International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP—89-95

Performance evaluation of query and query scrambling in distributed environment using probabilistic techniques

Prasad Suman Sourav[1], Mishra Sambit Kumar[2]
1(Department of MCA, Ajay Binay Institute of Technology, Cuttack)
Email: prasadsuman800@rediffmail.com
2 (Department of Computer Sc.&Engg., Gandhi Institute for Education and Technology, Baniatangi)
Email: sambitmishra@gietbbsr.com

ABSTRACT *: A large number of queries are posed on databases spread across the globe. In order to process these queries efficiently, optimal query processing strategies that generate efficient query processing plans are being devised. Usually, due to replication of relations at multiple sites, the relations required to answer a query may necessitate accessing of data from multiple sites. This leads to an exponential increase in the number of possible alternative query plans for processing a query. Though it is not computationally feasible to explore all possible query plans in such a large search space, the query plan that provides the most cost-effective option for query processing is considered necessary and should be generated for a given query. Relations may be replicated as well as fragmented at different sites in the system. The placement of data in the system may be determined by factors such as local ownership and availability concerns. When every site in the system runs the same database management software the system may be called homogenous, otherwise, it may be called as heterogeneous system. The complexity of query optimization is determined by a number of alternative query evaluation plans which grows exponentially with the number of relations involved in the query because a single query can be joined in several ways. Since all execution plans are equivalent in terms of their final output with a difference in cost and amount of time that they need to run, it is essential to optimize these query plans, join orders and join methods in modelling query processing. The query optimizer selects among the query evaluation plans responsible for generating the least estimated execution cost according to the given cost functions. Enumerative optimization strategies usually deal with the join queries to determine the best query plan to execute the query. In this paper, an attempt has been made to generate such optimal query plans and intended to focus on specific architecture i.e. relational database as a service which may package and deploy query evaluation plans The objective of this work is to unify the approaches of query scrambling and reductions to dynamic optimization of query execution plans at data integration stage. In particular we may remove the limitations of query scrambling to join expressions only and the limitations of reductions to computation of one operation at a time.*

KEYWORDS **-** *Query scrambling, distributed query, DBaaS, Virtual machine, metaheuristic optimization*

I. INTRODUCTION

In multi database systems distributed over wide area networks usually exhibits a number of complex performance problems. So selection of the best global query processing plan at pre processing stage and dynamic optimization of data integration operations at post processing stage have the most important impact on performance. A typical distributed multi database system consists of a central database system linked to a number of remote, autonomous and heterogeneous local database systems. A software layer installed at a central site makes distribution and heterogeneity of the local systems transparent to the end-users. In a typical multi database system users obtain a relational view of fully homogeneous and centralized database system. The queries submitted by the users are globally optimized, decomposed into the subqueries, and translated into the dialects of query languages available at the local systems. Then, a query processing coordinator optimizes a global processing plan and submits the subqueries to the local sites accordingly to this plan. Query optimization at post processing stage addresses the problem of effective integration of partial results into the final answer. In a multi database system where a global user’s view is the relational on a data integration procedure is formally represented as an expressions of relational algebra and called as data integration expression. Optimization of data integration expressions is conceptually different from the classical syntax based and cost based optimization of relational algebra expressions. A distributed database encompasses coherent data, spread across various sites of a computer network. A Distributed Database Management System deals with managing such distributed databases. DDBMS presents a simple and unified interface to users by providing them with access to the disparate databases, as if they were not distributed. The performance of a DDBMS is determined by its ability to process queries in an effective and efficient manner. The query processing problem is much more complicated in distributed environments, as there are various parameters affecting their performance. The required information for processing a distributed query is usually available at different sites. The query processing, thus, would involve transmission of data between these sites. These data transmissions, along with local data processing, constitute a distributed query processing (DQP) strategy for a query. Having data distributed across multiple sites is particularly advantageous for large organizations who have offices and their work force distributed across the world. As well as offering access to common data from multiple locations, having distributed data can offer other benefits including improving the availability of data. For example, by distributing data we avoid the single point of failure scenario where data resides at a single site that can fail. By replicating data across several sites in the event of a site failure there would be more chance of having an alternative copy of the required relation available. This would allow the service requesting the relation to continue without being disrupted. Replicating data not only improves availability but also improves load distribution. By creating copies of relations that are in high demand we reduce the load on owning sites, which may have an impact on reducing the number of site failures. By replicating relations to sites where they are most commonly used network accesses can be reduced, which can improve the performance for users at those sites. In distributed query processing, the distributed query is parsed before arriving at an effective query processing strategy for it. This strategy comprises of effective and efficient query processing plans that would decompose the distributed queries into local sub-queries to be executed at their respective sites. Also, the order and the site at which the results of the sub-queries are integrated is also part of this plan. The final integrated result is provided as the answer to the distributed query. Thus the DQP strategy aims to generate query processing plans that reduce the amount of data transfer between sites and thereby reduces the distributed query response time. The major costs incurred in DQP are CPU, I/O and the site-to-site communication cost. Among these, the site-to-site communication cost is the dominant cost. This cost can be reduced if fewer sites are involved in processing a distributed query. In order to process a distributed query, the data required may have to be obtained from several sites distributed over a computer network. Furthermore, as the number of sites containing the relations accessed by the query increase, the number of possible valid query plans also increases. So it becomes imperative to arrive at a query processing plan that entails an optimal cost for query processing. However, the number of such possible query plans increases exponentially with increase in the number of relations in the query and also with increase in the number of sites containing them. Thus, a large search space comprising all possible query plans needs to be explored in order to compute the optimal query plans. This exhaustive search is not computationally feasible. Further, this being a combinatorial optimization problem and can be addressed by techniques based on heuristics like greedy, evolutionary, and randomized. However, efficiency of these techniques is affected by the unconventional behavior, in specific instances, of the problem. An approach that generates close query plans with respect to the number of sites involved and the concentration of relations in the sites for a distributed relational query is given. Query processing over lesser number of sites would be more efficient and thus query plans involving fewer sites need to be generated. To support data processing for cloud users, it is important to provide scalable, reliable, highly available and highly efficient database services (DBaaS) in cloud. Currently users have three options when it comes to use databases. Cloud users currently have three DBaaS options to choose from. One) Run a traditional database management system (DBMS) inside a virtual machine (VM). The user must administer most system management activities, including software licensing, installation, configuration, management, backup, and recovery. This option is difficult to scale. These services are highly efficient and scalable for some types of large data, but the user is forced to deal with the lack of structured data and strong data integrity that is present in relational models.

II. Review of literature

Laurent Amsaleg et al.[1] have discussed that accessing data from widely distributed sources poses significant new challenges for query optimization and execution . Failures in the network can introduce highly variable response times for wide area data access. They introduced a class of dynamic run time query plan modification techniques called as query plan scrambling .They presented an algorithm that modifies execution plans on the fly in response to unexpected delays in obtaining initial requested tuples from remote sources

Tolga Urban et al.[2] have focused on remote data access from disparate sources across a wide-area network such as the Internet is problematic due to the unpredictable nature of the communications medium and the lack of knowledge about the load and potential delays at remote sites. Traditional , static, query processing approaches break down in this environment because they are unable to adapt in response to unexpected delays.

Bernice M. Purcell et al.[3] have focused on three major reasons for small to medium sized businesses to use cloud computing for big data technology implementation are hardware cost reduction, processing cost reduction, and ability to test the value of big data. The major concerns regarding cloud computing are security and loss of control.

Ku Ruhana et al.[4] have discussed about big data has the power to dramatically change the way institutes and organizations use their data. Transforming the massive amounts of data into knowledge will leverage the organizations performance to the maximum. Scientific and business organizations would benefit from utilizing big data. However, there are many challenges in dealing with big data such as storage, transfer, management and manipulation of big data. Many techniques are required to explore the hidden pattern inside the big data which have limitations in terms of hardware and software implementation. This paper presents a framework for big data clustering which utilizes grid technology and ant-based algorithm.

K.Sriprasadh et al.[5] have focused on retrieving the data and processing the query over cloud server by using Ant Hill optimization technique . An ant optimization technique for data retrieval is devised and tested on a well known suite of problems from the literature. It is shown that the ant colony method performs with little variability over problem instance. It is competitive with the best known heuristics for data retrieval.

Reena Jindal et al.[6] have focused on an application aiming to cluster a dataset with ACO-based optimization algorithm and to increase the working performance of colony optimization algorithm used for solving data-clustering problem, proposed two new techniques and shows the increase on the performance with the addition of these techniques. They bring out a new clustering initialization algorithm which is scale-invariant to the scale factor. Instead of using the scale factor while the cluster initialization, in this research we determine the number and position of clusters according to the changes of cluster density with the division an agglomeration processes. Experimental results indicate that the proposed DBSCALE has a lower execution time cost than DBSCAN, and IDBSCAN clustering algorithms. IDBSCALE ACO has a maximum deviation in clustering correctness rate of 95.0% and an error rate of deviation in noise data clustering of 2.62%.This algorithm is proposed to solve combinatorial optimization problem by using Ant Colony algorithm.

Divyakant Agrawal et al.[7] have discussed about Scalable database management systems (DBMS)— both for update intensive application workloads as well as decision support systems for descriptive and deep analytics—are a critical part of the cloud infrastructure and play an important role in ensuring the smooth transition of applications from the traditional enterprise infrastructures to next generation cloud infrastructures.

Badrish Chandramouli et al.[8] have proposed a new progressive analytics system based on a progress model called Prism that (1) allows users to communicate progressive samples to the system; (2) allows efficient and deterministic query processing over samples; and (3) provides repeatable semantics and provenance to data scientists.

Ms. Preeti Tiwari et al.[9] have discussed with the advancement of Computer Networks and increase in size of databases, the decentralization of databases has led to the development of Distributed Database over multiple machines where distribution of the database is Transparent to the users.

Khairul Munadi et al.[10] proposed a conceptual image trading framework that enables secure storage and retrieval over Internet services. The process involves three parties: an image publisher, a server provider, and an image buyer. The aim is to facilitate secure storage and retrieval of original images for commercial transactions, while preventing untrusted server providers and unauthorized users from gaining access to true contents.

Ranjan Kumar et al.[11] have focused on Cloud Computing as one of the paradigm in the field of IT. This technology uses the internet and central remote servers to maintain data and applications. There are heterogeneous environment, so, the utilization of resources are accessed and analyzed in real time manner.

Dr.D.Maruthanayagam et al.[12] have discussed about cloud computing as a technology which uses internet and one remote server to maintain data and various applications. Cloud Computing is one of the paradigm in the field of IT. This technology uses the internet and central remote servers to maintain data and applications.

Ratan Mishra et al.[13] have discussed about load balancing in cloud computing. The load can be CPU load, memory capacity, delay or network load.

Venkata Narasimha Inukollu et al.[14] have discussed security issues for cloud computing, Big data, Map Reduce and Hadoop environment. The main focus is on security issues in cloud computing that are associated with big data. Big data applications are a great benefit to organizations, business, companies and many large scale and small scale industries .They also discuss various possible solutions for the issues in cloud computing security and Hadoop .

Utkarsh Jaiswal et al.[15] have discussed about ant colony optimization as a new natural computation method from mimic the behaviors of ant colony. It is a very good combination optimization method. Ant colony optimization algorithm was recently proposed algorithm, it has strong robustness as well as good distributed calculative mechanism, and it is easy to combine with other methods, and the well performance has been shown on resolving the complex optimization problem.

Soorya M et al.[16] have focused on big data. The huge amount of data that cannot be processed using traditional methods are called big data. Useful information obtained after analyzing big data are used for various purposes such as predictive analytics, business applications etc.

Onkar S. Undale et al.[17] have discussed about cloud computing , due to rapid growth in information technology and mobile device technology. It is more important task, user's data privacy preservation in the cloud environment.

Henning Fernau et al.[18] have focused on convergence of the emerging trends, managing, querying and processing big data in Cloud environments, which have received a great deal of attention from the research community. Recently these play as leading role, and algorithmic implemented approaches to these challenges.

Deepak Puthal et al.[19] have aimed to precise the current open challenges and issues of Cloud computing. They have discussed the paper in three-fold: first we discuss the cloud computing architecture and the numerous services it offered. Secondly we highlight several security issues in cloud computing based on its service layer.

III. PROBLEM FORMULATION

To evaluate the query execution plans while associated with query scrambling. Scrambling is a process that modifies execution plans on the fly in response to unexpected delays in obtaining initial requested tuples from remote sources. Query scrambling consists of two different phases: a rescheduling phase, in which the scheduling of the operators of an active query plan is changed when a delay is detected, and an operator synthesis phase in which the query plan is restructured, typically by creating new operators that are not in the current query plan.

In case large data condition across wide area network, it may be associated with number of database servers known or partially known and each server may be associated with good number of databases or relations. A simulation technique, ant colony optimization (ACO) may be applied for this particular problem.

An ant optimization technique for data retrieval is devised and tested on a well known suite of problems from the literature. It is shown that the ant colony method performs with little variability over problem instance. It is competitive with the best known heuristics for data retrieval. This algorithm is evaluated using the simulated execution times for a cloud environment. Cloud Computing is dynamic in nature. So, prediction based analysis is not possible and performance monitoring of any application in cloud computing is very much important.

IV. APPLICATION OF ANT COLONYOPTIMIZATION IN PERFORMANCE EVALUATION OF QUERIES

The ant colony optimization algorithm (ACO) is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. This algorithm is a member of the ant colony algorithms family, in swarm intelligence methods, and it constitutes some metaheuristic optimizations. The first algorithm was aiming to search for an optimal path in a graph, based on the behaviour of ants seeking a path between their colony and a source of food. The original idea has since diversified to solve a wider class of numerical problems, and as a result, several problems have emerged, drawing on various aspects of the behaviour of ants.

Query Optimization in Distributed database is a difficult combinatorial optimization problem with complicated objective functions therefore powerful search algorithms are needed for it. ACO is a meta-heuristic, multi agent approach that simulates the foraging behavior of ants for solving difficult NP-hard combinatorial optimization problems. Ants are social insects whose behavior is directed more towards the survival of colony as a whole than that of a single individual of the colony . An important and interesting behavior of an ant colony is its indirect co-operative foraging process. Ant Colony Optimization takes inspiration from the foraging behavior of some ant species. These ants deposit pheromone on the ground in order to mark some favorable path that should be followed by other members of the colony. While walking from the food sources to the nest and vice versa, ants deposit a substance, called pheromone trail. Ants can smell pheromone. When choosing their way they tend to choose, with high probability, paths marked by strong pheromone concentration (shorter path) with the result that after some time the whole colony converges toward the use of the path.

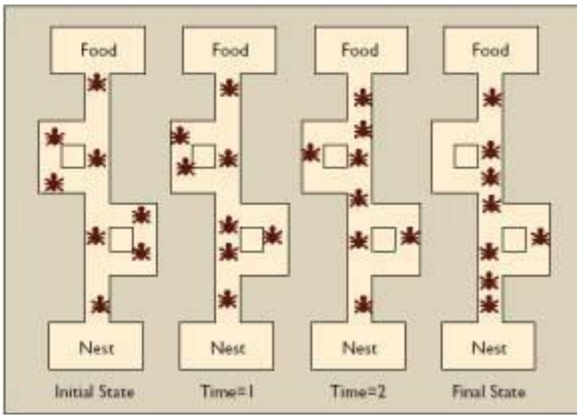
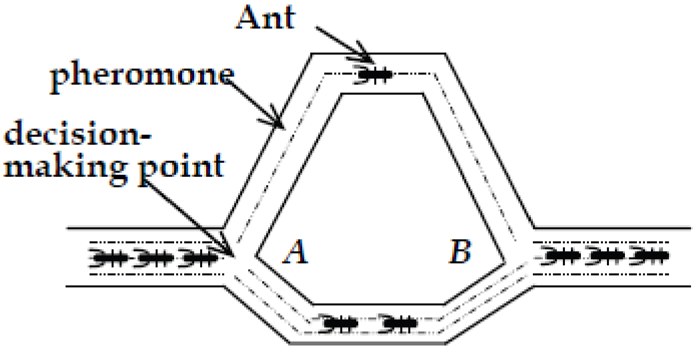


Figure shown above presents a decision-making process of ants choosing their trips. When ants meet at their decision-making point *A*, some choose one side and some choose other side randomly. Suppose these ants are crawling at the same speed, those choosing short side arrive at decision-making point *B* more quickly than those choosing long side. As a result, the quantity of pheromone is left with higher speed in short side than long side because more ants choose short side than long side. The number of broken line is in direct ratio to the number of ant approximately. Artificial ant colony system is made from the principle of ant colony system for solving kinds of optimization problems. Pheromone is the key of the decision-making of ants.



In ACO, an artificial ant builds a solution by traversing the fully connected construction graph GC (**V**, **E**), where **V** is a set of vertices and **E** is a set of edges. The artificial ants move from vertex to vertex along the edges of the graph building a partial solution. The first ant colony optimization algorithm is known as Ant System and was proposed in the early nineties. Since then, several other ACO algorithms have been proposed. Given below is an algorithm of ACO Metaheuristics that iterates over three phases.

a) ConstructAntSolution: A set of m artificial ants constructs solutions from elements of a finite set of available solution components

b) ApplyLocalSearch: Once solutions have been constructed, and before updating the pheromone, this function improves the solutions obtained by the ants through a local search.

c) UpdatePheromones: It increases the pheromone values associated with good or promising solutions, and to decrease those that are associated with bad ones. This is achieved (i) by decreasing all the pheromone values through pheromone evaporation, and (ii) by increasing the pheromone levels associated with a chosen set of good solutions.

ACO is designed and developed specifically to tackle continuous problems of Combinatorial Optimization. With the increasing number of relations in a query, much use of memory and processing is needed. DDBMS is now being used as a standard DBMS in all commercial applications which involve data from various sites. The path marking the behavior of ants is applied to direct the ants towards the unexplored areas of search space and visit all the nodes without knowing the graphic topology for generation of optimal solutions of distributed database queries. These ants calculate the running times of the execution plans of the given query and provide quick, high performance and optimal results in a cost effective manner.

V. QUERY EVALUATION TECHNIQUES

When considering the evaluation of query operators in a distributed setting we need to consider not only the I/O cost on local disk but also the network costs, which have a significant contribution to the overall query execution cost. Any cost model used in a distributed setting must also take into account the site that issues the query Sq so that we include the cost of shipping the result from the site at which it is materialized to Sq. In the following discussion we consider the case where relations are stored completely at sites and are not fragmented, as this is the approach taken in this thesis. In this case performing selections and projections is straight forward as we simply use centralized implementations of these algorithm on a relation at a particular site and then ship the results to Sq if necessary. When considering the join between two relations R1 and R2 in a distributed setting there are a number of situations that must be taken into consideration when determining the most efficient means of carrying out the join. Situations involving possible shipping of complete relations between sites are discussed below. Techniques that aim to reduce the network useage when performing cross site joins include the use of Semi-joins and Bloomjoins. If both relations are present at the same site the join can be performed entirely at the local site. As communication costs are more expensive than disk costs in general it is likely that this approach will result in a plan with the least cost. If both relations R1 and R2 reside at different sites, S1 and S2 respectively, the join can be performed at either site. If one of these sites S1 happens to be the site at which the query was issued, henceforth known as the query site, it is likely that the optimal query plan will involve shipping R2 from S2 to S1 and performing the join at S1. However, consider the situation where R2 is very large, R1 is very small and the result of joining R1 and R2 is very small. In this case the optimal plan may involve shipping R1 to S2 and performing the join at S2. The small result would then be returned to the query site.

VI. ALGORITHM

```
Step 1: Select data set, Ds
 Step 2: Evaluate average request time, Tavg
Step 3: Calculate application based instance, q
Step 4: Calculate expected arrival rate, y
Step 5: Evaluate current number of application instances,m
Step 6: while (m!=q) ynew=y/m
Step 7: if (m<q) then mnew=m
Step 8: else if(m==q) then m=mnew
Step 9: else m=m + m/2, mnew=m+1
```

VII. EXPERIMENTAL ANALYSIS

It is observed that the design usually contains cloud infrastructure created using Infrastructure as a service cloud based software that is responsible to build cloud environments. It may support multiple servers and may have the ability to build cloud environments with different servers with proper interfacing. In this case minimum two machines are usually required to implement private cloud. One machine may be used as management Server which may run on a dedicated server or a virtual machine. It controls allocation of virtual machines to hosts and assigns storage and IP addresses to the virtual machine instances.

VIII. CONCLUSION AND FUTURE WORK

Today big data is in boom and handling such a large volume and variety of data is a big challenge for us. Big data are implemented on this platform and uses its tools, softwares and hardwares for the manipulation of its data. In this paper the secure data retrieval from cloud database with the help of scrambling , without the loss of data is been elaborated .Without srambling, cloud computing is possible but it will be inefficient and inflexible. Scrambling provides flexibility scalability, and cost advantages to cloud computing. There are many levels and many ways to implement scrambling. Further in our research we are going to implement ant colony optimization technique on big data for faster manipulations of data.

References

[1] Laurent Amsaleg , Michael J. Franklin , Anthony Tomasic , Tolga Urhan , “ Scrambling Query Plans to Cope With Unexpected Delays “ .

[2] Tolga Urhan , Michael J. Franlin , Laurent Amsaleg , “ Cost-based Query Scrambling for Intial Delays “.

[3] Bernice M. Purcell , “ Big data using Cloud computing “ , Journal of technology Research .

[4] Ku Ruhana , Ku Mahamud , “ Big Data Clustering Using Grid Computing And Ant Based Algorithm “, ICOCI 2013 , 28-30 August 2013 , Sarawak , Malaysia .

[5] K.Sriprasadh , M.Prakash Kumar , “Ant Colony Optimization Technique for Secure Various Data Retrieval in Cloud Computing “,International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7528-7531.

[6] Reena Jindal, Samidha D.Sharma, Manoj Sharma , “ A New Technique to Increase the Working Performance of the Ant Colony Optimization Algorithm “ , International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-3, Issue-2, July 2013 .

[7] Divyakant Agrawal , Sudipto Das , Amr El Abbadi , ” Big Data and Cloud Computing: Current State and Future Opportunities “ .

[8] Badrish Chandramouli , Jonathan Goldstein , Abdul Quamar , “ Scalable Progressive Analytics on Big Data in the Cloud “ .

[9] Ms. Preeti Tiwari , Dr. Swati V. Chande , “ Optimization of Distributed Database Queries Using Hybrids of Ant Colony Optimization Algorithm “ , International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 6, June 2013 .

[10] Khairul Munadi , Fitri Arnia , Mohd Syaryadhi , Masaaki Fujiyoshi and Hitoshi Kiya , “ A secure online image trading system for untrusted cloud environments “ , SpringerPlus (2015) 4:277 DOI 10.1186/s40064-015-1052-1.

[11] Ranjan Kumar , G. Sahoo , K. Mukherjee , “ Performance Analysis of Cloud Computing using Ant Colony Optimization Approach “ , International Journal of Innovative Research in Science, Engineering and Technology Vol. 2, Issue 6, June 2013 .

[12] Dr.D.Maruthanayagam , T. Arun Prakasam , “ Job Scheduling in Cloud Computing using Ant Colony Optimization “ , International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, Issue 2, February 2014 .

[13] Ratan Mishra , Anant Jaiswal , “Ant colony Optimization: A Solution of Load balancing in Cloud “ , International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.2, April 2012 .

[14] Venkata Narasimha Inukollu , Sailaja Arsi , Srinivasa Rao Ravuri , “ Security Issues Associated With Big Data In Cloud Computing “ , International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014 .

[15] Utkarsh Jaiswal , Shweta Aggarwal , “Ant Colony Optimization “ , International Journal of Scientific & Engineering Research Volume 2, Issue 7, July-2011.

[16] Soorya M , Swaraj K.P , “Elliptic Curve Based Data Scrambling with Encryption for Security of Big data “,International Journal of Innovative Research in Science, Engineering and Technology , Volume 5, Special Issue 14, December 2016.

[17] Onkar S. Undale , Prof. Bharati Kale, ” Running Big Data Privacy Preservation in the Hybrid Cloud Platform “,International Journal on Recent and Innovation Trends in Computing and Communication , Volume: 4 Issue: 7.

[18] Alfredo Cuzzocrea , “Algorithms for Managing, Querying and Processing Big Data in Cloud Environments “,12 January 2016; Accepted: 27 January 2016; Published: 1 February 2016.

[19] Deepak Puthal , B. P. S. Sahoo , Sambit Mishra , Satyabrata Swain , “Cloud Computing Features, Issues and Challenges: A Big Picture “,2015 International Conference on Computational Intelligence & Networks (CINE 2015).