Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 06, June -2018

A SURVEY ON HEURISTIC QUERY BASED OPTIMIZATION USING DISTRIBUTED DATA MINING

Anusha Dasari, Ms. B. Jyothi

[1]Post Graduate Student, Dept. of CSE, Anurag Group of Institutions, Hyderabad, T.S., India.
[2]Assistant Professor, Dept. of CSE, Anurag Group of Institutions, Hyderabad, T.S., India.

ABSTRACT:-*New grid and clouds outputs for shared data extraction are required for finding the results of intensive workflows. In variation of the accepted workflows, the information between the jobs are transferred in form of files. The tasks are completed when they have processed the input data and produce the continuous flow of the outputs. Each task is working for longer period of time in order to get new data. Heuristic algorithm is used in cluster primitives' workflow. These clusters will act as one job so that the transferring of information between them is zero. Open grid service architecture is a distributed interaction and computing architecture based on services such as WSDL (Web Service Definition Language) and SOAP (Simple Object Access Protocol) is used to perform different operations between the sender and the receiver is in the form of request and response via internet.*

Index terms: *Grid computing, distributed data mining, scheduling algorithm.*

I. INTRODUCTION

Grid environment was developed addressing the needs of large scale to Data mining applications in e-Science. Digital data play major role in e-Science applications. The Grid provides both computational and data resources for scientific applications, it can equally be used by data mining applications in other domains Grid computing technologies, via dynamic and distributed virtual organizations, provide approach to use geographically distributed heterogeneous large scale resources. The key characteristics that are distinguishing Grid computing over the conventional technologies, according to the initial designers Foster & Kesselman, are: it is flexible, autonomous, secure, dynamic, service-oriented, robust, coordinated, scalable and it is a system which is transparent to the users.

Grid architectures are based on Grid standards and include components like: information system, resource broker, uniform access to the resources, security mechanisms, workflow composition and workflow scheduler, mechanisms for data access and data replication etc. Large scale data is the basis of the data mining applications on Grid, but due to predominantly file based organization of Grid data, there is an open issue of how to use data which is originally organized as shared and structured collections, stored in databases, in structured documents or in assemblies of binary files. These Grid application requirements require that the developers of Grid systems provide practical implementations of Grid databases (or data access integration) .Data mining algorithms have several stages, from preprocessing to generating output representation of the results ,and these stages should be executed in some form of a work flow .

Grid architecture for data mining should provide model for distributed data mining with easy creation of data mining workflows, providing most of the functionalities for distributed data mining and optimization of these workflows. Grid middleware should be adapted to be installed and to use the capabilities of today's commercial cloud computing solutions.

The Open Grid Service Architecture for Data Mining (OGSA-DM) is based on popular and well used OGSA-DAI distributed database which is built according to the Web Service Data Access and Integration and Grid Database Service Specification standards form Global Grid Forum. OGSA-DM consists from several subsystems and components which enhance the whole data mining process, such as: distributed and heterogeneous data organization, meta-learning model for knowledge acquisition, virtual data mining views, execution of external jobs, subsystem for monitoring workflow execution and measuring system performance, etc.

One of the main capabilities of OGSA-DM is the execution of data intensive workflows. Standard Grid system have mechanisms of executing workflows in which data are transferred between the jobs in form of files, job starts to execute when it receives all input files, and finally, job is finished when it processes the input files. In tightly-coupled architectures such as OGSA-DM, data intensive workflows are executed, in which: data is grouped in blocks and flows between jobs, i.e. edges in direct acyclic graph (DAG) as data streams, jobs are continuously running as new data arrives on input, jobs run for a long period of time (after one block of data is processed, the job is not finished, instead it continues with the next block), jobs are executed at the same time (they overlap) and they can process different (or the same) block of data according to this property, one server can be executing several jobs at the same time). The rest of the paper is structured as follows:In Section II, survey report. A comparison of surveyed Heuristic Query based Optimization using distributed data mining and concludes the paper and discusses several open research issues in section III.

II. LITERATURE SURVEY

This section is a brief review of the research papers to address various Optimization Approach and solutions. This conducting survey by the various research papers.

Related Work:

2.1 Achieving Privacy Protection Using Distributed Load Scheduling: A Randomized Approach

Engong Liu and Peng Cheng[1] developed achieving privacy Protection using distributed load scheduling to minimize a weighted sum of privacy leakage, electricity bill, and the user experience sacrifice by exploring the capacity of all end-devices of residential users including three categories of shift able devices, i.e., thermostatically controlled devices, rechargeable batteries and successive operation devices

To protect the residential users' privacy from malicious attacks and minimize the information leakage even when the attackers have gain access to the meter data. One typical way is to adopt encryption mechanism is novel Efficient and Privacy-Preserving Aggregation (EPPA) scheme for smart grid communication based on homomorphic Paillier cryptosystem. and the other is distributed data aggregation algorithm[1] in which data aggregation is performed at all smart meters. The third way is called battery-based method, which uses energy storage devices like Rechargeable batteries to mask the load signature or perturb the power consumption data

Different from most of the existing works which mainly rely on the charging/discharging scheduling of rechargeable batteries installed in the house, three kinds of shiftable devices to track a pre-specified aggregated load profile so that the raw smart meter data will be distorted and the risks of illegitimate inferring personal habits can be mitigated. The main objective is to reduce both residential user's electricity cost and potential privacy leakage while providing certain user satisfaction.

2.2 Scalability, Elasticity, and Efficiency in Cloud Computing

Sebastain et al.[2] developed Scalability, Elasticity, and Efficiency in Cloud Computing explains the common metrics for scalability, elasticity and efficiency. According to [2],

Scalability: Scalability is the ability of a cloud layer to increase its capacity by expanding its quantity of consumed lower-layer services. Data synthesis indicates that scalability is a well-established property used to describe the behavior of all types of distributed systems, not just cloud computing Found scalability metrics are generally not restricted to a particular role.

Elasticity: Elasticity is the degree a cloud layer autonomously adapts capacity to workload over time. Elasticity is the ability of a software system to dynamically scale the amount of the resources it provides to clients as their workloads increase or de- crease. They also state that elasticity consists of the temporal and quantitative properties of runtime resource provisioning and un provisioning", Elasticity requires scalability because a system that cannot scale is inherently inelastic. Elasticity also improves the efficiency of a system because it typically reduces operational cost, power consumption, etc.

Efficiency: the amount of resources consumed for processing a given amount of work. Existing efficiency metrics measure the amount of lower-layer services" over time based on the aspect of interest definitions focus on single cloud layers, e.g., power consumption.

Cloud providers and cloud consumers can use these concepts as common vocabulary and specify service level objectives based on metric.

2.3 A Multi-Agent Distributed Data Mining Model based on Algorithm Analysis and Task Prediction

Dan Zhou, et al.[3] developed A Multi-Agent Distributed Data Mining Model based on Algorithm Analysis and Task Prediction. [3] is implementing an approach Distributed Data Mining using Multi-Agent technology and proposes a Multi-Agent Distributed Data Mining mode. Each Agent is only responsible for specific duties. Agents communicate and coordinate with each other, which enhances the privacy and confidentiality of data. An Algorithm Analysis Agent is used to improve the intelligence and efficiency of Data Mining Agent in the model.To improve the performance of between agents and reduce the pressure of network bandwidth, Load Balancing Agent and Task Prediction agent. DDM is a process of discovering unanticipated knowledge from logically or physically distributed database using distributed computing technology. An Algorithm Analysis Agent and Task Prediction Agent are used improve the mining performance and intelligence, and reduce the communication costs.

A. Multi-Agent Distributed Data Mining Model based on Algorithm Analysis and Task Prediction

Depending on functions model can be divided into Data Warehouse Mining (DWM) sub-model, Load Balancing (LB) sub-model and User sub-model. DWM sub-model consists of Data Mining Agent, Data Preprocessing Agent, Algorithm Analysis Agent, Task Prediction Agent, and Cache used for storing temporary data. User sub-model comprises User Agent and Global Data Mining Agent. The coordination among users and data warehouses in transmitting temporary data and allocating tasks are implemented by LB sub-model (or LB Agent).

B. Service Layer

Service layer is supported by the DWM sub-model. The sub model mine data after receiving requests. Thus, Data Mining Agent which provides mining and decision-making information for both users and other Agents is the core of this sub model. The sub model includes Data Preprocessing Agent, Algorithm Analysis Agent, Task Prediction Agent, and Cache in addition. Those Agents are described as follows

1. **Data Preprocessing Agent**.

This Agent extracts data from the data warehouse and preprocesses them, such as normalizing structured or unstructured data. It mainly has four functions: data extraction, data clean, data transformation and data simplification. There are some preprocessing methods, including the data concentration method based on concept tree, the reduction method based on rough set theory, the attribute selection method based on statistical analysis and clustering preprocess by use of genetic algorithms, and information theory thought method based on universal knowledge discovery. The main target of this Agent is providing normalized data.

2. **Algorithm Analysis Agent**

The Algorithm Analysis Agent has a set of mechanism to manage these algorithms. During the idle time of system, the Agent tests all algorithms of algorithm lib. According to mined results of common tasks of local data, a most effective algorithm would be used by Data Mining Agent for mining. The main target of Algorithm Analysis Agent is offering a most effective algorithm for Data Mining Agent.

3. **Data Mining Agent.**

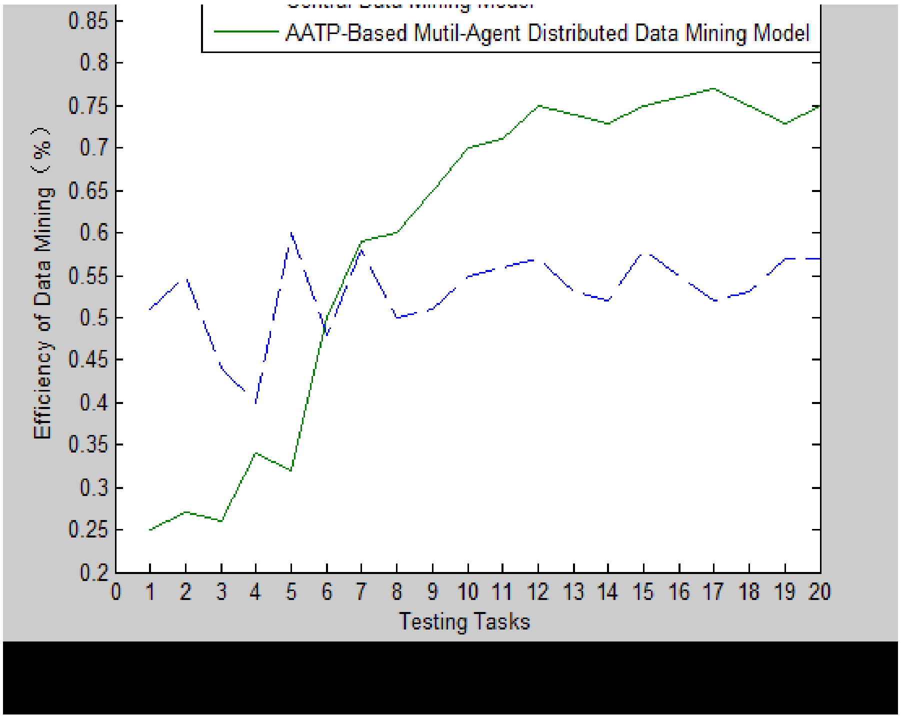
The Data Mining Agent accepts mining tasks distributed by LB Agent and other Agents requests. At the same time, the target Agent sends feedback information to the requesting Agents. After receiving tasks from LB Agent and other Agents, Data Mining Agent mines the normalized data which come from Data Preprocessing Agent, using the algorithm supplied by Algorithm Analysis Agent. When Data Mining Agent needs other agents' mined results, it can search the Cache. If the mined results cannot be found in Cache, a request to the target agent should be sent.

data mining is affected by the precision of data mining and the consuming time. The calculation of the efficiency of data mining can be expressed as follows:

$P = (1 - r) R1 / R2 + r *(T - t) / T \ (0 < r < 1)$ (1) (1) (1)

Notations are:

R1 — Results generated by current model.
R2 — The ideal mined results.
T — The average consuming time of current model.
t — The consuming time of current model.
r — The rate of influence which is chose from 0 to 1.
R1/R2 — The precision of current model



2.4 Integrated Queries over a Heterogeneously Distributed Scientific Database using OGSA-DQP

X.Xiang[4] developed Integrated Queries over a Heterogeneously Distributed Scientific Database using OGSA-DQP according to [4] explains regarding the slogan digital sky survey(SDSS) and the e-Science technologies OGSA-DAI and OGSA-DQP

Slogan Digital Sky Survey (SDSS)

Distributed SDSS explains distributed queries using UNION ALL operations with three nodes OGSADAI and OGSA-DQP is used to integrate the data across different sites for forming a logical distributed database system . OGSA-DQP coordinator service to create an OGSA-DQP
Coordinator

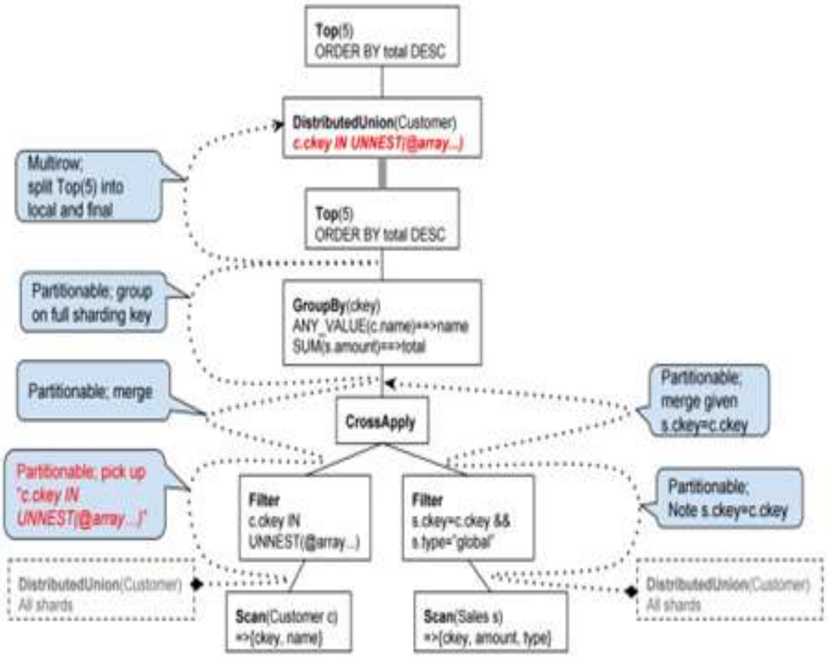


Figure 1: Distributed plan for a top-k aggregation query. The IN condition results in addition of sharding key predicate to Distributed Union. Dotted lines show the paths of operator pull-up.

The UNION operation is used to combine the results of two or more SELECT queries together. OGSA-DQP has two types of UNION operation: pure UNION and UNIONALL. The pure UNION operator only includes distinct values while UNION ALL operator includes all values. Because pure UNION must check for repeated values it is much more expensive, and these checks are not necessary. UNION or UNION ALL query must have the same number of fields in the result sets, and all corresponding columns must be of the same data types

2.5 A Heuristic Query for Heterogeneous Environments

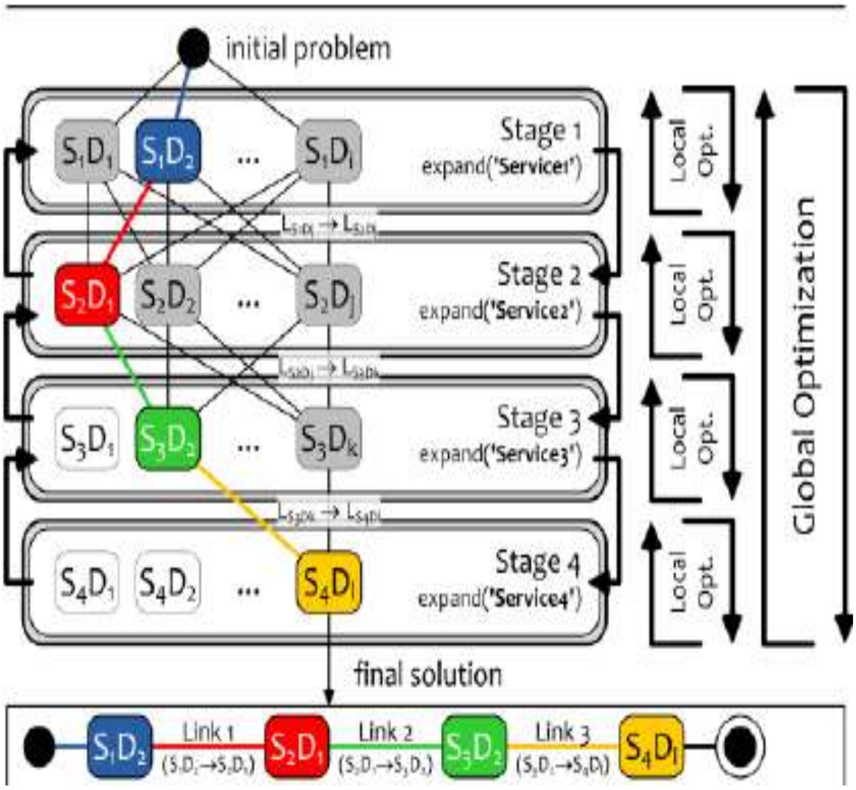
According to Peter Paul, et al. [6] main aim of heuristic query optimization multi-layered blackboard Mechanism is parallel query execution plans in heterogeneous environments. This survey mainly focus on the performance of distributed query depends on the quality of the constructed QEP. That is to say the performance of the used relational operations (e.g. sort,join) and their optimal sequence performed on data sets derived from a multitude of distributed databases

Blackboard Query Optimization

Global blackboard: It is an conventional database which explains regarding the sharing information about input data and partial solutions.

Knowledge base: It is a storage facility which contain independent units which r known as regions. Each region is maintained by a single expert and communication between regions is performed using the global blackboard.

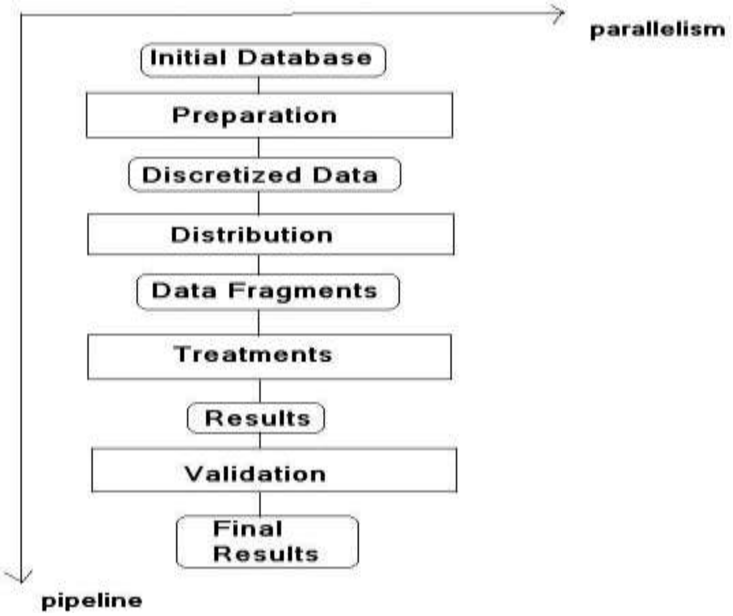
Control component: It is used to make reasonable decisions each region has to provide cost estimations for their applied operations to build up a cost-based decision tree. The decision tree consists of nodes, which resemble partial solutions to the problem and their estimated costs. In a step-wise approach the algorithm starts at the root and expands nodes whenever the estimated costs associated with that node are lower than the costs of any given solution that has already been found. Solutions which are heuristically known or proven to be inefficient are discarded automatically. After completion of a layer the process continues with the next layer.



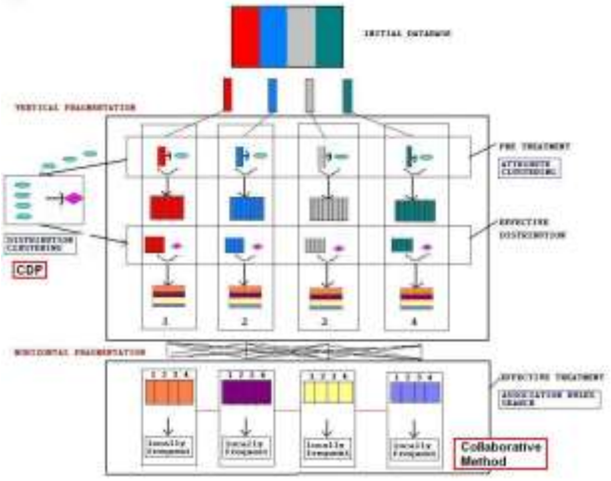
2.6 Optimal Grid Exploitation algorithms for Data Mining

Valerie FIOLET presented Optimal Grid Exploitation algorithms for Data Mining. According [6] Data Mining tasks have been parallelized and can be executed on dedicated clusters solutions. The DisDaMin mechanisms first implement a specific fragmentation of the data using

Clustering methods, and then realize asynchronous collaborative techniques according to the specifics of execution on grids. The use of this fragmentation method makes it possible to carry out optimal local processing on each node, with a minimum of communications. Using this, the distributed algorithm DICCoop is introduced an adaptation of *of DIC*. By parallelizing in a specific way, it is intended to obtain gain not only from parallel execution but also from complexity of technique.In the DisDaMin , an intelligent data distribution for the problem of association rules is used (data fragmentation on which the distribution of techniques depends). Computing fragments according to this intelligent distribution leads to a decrease of the global complexity of the problem. The aim is to allow the technique of huge quantities of data and to maximize the exploitation of resources (CPU,storage).The special configuration of a grid (without central memory)forces the distribution of the database into fragments and to handle these fragments in order to have optimal exploitation of parallelism for that kind of architecture. Becauseof the high communication cost in this environment,



The General Resolution Method Of Disdamin Project For The Association Rules Problem.



2.7 Assessing the Dependability of OGSA Middleware by Fault Injection

Nik Looker and Jie Xu developed Assessing the Dependability of OGSA Middleware by Fault Injection. According to [7] our research on devising a dependability assessment method for the upcoming OGSA3.0 middleware using network level fault injection. with the requirements of testing OGSA middleware and derive a new method and fault model. From this we have implemented an extendable fault injector framework and undertaken some proof of concept experiments with a simulated OGSA middleware system based around Apache SOAP and Apache Tomcat. We also present results from our initial experiments.

A. Synchronous and Asynchronous Transfers

DCE based distributed systems tend to be synchronous in nature, although both CORBA and COM+ now provide asynchronous capabilities. The main method used to implement asynchronous operations is polling. An asynchronous interface is compiled from the IDL with multiple stub methods generated for each call, i.e. one to send the message, one to poll for a response, on to receive a response.

B. Authentication and Encryption

OGSA systems utilize a much higher level of authentication and encryption than classical distributed systems because of the environment they are expected to run in. Whilst most classical distributed systems have mainly been run across private LANs, OGSA systems are envisaged to run across the Internet between geographically different sites, possibly run by different organisations. OGSA systems may utilize WS-Securityto provide end-to-end security.

2.8 Performance Effective and Low Complexity Task Scheduling for Scheduling for Heterogeneous Computing

Haluktopcuoglu,et.al [8] worked on performance and task scheduling .According to [2] A Heterogeneous computing system requires couple time and runtime support for executing applications. The efficient scheduling of the task of an application on the available resources is one of the key factors for achieving high performance. The general task scheduling problem includes the problem of assigning the task of an application. they presented two algorithms called HEFT and CPOP.HEFT algorithm selects the task with the highest upward rank value

III COMPARISON AND CONCLUSIONS

protocol	Highlighting features	requirement	Weakness or overhead
Engong Liu and Peng Cheng[1]	To reduce both residential user’s electricity cost and potential privacy leakage while providing certain user satisfaction.	To protect the residential users’ privacy from malicious attacks and minimize the information leakage even when the attackers have gain access to the meter data.	The feasibility of protecting users’ behavioral privacy by using method of load scheduling
Sebastain,et al[2]	It describes regarding defination and metrics	Cloud providers and cloud consumers specify service level objective based on metric	To externally execute and, analyzing its reproducibility of the metric and re executions can also vary parameter
Dan Zhou,et al[3]	Implementing of distributed data mining using multi agent in distributed data mining	for improving the performance of the communication	new connection between agents caused by communication delay and fault tolerance of system

Helen X.Xang[4]	integrating data results from data service resources based on heterogeneous underlying database manage ment systems.	examines the running of several OGSA-DQP queries against the database	To run the union queries across multiple machines efficiently
Peter Paul, et al[5]	Research focuses on query optimization of Distributed data base quires, considering a huge variety on different infra structure and algorithm	QET is used in domain distributed database	To cope with scalability for the query optimization process
Valerie FIOLET [6]	allow the technique of huge quantities of data and to maximize the exploitation of resources(CPU,storage).	implement a specific fragmentation of the data using clustering methods, and then realize asynchronous collaborative techniques according to the specifics of execution on grids.	observation results on load balancing and optimized placement of data fragments and components on the Grid.
Nik Looker and Jie Xu[7]	our research on devising a dependability assessment method for the upcoming OGSA 3.0 middleware using network level fault injection	defining a testing method for GRID middleware a new set of challenges are faced which require different solutions	Whilst this coding error is minor, its discovery is significant because it has gone unnoticed for a long period of time and was uncovered with a minimum of time and effort using this technique
Haluktopcuoglu,et al[8]	A Heterogeneous computing system requires couple time and runtime support for executing applications	They presented two algorithms called HEFT and CPOP.HEFTalgorithm selects the task with the highest upward rank value	Advancement in various algorithms

REFERENCES

[1]Endong Liu, Peng Cheng -“Achieving Privacy Protection Using Distributed Load Scheduling: A Randomized Approach” Member, IEEE,2017.

[2] Sebastian Lehrig, Hendrik Eikerling, Steffen Becker-“ Scalability, Elasticity, and Efficiency in Cloud Computing” Chemnitz University of Technology, Chemnitz, Germany,2015

[3] Dan Zhou, Wenbi Rao, Fangsu Lv-“ Multi-Agent Distributed Data Mining Model based on Algorithm Analysis and Task Prediction” dept. of Computer Science and TechnologyWuhan University of Technology ,2013

[4] Helen X.- Xiang-“Integrated Queries over a Heterogeneously Distributed Scientific Database using OGSA-DQP” ,Computer Science, University of Hertfordshire, UK,2011

[5] Peter Paul Beran, Werner Mach, Ralph Vigne, J¨urgen Mangler and Erich Schikuta –“A Heuristic Query for Heterogeneous Environments” Department of Knowledge and Business Engineering Rathausstr. 19/9, A-1010 Vienna, Austria 2010

[6] Valerie FIOLET, Richard Olejnik, Guillem Lefait, Bernard Toursel-“Optimal GridExploitation algorithms for Data Mining”, Computer Science Institute University of Mons-Hainault - Mons, Belgium, 2006

[7] *Nik Looker and Jie Xu* –“Assessing the Dependability of OGSA Middleware by Fault Injection”, Department of Computer Science, 2003

[8] Haluktopcuoglu, salim Hariri- “Performance Effective and Low Complexity Task Scheduling for Scheduling for Heterogeneous Computing “Min you senior in IEEE 2002.