Grid Based Distributed Data Mining–A Review

Darshna Tanwar
Department of Computer Science and Engineering
Guru Jambheshwar University Science & Technology, Hisar, India

Jyoti Vashishtha
Department of Computer Science and Engineering
Guru Jambheshwar University Science & Technology, Hisar, India

Abstract- Grid computing is modern and well defined structure for implementing distributed computing. Grid computing enables coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations. Distributed data mining intends to get the global knowledge from the local data at distributed sites. Distributed data mining addresses the impact of distribution of users, software and computational resources. Grid technologies, combined with distributed data mining techniques, obtain best results over heterogeneous data and jointly these approaches called as Grid based distributed data mining. Grid based distributed data miners are the software architectures for geographically distributed high-performance knowledge discovery applications such as Knowledge Grid, Data Mining Grid etc. These miners are designed on crest of different grid environments such as Globus, GridBus. This paper reviews different grid computing architectures, types of grids, distributed data mining and grid based distributed data miners.

Keywords – Grid Computing; Distributed Data Mining; Grid Based Data Miners

I. INTRODUCTION

In distributed computing environment, concept of distributed memory has been applied i.e. each processor has its own private memory[1], [2] and information is exchanged among the processors. Distributed computing environments can be implemented using one of the various architectures such as Web Technology (Client server), 3-tier Architecture, *N*-tier Architecture, Peer-to-Peer Computing, Cloud Computing, Cluster Computing (Highly Coupled), Virtualization (Space based), Grid Computing[3],.

In Web Technology (Client server) Smart client code contacts the server for data then formats and displays it to the user. Input at the client is committed back to the server when it represents a permanent change. 3-tier Architecture move the client intelligence to a middle tier so that stateless clients can be used that simplifies application deployment. Most web applications are 3-Tier. N-tier Architecture refers typically to web applications which further forward their requests to other enterprise service.Peer-to-Peer Computing is an architecture where there is no special machine or machines that provide a service or manage the network resources. Instead all responsibilities are uniformly divided among all machines, known as peers. Peers can serve both as clients and servers.Cloud Computing is a modern computing paradigm that providing IT infrastructure and essential services i.e. Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS)[4].Cloud computing is an important model for enabling ubiquitous, convenient, on demand network access to a shared pool of configurable computing resources like networks, servers, storage, applications, and services[5].Cluster Computing (Highly Coupled) refers typically to a cluster of machines that closely work together, running a shared process in parallel. The task is subdivided in parts that are made individually by each one and then put back together to make the final result[6]. Cluster is the combination of applications and networks that are parallel processed and distributed computation. Cluster is easily defined as the technique of linking between two or more computers into a local area network. Virtualization (Space based) refers to an infrastructure that creates the illusion of one single address-space. Data are transparently replicated according to application needs. Decoupling in time, space and reference is achieved [8].Virtualization means to create a virtual version of a resource or device, like storage device, server, network or even an OS where the framework divides the resource into one or more execution environments. Ganeti is an

important cluster virtualization system developed by Google which is very lightweight, simple to install as well as manage, and it does not demand any special storage hardware .

Grid Computing is a distributed computing infrastructure facilitating the coordinated sharing of computing resources within organizations and across geographically dispersed sites [9]**.** The Grid arrangements follow distributed and parallel computing paradigms together which allows heterogeneity, portability, resource cooperation and dynamic allocation of resources.

Data mining is often described as deriving knowledge from the stored data. Analyzing data in a distributed fashion is called as Distributed data mining. Grid environments can be used for computation related tasks as well data related tasks. Grid based distributed data mining addresses to a structure where data mining could be applied to geographically distributed environments. Grid-based architectures that support distributed knowledge discovery have been developed throughout the world for different kind of applications such as Knowledge Grid, GridMiner, AdaM toolkit etc.

This review paper is divided into five sections. Section II presents Grid Distributing computing environment including architecture and types of Grid technology, section III discusses about distributed data mining technique and how it works. Section IV describes about grid based distributed data mining and represent some architectures based on this concept. Final section concludes the research and sets new research directions.

II. 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[10]**.** For extensive resource sharing, pioneering applications and high-performance orientation a computational model has been projected known as Grid computing. Grid is represented as a single unified resource to the user which logically coupled an infinite number of computing devices ranging from high performance systems to specialized systems[11], [12].

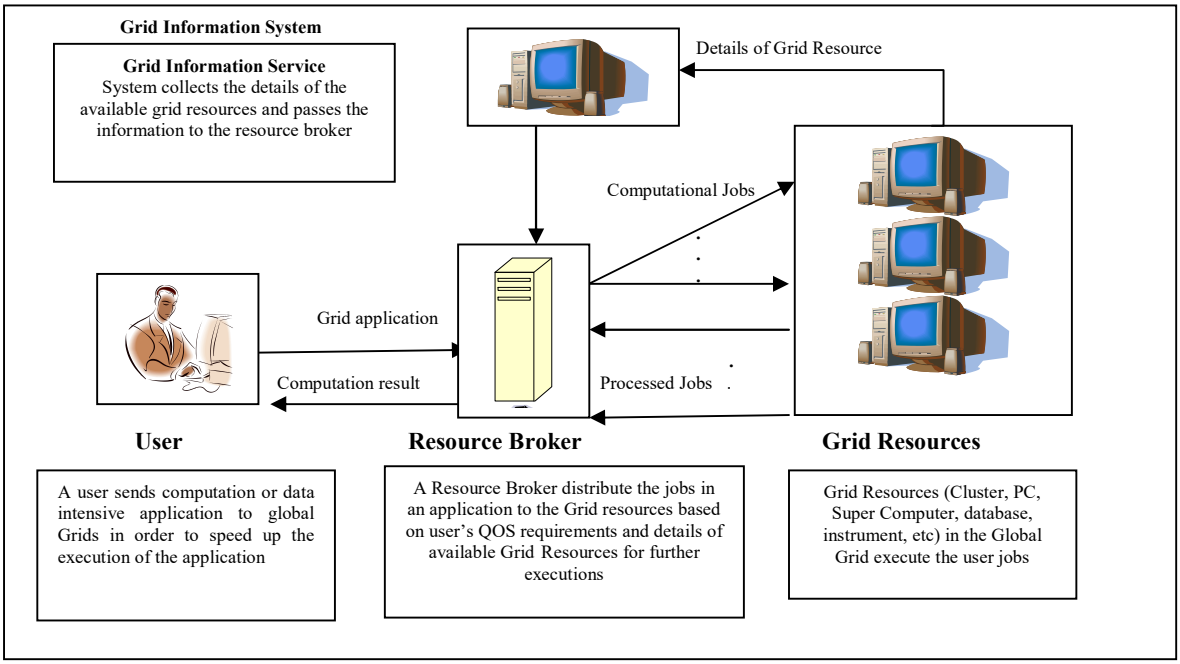

Figure 1. Grid Environment

For extensive resource sharing, pioneering applications, and high-performance orientation a computational model has been projected known as Grid computing[13]. Grid is presented as a single unified resource to the user which logically coupled an infinite number of computing devices ranging from high performance systems to specialized systems. Figure 1 shows that globally distributed Grid resources can be accessed by a Grid user simply by interacting with a Grid resource broker. The SOA (Service Oriented Applications) model is implemented using Open Grid Services Architecture (OGSA). The WS-Resource Framework (WSRF) has been adopted as an advancement of OGSA implementations[14].

A Grid Architecture

1. *The Layered Grid Architecture* organizes various grid capabilities and components such that high level services are built using lower-level services. Figure 2 depicts the layered grid architecture having four layers i.e Grid Fabric, Core middleware, user level middleware, Grid Applications explained in below section[15].

   1. 1 Grid fabric software layer provides management of resources as well as environment for execution at local Grid resources i.e. computers running a variety of operating systems, storage devices, and special devices such as radio telescope or heat sensor.

   1.2 Core Grid middleware layer provides Grid infrastructure and vital services. These services are related to storage access, trading, accounting, payment, security and information services. Resource trading based on the computational economy approach which is suitable for decentralization of Grids.

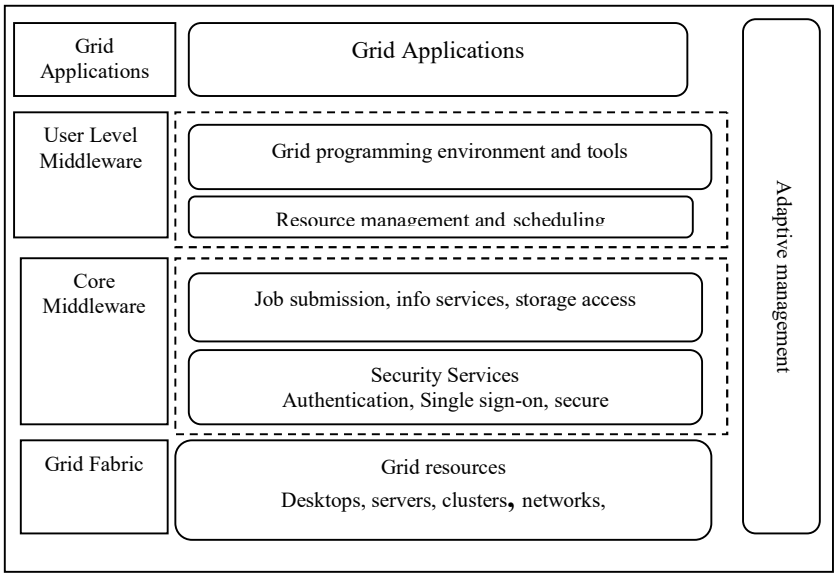

Figure 2. Layered Grid Architecture

   1.3 User-level middleware layer provides environments for programming in order to develop various types of applications and resource broker is used to select appropriate and exact resources in the context of applications.

   1.4 Grid applications Layer allows end-users to operate Grid services. Consequently, a number of web portals are being built since they allow users to universally access any resource from anywhere over any platform at any time.

2. *Hourglass Model* shown in figure 3 delineates a fundamental set of core construct and protocols, over which many different high-level actions can be applied and that further can be mapped onto many different underlying technologies [16].

   2.1 Grid Fabric (Interfaces to Local Control) layer provides the shared access to resources that is arbitrated by Grid protocols e.g. computational resources, storage systems, catalogues, network resources, and sensors.

2.2 Grid Connectivity (Communicating easily and securely) layer defines communication and certification protocols required for network transactions over Grid-specific applications.

2.3 Grid Resource (Sharing Single Resource) layer provides the security, initiation, supervising control, accounting, and payment of sharing operations on individual resources to the protocols of connectivity layers.

2.4 Grid Collective (Coordinating Multiple Resources) layer includes protocols and services in the form of APIs and SDKs that are associated to collections of resources.

2.6 Grid Application layer in hourglass model of Grid architecture comprises the user applications that operate within a VO (Virtual Organization) environment. Applications are constructed by using the services defined at any lower layer.

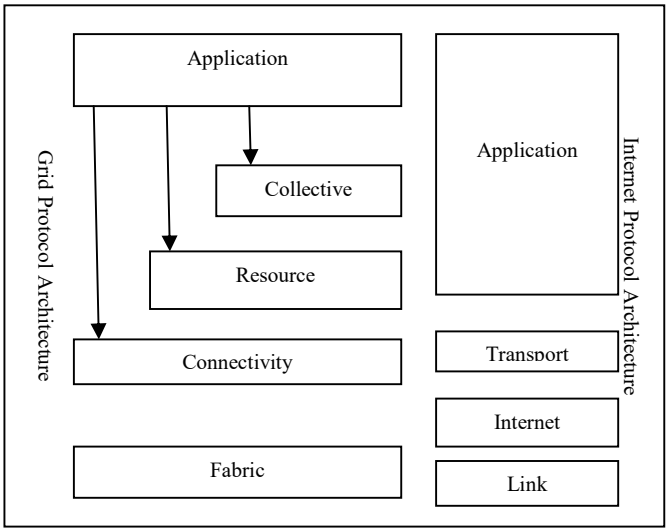

Figure 3. Hourglass Grid Architecture

B. Types of Grid

1. *On the basis of Nature of Grid* - in this type of classification, types of grid are categories on the basis respective nature or kind of work performed by the particular grid. Figure 4 shows that utility grid is related to user whereas computational grid shows relationship with infrastructure [17].

1.1 Computational Grid collects the computational power of globally distributed computers (e.g. TeraGrid, ChinaGrid).

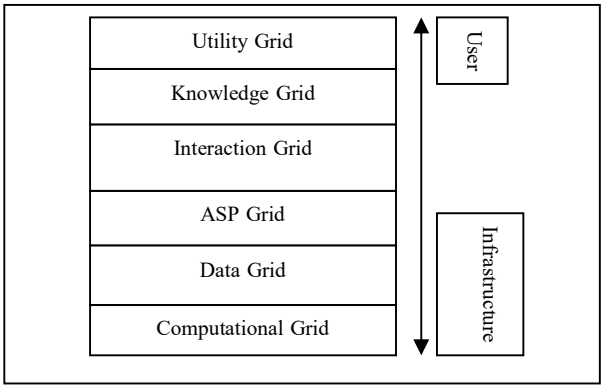

Figure 4. Types of Grid based on Nature

1.2 Data Grid focuses on global-level supervision of data in order to admission, amalgamation and processing of data contained in distributed data repositories (e.g. LHCGrid and GriPhyN).

1.3 Interaction Grid is related to interaction and joint revelation between participants (e.g. Access Grid).

1.4 Knowledge Grid focuses on knowledge acquirement, processing, management and provides analytical services driven by incorporated data mining services (e.g., Italian KnowledgeGrid and EUDataMiningGrid).

1.5 Utility Grid provides Grid services to end-users as IT utilities on a subscription basis. To meet contending demands from multiple users and applications allocation of resources has to be done (e.g., GridBus and Utility Data Center).

2. On the basis of levels of complexity for the enterprise

Grid classification based on the complexity in context of enterprise [18] such as infra grid talks about only a division where as inter grid talks about resource sharing between companies over web.

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

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

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

2.4 Inter-Grid enables 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.

III. DISTRIBUTED DATA MINING

Data mining algorithms are helpful in digging out hidden previously unknown information from existing data. Data mining algorithms can be applied using three different approaches i.e. Centralized approach, Parallel approach and last but not least is Distributed approach.

In *centralized approach*[19] of data mining, data is acquired and collected on a centralized store after cleaning and preprocessing. Task relevant data is selected from central store and mining techniques are applied. Xindong Wu [20] presented top 10 algorithms mostly used by the analysts of the world.

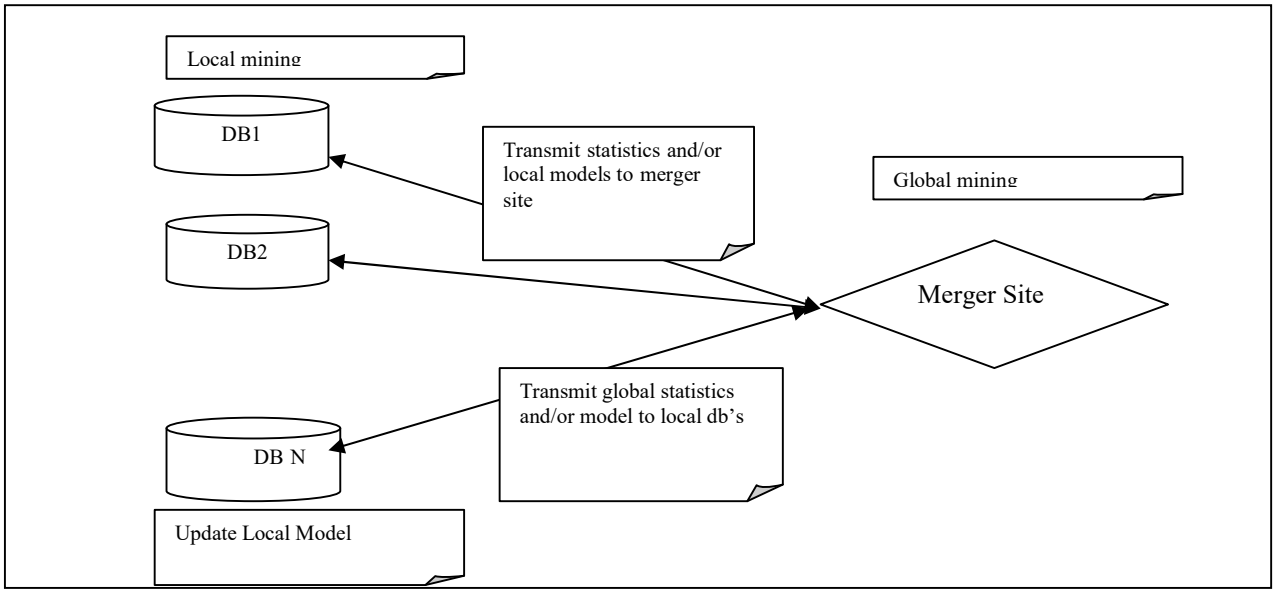

Figure 5. Distributed Data Mining Technique

In *Parallel approach* of Data Mining, multiple processors are used to perform data mining in a parallel environment. Parallel data mining [21] can be differentiate on the basis of task-parallel and data-parallel approaches.

In *Distributed approach* of Data Mining (DDM) proposes to get the global knowledge from the local data at distributed sites. In context of Distributed data mining [19], [22]–[24], approach concentrates on the impact of distribution of users, software and computational resources during the data mining process. Figure 5 depicts the procedure of distributed data mining, Local mining get done on local data bases then the result is transferred to the merger site and global mining performed. Now time to feed back and update local system with the result of global mining. In the DDM literature, data is distributed across sites: homogeneously and heterogeneously. In the homogeneous case, partition of global table is horizontal[18]. In the heterogeneous case, the global table is partitioned vertically, each site contains a collection of columns.

IV. DISTRIBUTING MINING WITH GRID COMPUTING

Grid based distributed data mining addresses to a structure where distributed data mining could be applied to geographically distributed environments. Following are the Grid-based architectures that support distributed knowledge discovery -

Knowledge Grid has been designed by Taila Domenico, Cannatro Mario[28]–[31]. 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

GridMiner, Brezany P. et al. [32], 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.

DataMiningGrid has been developed by Stankovski et al.[33],[34] 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.

GATES (Grid based Adaptive Execution on Streams) GATES[35] is a grid based middleware for processing of data streams built on top of Globus 3.0 and uses the concept of OGSA (Open Grid Services Architecture). GATES is easy to deploy in distributed environment and meets the real time constraint by self adaption and analyzing of one or more than two data streams with the help of at least two stages.

Multi-Domain Architecture-for distributed data streams Architecture developed by Talia et al.[36], [37] combines parallel and distributed paradigms for mining of continuous data streams from various nodes. They calculated frequent items in the streams using sketch algorithm and frequent itemsets by hybrid multipass analysis.

Environmental Scenario Search Engine (ESSE) for Data grids [15] provides uniform access to heterogeneous distributed environmental data archives and allows the use of human linguistic terms while querying the data. Environmental data access scenarios include mostly the retrieval of one-dimensional time series or two-dimensional geographic grids from three- or four-dimensional arrays of data. To support these kinds of operation they have developed a separate OGSA-DAI data resource component, the ESSE Data Resource, and a set of corresponding activity components. They have added a data mining activity fuzzy Search.

Anteater [15] is a service-oriented architecture for data mining that relies on Web services to achieve extensibility and interoperability. However, unlike the other systems, it doesn't support grid standards such as WSRF or OGSA. Moreover, Anteater requires data mining applications to be converted into a*lter-stream* structure which increases scalability.

Algorithm Development and Mining System (ADaM) toolkit has been designed by Ramachandran et al. [38]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.

Various grid based distributed data mining systems have been designed by the researchers for diverse type of applications such as Knowledge Grid for Business specific tasks, Grid Miner for e-Science, ADaM toolkit mainly used in earth observation projects etc. In the past decade, with the advancement in data collection and generation technologies, a new class of application has emerged that requires managing data streams, i.e. data composed of continuous sequence of items. A significant research effort has already been devoted to stream data management and data stream mining. Many systems use a centralized model in which the distributed data streams are directed to one central location before they are mined. Such a model is limited in many aspects such as Computational cost, Communication cost, Storage cost[39]-[43].

V. CONCLUSION

Distributed Data mining is described as deriving global knowledge from the local input data at distributed sites. Grid based distributed data mining addresses to a structure where data mining could be applied to geographically distributed environments. Designing such a structure is a computational challenge in the field of data mining. 'Grid' refers to persistent computing environments that enable software applications to integrate processors, storage, networks, instruments, applications and other resources which are managed by diverse organizations in widespread locations. However, there are many potential extensions of this work is available in research literature of data mining towards comprehensive, productive and high-performance analysis of various data sets.

In today's scenario, there is a strong need to mine distributed data streams for network analysis, sensor network monitoring, moving object tracking, financial data analysis and scientific data processing. Centralized models for distributed data streams have higher computation cost, communication cost and storage cost. So, we can conclude that we need to build an efficient integrated model for distributed stream mining using Grid environment.

REFERENCES

[1] "Distributed computing - Wikipedia.", https://en.m.wikipedia.org/wiki/Distributed_computing.

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

[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 , vol. 5, no. 2, pp. 144–148, Feb. 2015.

[4] R.-S. Petre and others, "Data mining in cloud computing,", *Database Systems Journal.*, vol. 3, no. 3, pp. 67–71, 2012.

[5] R. Vrbić, "Data Mining and Cloud Computing,", *Journal of Information Technology and Applications (Banja Luka),* vol. 4, no. 2, pp.75-87,Dec. 2012.

[6] 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", *Future Generation Computer System*, vol. 25, no. 6, pp. 599–616, Jun. 2009.

[7] H. AlHakami, "Comparison Between Cloud and Grid Computing: Review Paper", International Journal on Cloud Computing: Services and Architecture, vol. 2, no. 4, pp. 1–21, Aug. 2012.

[8] S. Jafari and M. Raeisi, "Compare Cloud and Grid Computing", International Journal of Computer and Electronics Research , vol. 3, no. 3, pp. 142–146, 2014.

[9] G. Marin, "Grid Computing Technology," *Database Systems Journal*, vol. 2, no. 3, pp. 13–22, 2011.

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

[11] P. Plaszczak and R. Wellner, *Grid computing*. Amsterdam ; Boston: Elsevier/Morgan Kaufmann, 2006.

[12] M. Joshi, "Grid computing," presentation, Apr-2005.

[13] A. Sanchez, J. Montes, W. Dubitzky, and P. de miguel, "Data Mining Meets Grid Computing: Time to Dance? ", *Data MIning Techinques in Grid Computing Environments*, 1st ed., pp. 1–16, 2009.

[14] M. A. Hussain, M. Naser, A. Begum, N. Shaik, and M. Shaik, "DataMining with Grid Computing Concepts," American Journal of Engineering Research, vol. 4, no. 7, pp. 256–260, 2015.

[15] M. Zhizhin, A. Poyda, D. Mishin, D. Medvedev, E. Kihn, and V. Lyutsarev, "Grid-Based Data Mining With the Environmental Scenario Search Engine (ESSE)" , *Data Mining Techniques in Grid Computing Environments*, W. Dubitzky, Ed. Chichester, UK: John Wiley & Sons, Ltd, pp. 221–245,2009.

[16] H. Bidgoli, "*The Handbook of Computer Networks,Distributed Networks, Network Planning, Control, Management, and New Trends and Applications "*, vol. 3. 2007.

[17] M. S. Pérez, A. Sánchez, V. Robles, P. Herrero, and J. M. Peña, "Design and implementation of a data mining grid-aware architecture", *Future Generation Computer Systems*, vol. 23, no. 1, pp. 42–47, 2007.

[18] G. Pirró, C. Mastroianni, and D. Talia, "A framework for distributed knowledge management: Design and implementation", *Future Generation Computer System*, vol. 26, no. 1, pp. 38–49, Jan. 2010.

[19] B. B. Ahamed and S. Hariharan, "A Survey On Distributed Data Mining Process Via Grid", *International Journal of Database Theory and Application*, vol. 4, 2011.

[20] X. Wu, vipin Kumar, and others, "The Top Ten Algorithms in Data Mining", *Knowledge and information systems*, vol. 4, no. 1, pp. 1–37, Dec. 2007.

[21] S. Devi, "A survey on distributed data mining and its trends", *International Journal of Research in Engineering & Technology*, vol. 2, pp. 107–120, 2014.

[22] G. Tsoumakas and I. Vlahavas "Distributed Data Mining of Large Classifier Ensembles" in Proceedings of Companion Volume of the Second Hellenic Conference on Artificial Intelligence , pages 249–256, Thessaloniki, Greece, April 2002.

[23] S. Masih and S. Tanwani, "*Data Mining Techniques in Parallel and Distributed Environment-A Comprehensive Survey"*, International Journal of Emerging Technology and Advanced Engineering, 3rd ed., vol. 4, 2014.

[24] S. Masih and S. Tanwani, "Distributed Framework for Data Mining As a Service on Private Cloud", International Journal of Emerging Technology and Advanced Engineering, vol. 4, no. 11, pp. 65–70, Nov. 2014.

[25] S. Masih and S. Tanwani, "Data Mining Techniques in Parallel and Distributed Environment-A Comprehensive Survey" , International Journal of Emerging Technology and Advanced Engineering, vol. 4, no. 3, p. 9, 2014.

[26] G. Tsoumakas and I. Vlahavas, "Distributed data mining", *Encyclopedia Data Warehousing and Mining*, 2009.

[27] M. Cannataro, D. Talia, and P. Trunfio, "Distributed data mining on the grid", *Future Generation Computer Systems*, vol. 18, no. 8, pp. 1101–1112, 2002.

[28] 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.

[29] 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.

[30] G. Bueti, A. Congiusta, and D. Talia, "Developing Distributed Data Mining Applications in the Knowledge Grid Framework", in *High Performance Computing for Computational Science - VECPAR 2004*, vol. 3402, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 156–169, 2005.

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

[32] P. Brezany, J. Hofer, A. M. Tjoa, and A. Woehrer, "Gridminer: An infrastructure for data mining on computational grids," in *Proceedings of Australian Partnership for Advanced Computing Conference (APAC)*, 2003.

[33] V. Stankovski, J. Trnkoczy, M. Swain, W. Dubitzky, V. kravtsov, A. Schuster, T. Niessen, D. Wegener, M. May, M. Rohm, and J. Franke, "Digging Deep into the Data Mine with DataMiningGrid," *IEEE Computing Society*, pp. 69–76, Dec. 2008.

[34] V. Stankovski, M. Swain, V. Kravtsov, T. Niessen, D. Wegener, J. Kindermann, and W. Dubitzky, "Grid-enabling Data Mining Applications with DataMiningGrid: An architectural perspective", *Future Generation Computer System*, vol. 24, no. 4, pp. 259–279, Apr. 2008.

[35] L. Chen, K. Reddy, and G. Agrawal, "GATES: a grid-based middleware for processing distributed data streams",in *13th IEEE International Symposium Proceedings on High performance Distributed Computing*, pp. 192–201, 2004.

[36] E. Cesario, C. Mastroianni, and D. Talia, "A Multi-Domain Architecture for Mining Frequent Items and Itemsets from Distributed Data Streams", *Journal of Grid Computing*, vol. 12, no. 1, pp. 153–168, Mar. 2014.

[37] E. Cesario, A. Grillo, C. Mastroianni, and D. Talia, "A Sketch-Based Architecture for Mining Frequent Items and Itemsets from Distributed Data Streams", pp. 245–253, 2011.

[38 R. Ramachandran, S. Graves, J. Rushing, K. Keyze, M. Maskey, H. Lin, and H. Conover, Eds., "ADaM Services: Scientific data mining in the service-oriented architecture paradigm", in Data mining techniques in grid computing environment, Oxford: John Wiley & Sons, Ltd, pp. 57–69, 2009.

[39] L. Golab and M. T. Özsu, "Issues in data stream management," *ACM Sigmod Record*, vol. 32, no. 2, pp. 5–14, 2003.

[40] M. Charikar, K. Chen, and M. Farach-Colton, "Finding frequent items in data streams," in *Automata, languages and programming*, Springer, pp. 693–703,2002.

[41] A. Ghoting and S. Parthasarathy, "Facilitating interactive distributed data stream processing and mining," in *Proceedings of 18th International Symposium on Parallel and Distributed Processing* , p. 86, 2004.

[42] B. Babcock, S. Babu, R. Motwani, and M. Datar, "Chain: Operator scheduling for memory minimization in data stream systems," in *Proceedings of ACM SIGMOD international conference on Management of data*, pp. 253–264, 2003.

[43] G. Krempl, I. Žliobaite, D. Brzeziński, E. Hüllermeier, M. Last, V. Lemaire, T. Noack, A. Shaker, S. Sievi, M. Spiliopoulou, and others, "Open challenges for data stream mining research," *ACM SIGKDD Explorer Newsletter*, vol. 16, no. 1, pp. 1–10, 2014.

[44] C. C. AGGARWAL, *DATA STREAMS: MODELS AND ALGORITHMS*, 1st ed. NEWYORK: Kluwer Academic Publishers, 2007.