDMiner provide the vast majority of a series of parallel mining algorithms and ETL operations components, development of ETL algorithm to achieve a linear speedup, meanwhile has good fault tolerance[23]. PDMiner has open architecture that allows the user to system through a simple configuration.

Robert Grossman et al. [24] propose the Sector/Sphere framework which is used to mine large-scale datasets using computer clusters connected with wide-area high-performance networks. While high-level data parallel frameworks like Sector/Sphere simplify the design and implementation of large-scale data processing systems, they do not naturally support many important data mining algorithms and can lead to inefficient learning systems.

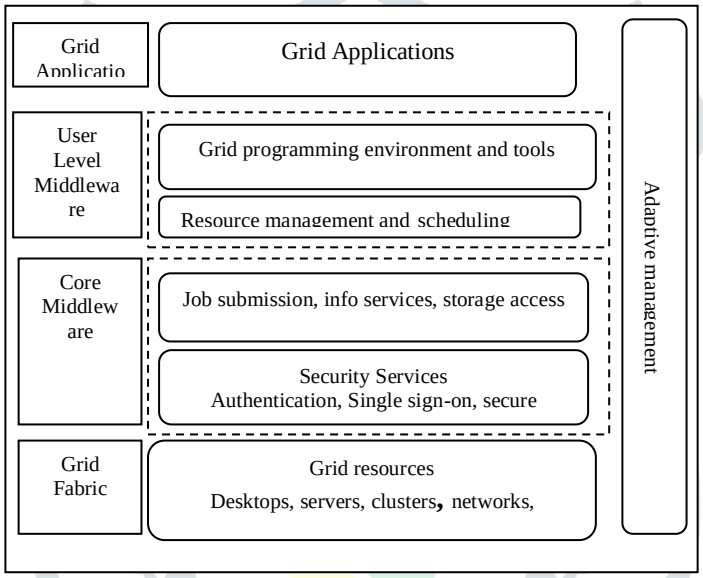
Leila Ismail et al.[25] designed Cloud computing framework, which is responsible to distribute and schedule a Cluster-Based data mining application and its data set. The main goal of the proposed framework for scheduling of Big Data Mining (FSBD) is to decrease the overall execution time of the application with minimum loss in mining quality. They consider the Cluster-based data mining technique as a pilot application for our framework. The results show an important speedup with a minimum loss in quality of mining.

III. GRID COMPUTING

The Grid concept is related to sharing of resources for the purpose of solving complex and collaborative problems and resource brokering strategies emerging in industry, science, and engineering. It is not only related to file exchange rather direct access to computers, software, data and other resources[26].For extensive resource sharing, pioneering applications, and high-performance orientation a computational model has been projected known as Grid computing. Grid can be viewed as a cohesive resource to the grid user, these resources are reasonably tied an infinite number of computing devices which ranged from high speed devices to specialized devices.

3.1Grid Architecture

The layered Grid architecture [27] organizes various grid capabilities and components such that high level services are built using lower-level services.



Grid fabric software layer provides local Grid resources with resource management and execution environment i.e. computers with different operating systems, storage devices and embedded devices such as heat sensor.

Core Grid middleware layer provides Grid infrastructure i.e. services are related to storage access, trading, accounting, payment, security and information services. Trading of resources is beneficial for decentralization of Grids because it is based on the computational economy approach.

User-level middleware layer provides programming environments for different applications .Resource broker is an application which is used to for selection of appropriate resources in the context of applications.

Grid applications Layer makes available Grid services to the end-users. Consequently, to allow users for universally access of any resource from anywhere by using any platform at any time, a number of web portals are being built.

3.2 Types of Grid

3.2.1 On the basis of Nature of Grid [26]

Computational Grid is collection of globally scattered computers whose computational power is used. Example of computation Grids are TeraGrid, ChinaGrid.

Data Grid focuses on data stored in distributed data repositories, data's global-level supervision during accessing, integration and processing of it at distribiuted sites. Example of Data Grids are LHCGrid and GriPhyN.

Application Service Provisioning (ASP) Grid provides access to applications situated at remote location whether it is hosted on Data Grids or Computational Grids. Example of ASP Grids are Net Solve, GridSolve.

Interaction Grid is provides interface and set up a joint revelation between different users of grid. Example of Interaction Grid is Access Grid.

Knowledge Grid focuses on knowledge acquirement, processing, supervision as well as it provides various analytical services motivated from integrated data mining services .Example of Knowledge grid are Italian KnowledgeGrid and EUDataMiningGrid.

Utility Grid is able to provide Grid services as utility to the users on subscription basis. Allocation of resources to the applications from multiple users done by Utility Grid. Example of Utility Grids are Gridbus and Utility Data Center.

3.2.2 On the basis of levels of complexity for the enterprise [28]

Infra-Grid allows optimizing the resource sharing within division of the organization's departments.

Intra-Grid focused on collecting diverse resources from numerous departments and divisions of an enterprise.

Extra-Grid related to resource sharing to / from foreign partners of an organization due to certain relationships.

Inter-Grid enabled sharing and storage of resources and data using the Web and enabling the collaborations between various companies and organizations. Inter-Grid involves all facilities of last three types of grid.

3.3 Distributed data mining using Grid Computing

Knowledge Grid has been designed by Taila Domenico, Cannatro Mario [29]. It is a software framework, defined on top of Globus Toolkit and services, for implementing knowledge discovery tasks for geographically distributed applications with high performance.

Brezany P. et al.[30] have developed *GridMiner* framework which is defined to deal with all the tasks related to knowledge discovery process on grids and integrate this knowledge discovery process in an advanced service-oriented grid applications. Grid Miner framework is divided in two parts: first part consists of tools and technologies whereas second one consists of use cases which demonstrate the combination of technologies and tools as well as make use of it in realistic situations.

*DataMiningGri*d has been developed by Stankovski et al.[31] It facilitates a generic system for the development and deployment of grid enabled data mining applications. It is based on a service-oriented architecture (SOA) through which modern distributed data mining scenarios could be implemented.

Algorithm Development and Mining System (ADaM) toolkit has been designed by Ramachandran et al. [27]for mining of large scientific data sets such as for geophysical phenomena detection and feature extraction. The original design of ADaM was a comprehensive system which contained different key software components for data mining over distributed computing environments including a mining daemon, a mining database, a mining scheduler, a set of mining operations and last but not least a mining engine to handle mining requests, to fetch and stage appropriate data, to schedule different mining jobs, to parse mining plans or workflows respectively.

IV. COMPARATIVE STUDY

Grid infrastructure and Cloud infrastructure both looks same but there is a huge difference between their technology and this totally depends on type of work that user wants to apply.Table1 shows the difference between these two major technologies[32]–[39]. Grid Computing shares computing power and storage capacity of distributed resources. Researchers uses this infrastructure for short term jobs with large datasets because it is reliable, secure, transparent, resilience, non profitable but little complex.

Cloud Computing provides its computing power and resources on lease which stores at centralized locations. Small or medium scale businesses use this technology for long term jobs with small datasets because it is easy to use, little reliable, beneficial for providers, quick response to sudden needs.

Table 1 Comparison between Grid Computing and Cloud Computing for the benefit of Distributed Data Mining

Parameters	Grid Computing	Cloud Computing
When?	This Concept was proposed in 1995 by starting Open Science Grid Started in 1995.	Oracle suggests the concept of Cloud in the late 1990's but actually this concept comes into existence after the announcement of first cloud by amazon in 2007.
What?	Shared Computing power and storage capacity.	Leased Computing power and storage Capacity.
Why?	For Short term jobs which uses or generates huge amount of data.	For Long term services and extensively running jobs where needs quick access of extra resources.
Who Provides? *Or* *Providers*	By research institutes and universities or academia's.	By Large Individual Companies.
Who Uses? *Or* *Users*	Virtual organizations (Research Collaborations) that bring together researchers or analyzers around the world.	Small, Medium businesses or researcher with generic IT needs.
Where Resources? *Or*	In Computing centers distributed across different sites ,countries or continents	In Centralized data centers on some specific locations

Resources Resides		
How Do they Work? *Or* *Mode of Working*	Grids are an Open-Source Technology where Grid Users and Providers both can understand and manage grid.	Clouds are a proprietary technology. Only resource Provider knows exactly how their cloud manages data, job, queue, security requirements etc.
Distributed Data Mining + Technology	By Combining the concept of DDM and Grid, some architectures are already build such as Knowledge Grid, DataMiningGrid, Grid Miner etc.	Architectures based on DDM And Cloud computing are very few such as Distributed Graph LAB. But on Data Mining using Cloud computing are DMCC, GraphLab, Hadoop/MaPReduce etc.
Collaboration	Grids Offers a federated platform for distributed and collective work.	No such platform provided by Cloud
Flexibility	Grid resources are also accessible to all users whenever they demand.	Users can quickly outsource peaks of activity without long term commitment.
Reliability	Grids rely on distributed services maintained by distributed staff, often resulting in inconsistency in reliability across individual sites, although the service itself is always available.	Provider has financial incentive to guarantee service availability
Ownership	Resource providers maintain ownership of the resources they contribute to the grid.	Cloud provider maintain ownership of the resources.
Transparency	The Technologies used are open source, encouraging trust and transparency.	The technologies used in cloud computing is only known to cloud providers.
Ease of Use/ Complexity	Grids are complicated to build and use, and currently users require some level of expertise.	Relatively quick and easy for non expert users to get started but setting up sophisticated virtual machines to support complex applications is more difficult.
Resilience/Rigidity	Grids are located at multiple sites, reducing the risk in case of a failure at one site that removes significant resources from the infrastructure.	The clouds are generally located at a single site, which increases risk of complete cloud failure.
Commercial	Grids are generally only available for not-for-profit work and for proof of concept in the commercial sphere.	Clouds are use the concept of pay-per-use.
Generality	Grid offer specific high-level services.	Clouds do not offer many of the specific high-level services currently provided by grid technology.
Security	Grid Users can access sensitive data if only they have authorization and authentication.	Users with sensitive data may be reluctant to entrust it to external providers or to providers outside their borders.
Opacity	The technology used to guarantee reliability and safety of Grid operations are well known.	The technology used to guarantee reliability and safety of cloud operations are not made public.
Provider Lock-in	There is no risk of being locked in to services provided by a small group of suppliers.	There is a risk of being locked in to services provided by a very small group of suppliers.

V. Conclusion and Future Work

Now days, data mining is a necessity and essential field for business intelligence to complex scientific data. Centralized data mining techniques are not suitable in case of complex scenarios in which virtual organization are concerned. In this study we have observed various parameters for the comparison of two well known computing techniques grid computing and cloud computing. Distributed data mining gradually relying on distributed computing environment i.e. Grid infrastructure but cloud is less suitable for distributed computing environment so for distributed data mining. There can be many possible extensions of this work towards comprehensive, productive and high- performance analysis of scientific data sets using Grid Computing or Cloud Computing.

REFERENCES

[1] "Chapter 4: Distributed and Parallel Computing." [Online]. Available: http://wla.berkeley.edu/~cs61a/fa11/lectures/communication.html.

[2] "Distributed computing - Wikipedia." [Online]. Available: https://en.m.wikipedia.org/wiki/Distributed_computing.

[3] E. R. Kaur, "A Review of Computing Technologies: Distributed, Utility, Cluster, Grid and Cloud Computing," *International Journal of Advanced Research in Computer Science and Software Engineering*, volume 5, no. 2, pp. 144–148, Feb. 2015.

[4] G. Tsoumakas and I. Vlahavas, "A survey on distributed data mining and its trends," *International Journal of Research in Engineering and Technology*, volume 2, pp. 107–120, 2014.

[5] T. Hu, H. Chen, L. Huang, and X. Zhu, "A survey of mass data mining based on cloud-computing," in *Anti-counterfeiting, Security, and Identification*, pp. 1–4, 2012.

[6] M. A. S. Patil and M. A. S. Patil, "A review on Data Mining based Cloud Computing" *International Journal of Research In Science & Engineering*, Volume 1, 2015.

[7] T. Zhu, J. Yu, and B. Du, "RTIC-C: A Cloud Computing Platform for History Data Mining of Traffic Information," in *Proceeding of International Conference on Connected Vehicles and Expo (ICCVE)*, pp. 282–283, 2012.

[8] I. Foster, C. Kesselman, and S. Tuecke, "The anatomy of the Grid," in *International Journal of Supercomputer Applications*, pp. 171–197, 2001.

[9] E. Cesario, M. Lackovic, D. Talia, and P. Trunfio, "Programming knowledge discovery workflows in service-oriented distributed systems," in *Concurrency and Computation: Practice and Experience*, vol. 25, no. 10, pp. 1482–1504, 2012.

[10] M. Cannataro, D. Talia, and P. Trunfio, "Design of distributed data mining applications on the knowledge grid," in *Proceedings National Science Foundation Workshop on Next Generation Data Mining*, pp. 191–195,2002.

[11] G. Bueti, A. Congiusta, and D. Talia, "Developing Distributed Data Mining Applications in the Knowledge Grid Framework," in *Proceedings of International Conference on High Performance Computing for Computational Science - VECPAR 2004*, vol. 3402, pp. 156–169, 2005.

[12] "Cloud computing," *Wikipedia*. 05-Dec-2016.

[13] "Concept:Cloud computing." [Online]. Available: http://www.wikinvest.com/concept/Cloud_computing.

[14] "What is cloud computing? A beginner's guide | Microsoft Azure." [Online]. Available: https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/.

[15] "Cloud management and monitoring news, help and research - SearchCloudComputing." [Online]. Available: http://searchcloudcomputing.techtarget.com/resources/Cloud-management-and-monitoring. [Accessed: 09-Dec-2016].

[16] Y. Jadeja and K. Modi, "Cloud computing-concepts, architecture and challenges," in *Computing, Electronics and Electrical Technologies (ICCEET), 2012 International Conference on*, 2012, pp. 877–880.

[17] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," in *Future Generation Computer System*, vol. 25, no. 6, pp. 599–616, 2009.

[18] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein, "Distributed GraphLab: a framework for machine learning and data mining in the cloud," *The Proceedings of the VLDB Endowment*, vol. 5, no. 8, pp. 716–727, 2012.

[19] R. Vrbić, "Data Mining and Cloud Computing," *Journal of Information Technology and Applications (Banja Luka) – APEIRON*, vol. 4, no. 2, 2012.

[20] M. Butwall and S. Kumar, "A Review on Data Mining Algorithms in Cloud Environment," *International Journal of Computer Applications & Information Technology*, vol. 6, no. 2, pp. 129–135, 2014.

[21] M. D. Dikaiakos, D. Katsaros, P. Mehra, G. Pallis, and A. Vakali, "Cloud computing: Distributed internet computing for IT and scientific research," in *IEEE Internet Computing*, vol. 13, no. 5, pp. 10–13, 2009.

[22] Q. Ren, "The Development of Data Mining System Based on Cloud Computing," *in Advanced Materials Research*, vol. 926–930, pp. 2280–2283, 2014.

[23] B. Ambulkar and V. Borkar, "Data mining in cloud computing," in *MPGI National Multi Conference*, pp. 7–8, 2012.

[24] R. Grossman and Y. Gu, "Data mining using high performance data clouds: experimental studies using sector and sphere," in *Proceedings of international conference on Knowledge discovery and data mining*, pp. 920–927, 2008.

[25] L. Ismail, M. M. Masud, and L. Khan, "FSBD: A Framework for Scheduling of Big Data Mining in Cloud Computing," -in Proceeding of 2014 IEEE confress on Big Data, pp.514-521, 2014.

[26] C. S. Yeo *et al.*, "Utility computing and global grids," in *Handbook CN Utility GRID*, pp. 1–26, 2006.

[27] M. Zhizhin, A. Poyda, D. Mishin, D. Medvedev, E. Kihn, and V. Lyutsarev, "Grid-Based Data Mining With the Environmental Scenario Search Engine (ESSE)," in *Data Mining Techniques in Grid Computing Environments*, pp. 221–245, 2009.

[28] P. Plaszczak and R. Wellner, *Grid computing*. Book ,pp.1-288, 2006.

[29] M. Cannataro, D. Talia, and P. Trunfio, "Knowledge grid: high performance knowledge discovery services on the grid," in *Grid Computing—GRID 2001*, vol. 2242, Springer, pp. 38–50, 2001.

[30] Brezany, Peter & Hofer, Jürgen & Tjoa, A Min & Wöhrer, Alexander, "GridMiner: An Infrastructure for Data Mining on Computational Grids, "2009 .

[31] Vlado Stankovski et al."Digging Deep into the Data Mine with DataMiningGrid" in *IEEE Internet Computing* volume: 12 , pp.69-76 , 2008.

[32] R. Kumar and S. Charu, "Comparison between Cloud Computing, Grid Computing, Cluster Computing and Virtualization," *International Journal Modern Computer Science Application*, vol. 3, no. 1, pp. 42–47, 2015.

[33] S. Jafari and M. Raeisi, "Compare Cloud and Grid Computing," in *International Journal of Scientific & Engineering Research*, vol. 3, no. 3, pp. 142–146, 2014.

[34] S. M. Hashemi and A. K. Bardsiri, "Cloud computing Vs. grid computing," in *ARPN Journal of System and Software*, volume 2, no. 5, pp. 188–194, 2012.

[35] H. AlHakami, "Comparison Between Cloud and Grid Computing: Review Paper," in *international journal on cloud computing services and architecture*, volume 2, no. 4, pp. 1–21, Aug. 2012.

[36] "Difference between Cloud Computing and Grid Computing." [Online]. Available: http://www.rapidprogramming.com/questions-answers/difference-between-cloud-computing-and-grid-computing-1542. [Accessed: 09-Dec-2016].

[37] "Difference between Cloud computing and Grid computing," *The Windows Club*, 12-Nov-2011. .

[38] "Cloud computing versus grid computing," 03-Mar-2009. [Online]. Available: http://www.ibm.com/developerworks/library/wa-cloudgrid/index.html. [Accessed: 09-Dec-2016].

[39] "Clouds and grids compared - Cloud lounge." [Online]. Available: http://www.cloud-lounge.org/clouds-and-grids-compared.html. [Accessed: 09-Dec-2016].