Journal of Theoretical and Applied Information Technology
30th April 2017. Vol.95. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TEACHER-LEARNER & MULTI-OBJECTIVE GENETIC ALGORITHM BASED QUERY OPTIMIZATION APPROACH FOR HETEROGENEOUS DISTRIBUTED DATABASE SYSTEMS

S.VENKATA LAKSHMI [1*], VALLI KUMARI VATSAVAYI [2]
[1]Assistant Professor, Department of Information Technology, GITAM University, Visakhapatnam - 530045, Andhra Pradesh, India
[2]Professor, Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam - 530003, Andhra Pradesh, India
svlakshmi2014@gmail.com[1], vallikumari@gmail.com[2]

ABSTRACT

Growing database demands more technological developments and computing paradigms, as Grid and Cloud computing, unleashed new developments in the database technology sector. Query Optimization is essentially a complex search task to obtain the best possible plan from the enormously increasing databases. Heterogeneous Distributed database management systems (DDBMS) are amongst the most important and successful software developments where the query processing is more difficult since large number of parameters effect the performance of the queries. Thus, the author attempted to introduce a new approach for Query Optimization in Heterogeneous DDBMS both for local and global optimization separately. In this paper, two stochastic approaches such as multi-objective genetic algorithm for local optimization and teacher-learner based optimization for global optimization is employed. The local optimization approach deals with optimization within the local sites whereas global optimization works with the sites at different locations globally. Join ordering cost (JOC), Total Local Processing Cost (TLPC) and Total Communication Cost (TCC) are used to obtain the optimal query plans amongst the relation between the query sites. The Experimental Analysis of the proposed approach showed that it has better performance i.e. less cost when compared with other heterogeneous DDBMS and has more cost when compared with the other existing homogeneous approaches.

Keywords: *Query Optimization, Heterogeneous Distributed Database Systems, Multi Objective Genetic Algorithm, Teacher-Learning based Optimization*

1. INTRODUCTION

Growing database demands, new technological developments and new computing paradigms, as Grid and Cloud computing, unleashed new developments in the database technology sector. Meanwhile a steadily increasing proliferation of more and more inexpensive resources, such as processor, memory and hard disk, led to significant developments of sophisticated parallel database systems in the last decades [1]. These databases operate on row-oriented or column-oriented schemes. Representatives of row-oriented databases are Oracle DB, MySQL and PostgreSQL. Usually they provide a good performance for a generic set of use cases, but do not specifically focus on high performance. In contrast, column-oriented databases such as Vertica, Mariposa [2] or VoltDB establish mechanisms to provide high performance, but are mostly bound to a very particular execution environment. In generally parallelism in both types is achieved by facilitating intra- and inter-operator parallelism [3] and making use of computational resources (CPUs) to support a multitude of incoming query requests in parallel.

Query optimization [4] is essentially a complex search task for the best possible plan among the semantically equivalent plans that can be generated for any given query. Various Optimizations approaches have been applied by researchers to find an optimal plan for query execution. With queries getting more and more complex the search complexity is increasing. Heterogeneous Distributed database management systems (DDBMS) are amongst the most important and successful software developments in this decade. Query Processing is much more

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

difficult in distributed environment than in centralized environment because a large number of parameters affect the performance of distributed queries like data translation, relations may be fragmented and replicated and considering many sites to access, query response time may be high. In Heterogeneous DDBMS represented in Figure1, Communication cost is dominant factor. Various algorithms for query optimization were already devised which attempts to reduce the quantity of data transferred.

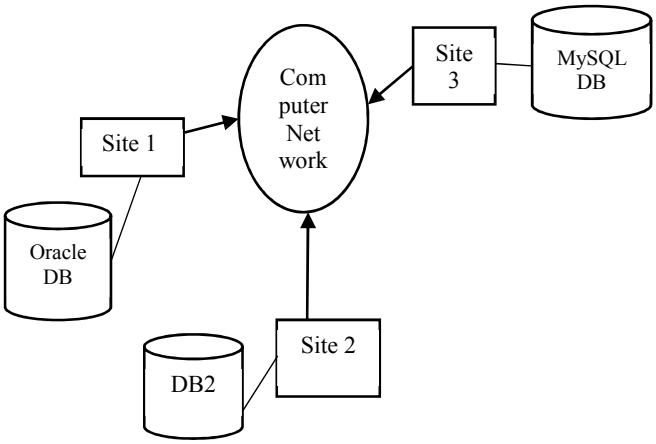

Figure 1: Heterogeneous Distributed Database System

The distributed query optimization has several problems related to cost model, large set of queries, optimization cost and optimization interval. In case of Query Optimization in Heterogeneous Distributed databases [5], there are different types of database management system present on different sites as some sites may have Oracle, some sites may have db2 and some other may have Sybase .Therefore there is a problem of integrating all the databases into one common interface and doing querying optimization of heterogeneous data sources. Owing to this issue, numerous Stochastic, Evolutionary and Combinatorial optimization techniques have been introduced recently. However, Exhaustive techniques are adequate for trivial instances only, while combinatorial optimization techniques are vulnerable to the peculiarities of specific instances.

Genetic Algorithm is one such optimization technique that are widely used and accepted for very difficult optimization problems. Even though, the application of Genetic Algorithm to query optimization is motivated by its robustness and efficiency, this approach could not completely handle the issues while working with Heterogeneous Distributed Database for Query Optimization. Thus, the author attempt to introduce a new approach for Query Optimization in Heterogeneous DDBMS, for both local and global optimization separately. In this paper, two stochastic approaches such as multi-objective genetic algorithm for local optimization and teacher-learner based optimization for global optimization is employed. The local optimization approach deals with optimization within the local sites whereas global optimization works with the sites at different locations globally. The join ordering cost (JOC), Total Local Processing Cost (TLPC) and Total Communication Cost (TCC) are used to obtain the best optimal query plans in the proposed approach amongst the relation between the query sites.

1.1 Organization of the Paper

An introduction for the query optimization, heterogeneous distributed database systems and motivation for the proposed Hybrid approach is given in this section. Brief description of existing query optimization approaches in distributed database systems and the different hybrid approaches for the distributed systems are given in section 2. The detailed explanation of the proposed Hybrid Approach of Teacher-Learner Based Optimization and Multi Objective Genetic Algorithm for Heterogeneous DDBMS is given in section 3. The experimental results and its detailed analysis is discussed in section 4 followed by conclusions and references given in section 5 and section 6 respectively.

2. LITERATURE SURVEY

An evolutionary query optimization mechanism in distributed heterogeneous systems has been suggested in [6] using multi-agent architecture and genetic algorithm approach to meet demand for high dimensional queries that has the big disadvantage of its exponential order and also compare the result with some commonly used query optimization algorithms. A comprehensive state of the art concerning the evolution of query optimization methods [7] from centralized database systems to data Grid systems through parallel, distributed and data integration systems is given. For each environment, synthetically some methods are described that pointed out its main characteristics.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A novel heuristic framework for the optimization [8] of query execution plans (QEP) on a world-wide scale is given. This work is based on a multi-staged blackboard mechanism to determine which available data, resources and operations have to be considered to perform a query optimally. The research effort addressed the query optimization issue of distributed database queries and a variety of different heterogeneous and homogeneous infrastructures, parallel algorithms, and huge datasets are considered which span across several virtual organizations (VOs) with usually no centralized authority.

An approach for multi-objective parametric query optimization [9] (MPQO) is specified for advanced database systems such as distributed database systems (DDBS). MPQO builds a parametric space of query plans and progressively explores the multi-objective space according to user tradeoffs on query metrics. In heterogeneous and distributed database system, logically unified data is replicated and distributed across multiple distributed sites to achieve high reliable and available data system; this imposed a challenge on evaluation of Pareto set. An MPQO attempt exhaustively determines the optimal query plans on each end of parametric space.

K-QTPT [10] is introduced to reduce the high optimization cost incurred by QTPT (Query Trading with Processing Task Trading). In k-QTPT, only k nodes participate in generating optimal plans. The implementation details of Query Trading, QTPT algorithm and the k-QTPT solutions are discussed. K-QTPT is evaluated through emulation and shown that the cost of optimization reduces substantially in k-QTPT as compared to QT and QTPT. An adaptive cost-based query optimization [11] to meet the requirements while taking network topology into consideration.

Parallel Genetic Algorithm-Max-Min Ant System [12] was proposed to seek a best query execution plan, which combines faster convergence of Genetic Algorithm, globally search ability of Max-Min Ant System and parallel property for both of them. The experiment results show that the proposed algorithm is effective for query processing of multi-join, and plays important role in improving the performance of distributed database. Quadratic assignment problem [13] heuristics were designed and implemented for the data allocation problem. The proposed algorithms find near-optimal solutions for the data allocation problem. In addition to the fast ant colony, robust tabu search, and genetic algorithm solutions to this problem, we propose a fast and scalable hybrid genetic multi-start tabu search algorithm that outperforms the other well-known heuristics in terms of execution time and solution quality.

An optimization problem [14] is defined that combined the iterative join and the graph exploration methods to minimize the evaluation time of distance join queries. Without sacrificing a system's scalability, our technique exploits a light-weight vertex centric encoding schema built on a distance-aware partition of the entire graph. Extensive experiments over both real and synthetic large graphs show that, by employing an adaptive query plan generation and scheduling method, we can effectively reduce the redundant message passing and I/O costs. Compared to simply using iterative join or graph exploration method, our solution achieves as many as one order of magnitude of time saving for the query evaluation.

3. PROPOSED HYBRID QUERY OPTIMIZATION APPROACH USING TEACHER-LEARNING BASED OPTIMIZATION AND GENETIC ALGORITHM

In this section, a hybrid optimized approach is suggested for the Distributed Heterogeneous Database Systems. The proposed approach is hybridized by using two different algorithms such as Parametric and Parametric-less techniques. The query optimizations in large heterogeneous distributed systems are accomplished by means of four stages. They are: Query decomposition, Data localization, Global optimization, Local optimization. However, in this paper, the main concentration is on last two stages that is Global Optimization and Local Optimization. The Local Optimization of the queries within the query sites are performed by means of parametric technique known as Genetic Algorithm whereas the Global Optimization between different query sites are accomplished by means of Parametric-less Technique known as Teacher-Learning Based Optimization Approach. The block Diagram for the proposed approach is given in Figure 2 and Flow Chart for the Query Optimization is given in Figure 3.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

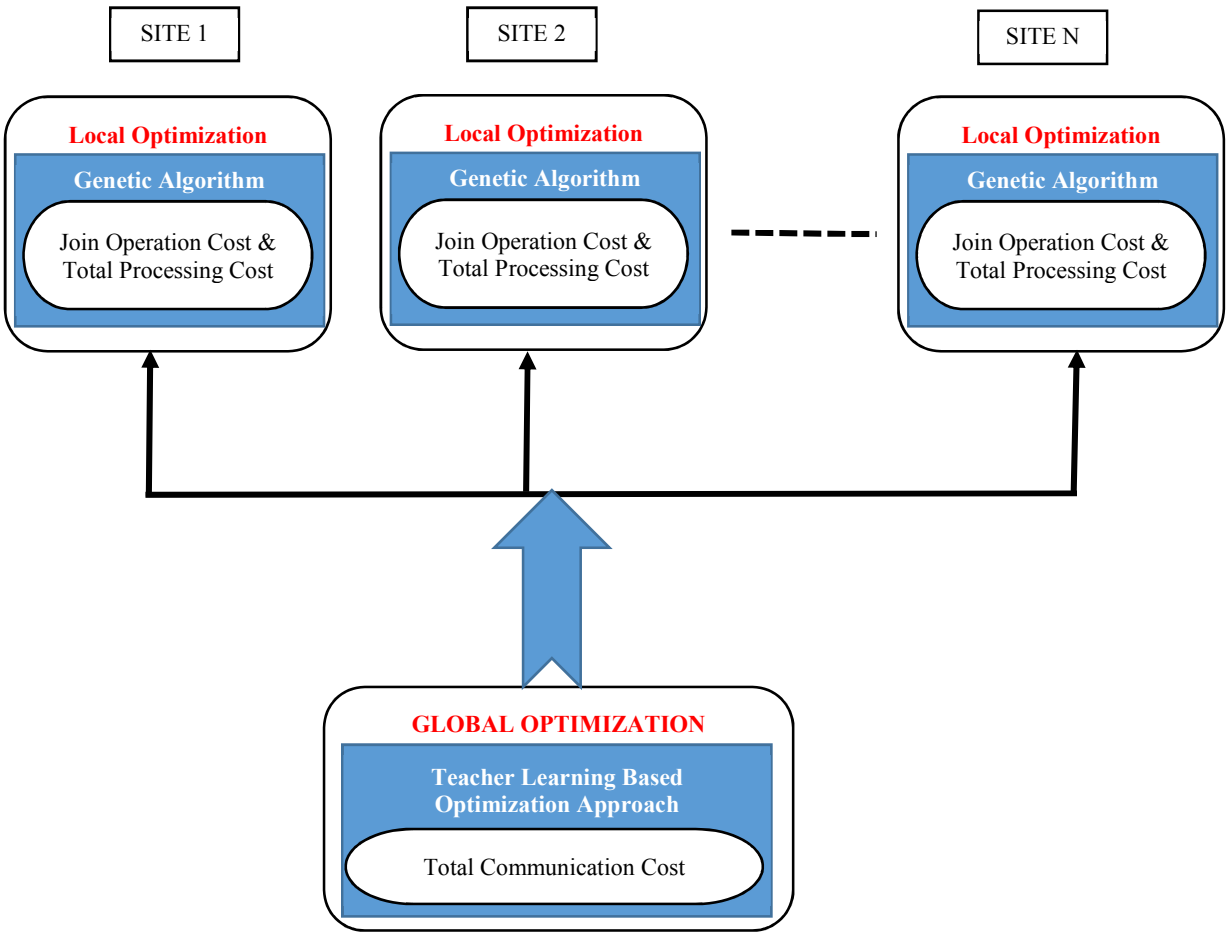

Figure 2: Block Diagram For The Proposed Query Optimization Approach

3.1 Query Cost Model for Heterogeneous Distributed Systems

The Query Evaluation Plans in Distributed Database are selected pertaining to the size of query relations along with the size of the intermediate results of two join relations also that are employed as cost estimation., For a join $J = R_1 \, join \, R_2$ in distributed database, the implementation cost is necessary to consider its computational cost in two different cases. They are:

- When two join relations are in the same site, the join operation cost is calculated within the site
- When they are not in the same site, Communication cost needs to be considered additionally.

1. Join operation cost in equation (1) is given as:

$$Cost(J) = \frac{|R_1| \times |R_2|}{\prod_{c_i \in C} \max(V(c_i, R_1), V(c_i, R_2))} \quad (1)$$

$|R|$ is the size of relation R, C stands for public attribute set of relations R_1 and R_2, $V(c, R)$ is numbers of the distinct values of attribute c in R. In a join that contains n relations, (J_1, J_2, J_i) are intermediate nodes of join tree, the total cost estimation model is given as:

$$COST = \sum_{i=1}^{n-1} Cost(J_i) \quad (2)$$

Total Local Processing Cost is the cost between the participating sites.
This is given as:

$$TLPC = \sum_{i=1}^{u} LPC_i \times a_i \quad (3)$$

where u is the number of locations accessed through the query plan in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ascending order of cardinality per sites, LPC_i is the local processing cost per byte at location i, a_i is the bytes to be processed at site i. If a site contains a single relation, its LPC is considered to be zero.

2. Total Communication Cost is given as:

$$TCC = \sum_{i=1,j=i+1}^{i=u-1,j=u} CC_{ij} \times b_i \quad (4)$$

Where$u$ is the number of locations accessed through the query plan in ascending order of cardinality per sites, CC_{ij} is the communication cost per bytes amongst locations i and j, b_i is the bytes to be broadcasted from location i. If a site contains a single relation, its LPC is considered to be zero. Here,

$$Card_i = \frac{Card(R_t) \times Card(R_s)}{Dist_{ts} \times \min(Card(R_t), Card(R_s))} \quad (5)$$

$Dist_{ts}$ is the number of distinct tuples in the minimum relation amongst R_t and R_s. The dimension of the resultant relation R_i at site i is given as:

$$Size_i = Size(R_t) + Size(R_s) \text{ and}$$
$$a_i = Card_i \times Size_i ,$$
$$b_j = Card_j \times Size_j \quad (6)$$

3.2 Local Optimization using Multi Objective Genetic Algorithm

The Local Optimization within the Database Site is performed by means of Multi Objective Evolutionary Algorithm known as Genetic Algorithms. In this section, two objectives are considered for the optimization. The detailed description of the objectives that are to be employed are given in section 3.1. They are Total Local Processing Cost (TLPC) and Join Operation Cost (JOC). The individuals Chromosomes of the Multi Objective Genetic Algorithm are arranged in n different front depending on the objective function. The first front being completely non-dominant set in the current population and the second front being dominated by the individuals in the first front only and the front goes so on. Each individual in the each front are assigned rank (fitness) values or based on front in which they belong to. Individuals in first front are given a fitness value of 1 and individuals in second are assigned fitness value as 2 and so on. In addition to fitness value a new parameter called crowding distance is calculated for each individual. The crowding distance is a measure of how close an individual is to its neighbors. Large average crowding distance will result in better diversity in the population.

Multi-Objective Genetic Algorithm based Local Optimization Algorithm:

1. **Population Initialization:** Each Relation $R_1, R_2, \ldots \ldots R_n$ of every individual site are initial individuals or chromosomes for this approach. The Population is initialized based on the problem range and constraints.

2. **Non Dominated Sort:** In order to perform a non-dominated ranking, each query plan is matched with every other query plan in the population to detect if it is dominated. The fast non dominated sorting procedure takes the current population as input and produces a list of non-dominated ranks as output. For every query plan "i", the following entities need to considered:

R_i: The number of Relations that dominate the Relation i
S_i: The set of Relations that the Relation I dominates.

- Compute the TPLC and JOC for each and every Relation in the Site.
- Now check for the condition
  - $initialize\ set_i = \{\emptyset\}\ and\ R_i = 0\ and\ rank = 0$
  - $for\ i = 1{:}\,pop\_size$
    $for\ j = 2{:}\,pop\_size$
    $if\ JOC_i < JOC_j \&\& TLPC_i < TLPC_j$
    $Set_i = Set_i \cup \{j\}$
    $R_i = R_i + 1$
  - $for\ j = 1{:}\,pop\_size$
    $for\ all\ R_i = 0, rank = 0,$
    $then\ reduce\ all\ n_j\ by\ 1\ and\ prioritize\ the\ rank\ to\ the\ next\ increased\ number\ when\ ever\ n_j = 0\ or\ length\ \{Set_i\} = 1\ untill\ all\ the\ values\ are\ ranked\ with\ the\ numbers$
- Finally the ranked query plans are stored.

3. **Crowding Distance:** The Crowding Distance is assigned after the non-dominated sort is performed. Since the individuals are selected based on rank and crowding distance all the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

chromosomes in the population are assigned a crowding distance value and is calculated in the same front. The crowding distance is evaluated as:

- For each front F_i, n is the number of Relations.
  - Initialize the distance to be zero for all the individuals i.e. $F_i(d_i) = 0$,where j corresponds to the j^{th} individual in front F_j.
  - For the objective function JOC and TLPC
    a. Sort the individuals in front F_i based on two objectives one by one i.e. $I = sort(F_i, JOC) \& I = sort(F_i, TLPC)$
    b. Assign infinite distance to boundary values for each individual in F_i i.e. $I(d_1 = \propto) \; and \; . I(d_n = \propto)$
    c. For k = 2 to (n-1)

$$I(d_k) = I(d_k) + \frac{I(K+1)_m - I(K-1)_m}{{f_m}^{max} - {f_m}^{min}}$$

Here, $I(K)_m$ is the value of the m^{th} objective function of k^{th} query plan in the rank and ${f_m}^{max}$ and ${f_m}^{min}$ are the maximal and minimal values obtained for the objective function m.

The basic idea behind the crowding distance is finding the Euclidian distance between each individual in a front based on their m objectives in the m dimensional hyper space.

4. **Selection Operation:** The Selection Operation is carried out using Crowed Comparator Operator $\prec_n$ i.e. based on the non-domination rank p_{rank}individuals in front F_i will have their rank as $p_{rank} = i$. The individuals are selected by using a binary tournament selection with crowed comparison-operator.
   For $p \prec_n q$, if $p_{rank} < q_{rank}$ or if p and q belong to the same front F_i then $F_i(d_p) > F_i(d_q)$ i.e. the crowding distance should be more.
5. **Genetic Operations:** The two genetic operations that employed in the algorithm are crossover and mutation. The single point binary crossover is performed on the selected random relation and an arithmetic mutation is performed on the obtained parents from crossover.
6. **Termination Criteria:** The offspring relation is combined with the current generation population and selection is performed to set the individuals of the next generation. All the previous and current best individuals are added in the population that ensured elitism. The individuals are now sorted based on non-domination. The new generation is filled by each front subsequently until the population size exceeds the current population size. If by adding all the individuals in front F_j, the population exceeds N then individuals in front F_j are selected based on their crowding distance in the descending order until the population size is N. The process repeats to generate the subsequent generations.

3.3 Global Optimization using Elite Teacher-Learning Based Optimization

The Global Optimization between different Database Site is performed by means of Parametric less Stochastic Algorithm known as Teacher-Learning based Optimization Algorithms [15]. In this section, only single objective is considered for the optimization known as Total Communication Cost is given in detailed in section 3.1. TLBO algorithms aim to find global solutions for real world problem with less computational effort and high reliability. The principle idea behind TLBO is the simulation of teaching–learning process of a traditional classroom in to algorithmic representation with two phases called teaching and learning. Elitist TLBO was pioneered with a major modification to eliminate the duplicate solutions in learning phase. In this optimization algorithm, the fraction of learners are assumed as population and diverged configuration of variables are treated as distinctive subjects accessible to the learners, and their result is comparable to the fitness estimation value of this optimization issue. In the whole population, the best solution is treated as the teacher.

Elite Teacher-Learning Based Optimization Algorithm:

1. **Initialization Phase:** Initialize the Relations of different sites as Population which are also known as Learners, number of sites as design

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

variables and the termination criteria as number of generations.

2. **Elite Teaching Phase:** The best join relation is selected as teacher from every site in heterogeneous distributed database for and the mean result of learner in each site is evaluated.

   i. Initially, the optimal relations obtained from local optimization using Multi Objective Genetic Algorithm is kept as the elite solution
   ii. The best solution for the teacher phase is calculated by means of the objective function i.e. Total Communication Cost between the different relations in the different sites. The obtained result is retained as best solution TCC_{best} for the current iteration in the teaching phase.

   iii. Using the Best Solution, the existing Join Relations are modified as given below as:

   $$TCC_{modified} = r(TCC_{best} - TF(TCC_{existing})) \quad (7)$$

   The random value r is in between [0, 1] and $TF = round[1 + rand(0,1), 2 - 1]$

   iv. If the modified values are better than the existing then existing values are replaced the new ones, else the existing solutions remains similar

3. **Elite Learning Phase:** In this phase, the learner's knowledge is updated by means of teacher's knowledge.
   i. The two relations such as R_1&$R_2$are selected randomly by means of modified Relations in teachers phase and TCC values are evaluated.
   ii. If $TCC_1 > TCC_2$then for R_1&R_2compute: ${TCC_1}^{new} = {TCC_1}^{old} + r({TCC_2} - {TCC_1}^{old})$
   iii. Else Compute: ${TCC_2}^{new} = {TCC_2}^{old} + r({TCC_1} - {TCC_2}^{old})$
   iv. If the modified values are better than the existing then existing values are replaced the new ones, else the existing solutions remains similar.
   v. Now replace the worst solutions in the learner phase with the elite solutions obtained in the teaching phase.

4. **Termination Criteria:** Once the single iteration of Teacher Phase and Learning Phase is completed, then check for the termination criteria with maximum number of iterations.

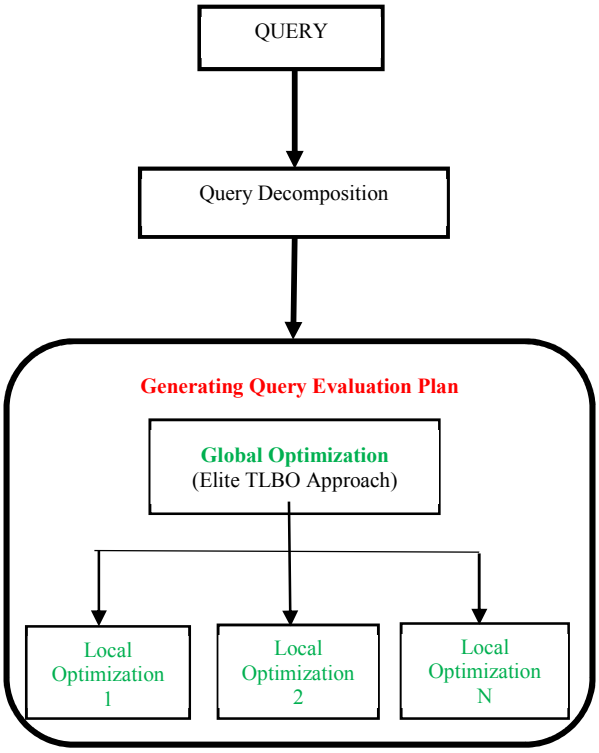

Figure 3: Flow Chart For The Proposed Hybrid Optimization Approach

4. Experimental Results and Analysis

The Experiment for the proposed hybrid Query Optimization Approach using Elite Teacher-Learner and Multi Objective Genetic Algorithm was implemented in Matlab 14a in Windows 8 professional 64 bit OS. The Query Optimization Dataset comprises of 20 different queries with 15 relations distributed over 9 different database sites in the heterogeneous environment. The heterogeneous distributed database system in this paper employs three different database storages like MySQL Database, Oracle Database and PostgreSQL with row-oriented representation. The Average Query Execution Cost for the optimal Query Execution Plan (QEP) in the proposed approach is evaluated using Join Operation Cost (JOC) and Total Local Processing Cost (TLPC) for Local Optimization and Total Communication Cost (TCC) for Global Optimization.

Journal of Theoretical and Applied Information Technology
30th April 2017. Vol.95. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

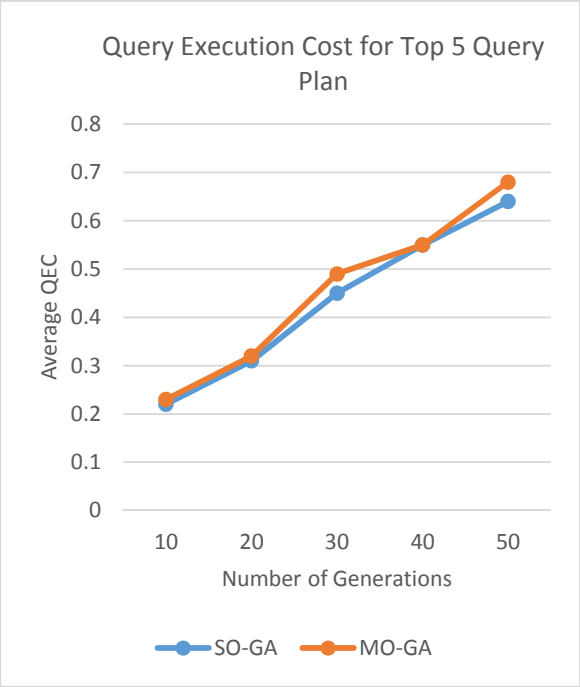

Figure 4: Average Query Execution Cost For Top 5 Query Plans

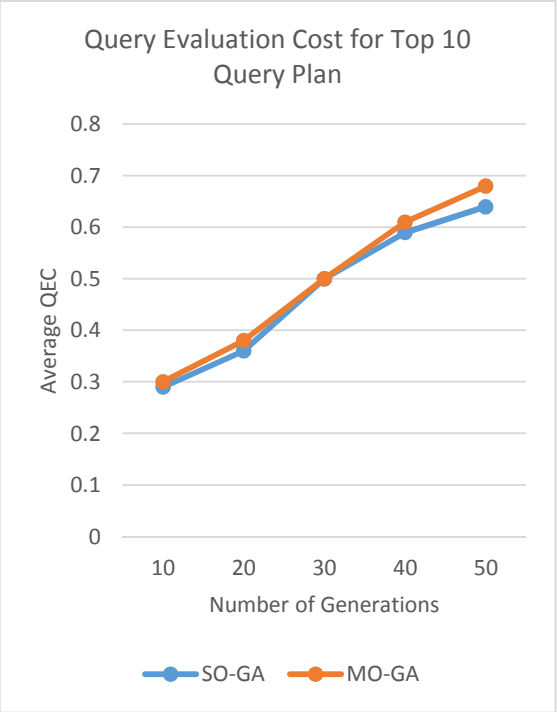

Figure 5: Average Query Execution Cost For Top 10 Query Plans

The performance analysis of the proposed approach is carried out in different ways by comparing the Average Query Execution Cost (QEC) of the proposed hybrid approach with the existing optimization Techniques against the Top-3, Top-5 and Top-10 Query Execution Plans. Figure 4 and Figure 5 represents the Average Query Cost for Top-5 and Top-10 Queries respectively for number of Generations. From the Figures, it is shown that the proposed MOGA has more and some places relatively equal Average QEC when compared with the existing SOGA. Since in MOGA, the Computation is more as it has performed for two different objectives and also employed the Non-Dominant and Crowded Distance Approach.

The local optimization carried out using Multi-Objective Genetic Algorithm is compared with the existing Single-Objective Genetic Algorithm for the Top Query Plans as shown in Table 1 and Figure 6. From this, it can be inferred that the Top-3 Query Plan has less QEC for the MOGA and for Top-5 and Top-10, the QEC values are relatively similar to the SOGA. Table 2 and Figure 7 demonstrate the Average QEC for the global Optimization technique carried out using three different approaches. They are Single-Objective TLBO, Non-Dominant TLBO and Proposed Elite based TLBO fir the Top queries. From the results, it can be inferred that Average QEC cost for the Proposed approach is comparatively less when compared to other optimization techniques.

Table 1: Average QEC In Local Optimization For Single And Multi Objective Genetic Algorithm

Top K Query Plans	SO-GA	MO-GA
Top 3	0.45	0.44
Top 5	0.55	0.56
Top 8	0.69	0.72
Top 10	0.75	0.75

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

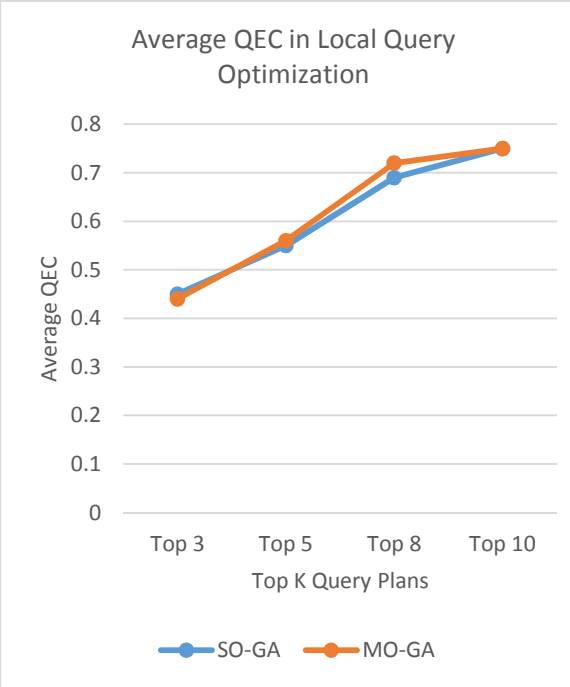

Figure 6: Average QEC in Local Optimization for Single and Multi Objective Genetic Algorithm

Table 2: Average QEC in Global Optimization for different variants of TLBO Approach

Top K Query Plans	SO-TLBO	ND-TLBO	Elite - TLBO
Top 3	0.35	0.34	0.33
Top 5	0.43	0.43	0.42
Top 8	0.57	0.56	0.56
Top 10	0.69	0.7	0.71

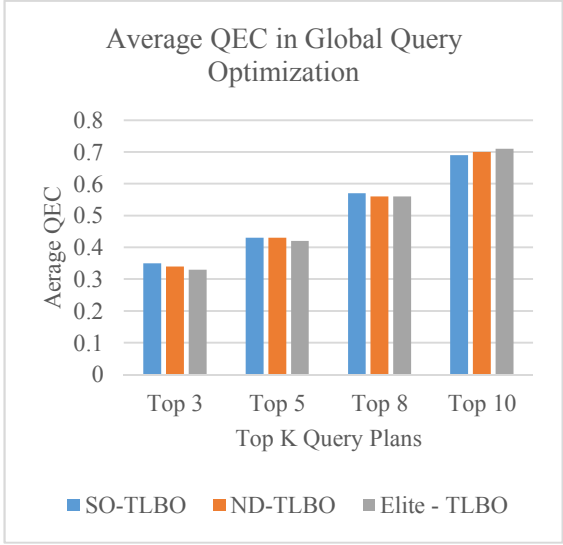

Figure 7: Average QEC in Global Optimization for different variants of TLBO Approach

The differences in the heterogeneous and homogeneous distributed database systems is analyzed and shown in Table 3 and Figure 8. In this, the average QEC of the Proposed Heterogeneous Distributed Database System is compared with existing Homogeneous Distributed Database Systems for different query plans and it can be clearly inferred that the QEC is more for heterogeneous DDBMS compared to Homogeneous since the DDBMS involves different data bases each with different execution cost.

Table 3: Average QEC Of Query Optimization For Heterogeneous And Homogeneous DDBMS

Top K Query Plans	Query Optimization Using Homogeneous DDBMS	Query Optimization Using Heterogeneous DDBMS
Top 3	0.54	0.56
Top 5	0.67	0.69
Top 8	0.72	0.74
Top 10	0.75	0.78

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

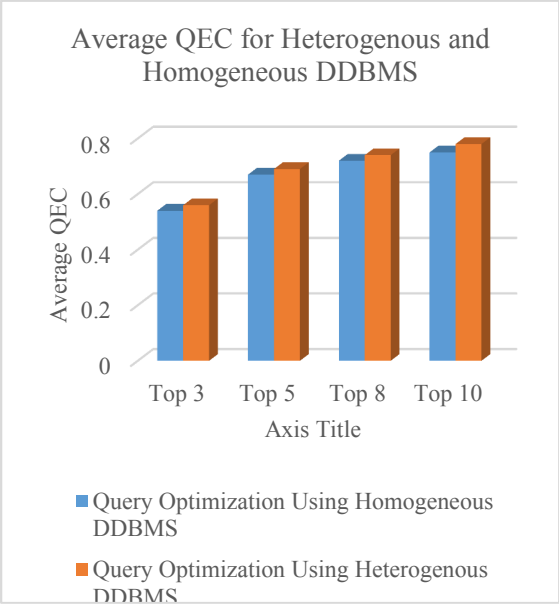

Figure 8: Average QEC of Query Optimization for Heterogeneous and Homogeneous DDBMS

In Table 4 and Figure 9, the comparison of three different approaches such as Clustering based GA Query Optimization [16], Non-Dominant TLBO Query Optimization [17], and the Proposed Hybrid Elite TLBO and GA Query Optimization proposed by author are given. The Comparison is accomplished in Top Query Plans with the Average Query Execution Cost of the queries in the Distributed Database Systems. From the fig and table, it can be inferred that Average QEC for the prior existing approaches are less when compared to the proposed approach. However, it can be observed that the performance is carried on the Heterogeneous Distributed Database System. Thus, in Fig the proposed Heterogeneous DDBMS is compared with the existing Heterogeneous DDBMS and is shown that the Query Execution Cost in the proposed approach is less when compared with the existing approach.

Table 4: Average Query Execution Comparison For Three Different Query Optimization Approaches

Top K Query Plans	Clustering Based GA	ND-TLBO	Proposed Hybrid ELITE TLBO-GA
Top 3	0.35	0.39	0.4
Top 5	0.42	0.45	0.46
Top 8	0.55	0.57	0.58
Top 10	0.69	0.7	0.72

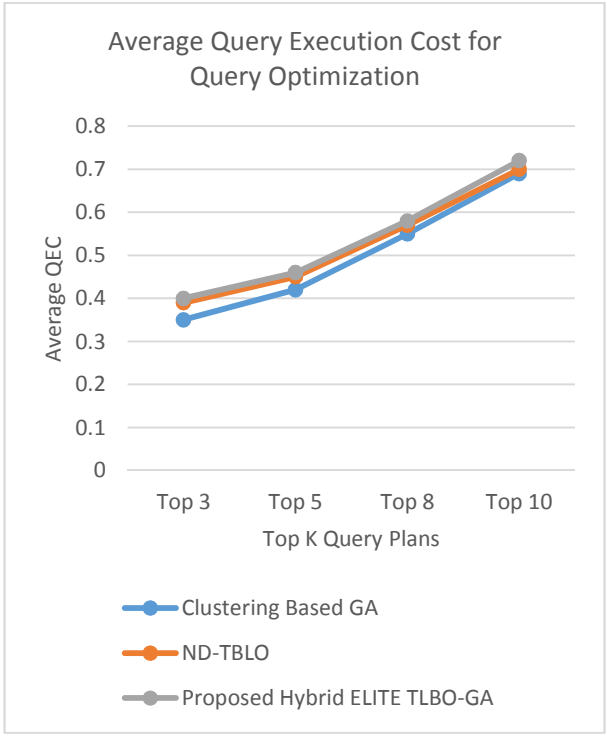

Figure 9: Average Query Execution Comparison for three different Query Optimization Approaches

5. CONCLUSIONS

In this paper, a new approach for Query Optimization in Heterogeneous DDBMS is addressed, for both local and global optimization separately. In this paper, two stochastic approaches such as multi-objective genetic algorithm for local optimization and teacher-learner based optimization for global optimization is employed. The local optimization approach deals with optimization within the local sites whereas global optimization works with the sites at different locations globally. The join ordering cost (JOC), Total Local Processing Cost (TLPC) and Total Communication Cost (TCC) are used to obtain the best optimal query plans in the proposed approach amongst the relation between the query sites. The Experimental Analysis of the proposed approach showed that it has better performance i.e. less cost when compared with other heterogeneous DDBMS whereas has more cost when compared with the other existing homogeneous DBMS and existing approaches.

The future scope could be to introduce an intelligent approach for the heterogeneous distributed database system with less time and less cost along with the accurate query plans.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES

[1] Bitton D, Boral H, DeWitt D J. Wilkinson W K, Parallel algorithms for the execution of relational database operations. *ACM Transactions on Database Systems*. 1983, 8(3), pp. 324-353.

[2] Stonebraker M, Aoki P M, Litwin W, Pfeffer A, Sah A, Sidell J, Staelin C, Yu Mariposa A. A wide-area distributed Database System. *The VLDB Journal*. 1996, 5, pp. 048–063.

[3] DeWitt D, Gray J. Parallel Database Systems: The Future of High Performance Database Systems. *Communications of the ACM*. 1992, 35(6), pp. 85–98.

[4] Hongbin Dong, Yiwen Liang, Genetic Algorithms for Large Join Query Optimization, *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*, 2007 July, pp. 07–11.

[5] Tamer Ozsu M, Patrick Valduriez, Principles of Distributed Database Systems. 2nd ed. Prentice-Hall Inc., Upper Saddle River, NJ, 1999.

[6] Reza Ghaemi, Amin Milani Fard, Hamid Tabatabaee, and Mahdi Sadeghizadeh, Evolutionary Query Optimization for Heterogeneous Distributed Database Systems, *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 2008, 2(7).

[7] Abdel kader Hameurlain, Evolution of Query Optimization Methods: From Centralized Database Systems to Data Grid Systems, *Database and Expert Systems Applications*, springer, 2009, pp. 460-470.

[8] Peter Paul Beran, Erich Schikuta, A Multi-Staged Blackboard Query Optimization Framework for World-Spanning Distributed Database Resources, *Proceedings of the International Conference on Computational Science*, 2011, 4, pp. 156-165.

[9] Vikram Singh. Multi-objective Parametric Query Optimization for Distributed Database Systems. *Proceedings of 5th International Conference on Soft Computing for Problem Solving*.2016, 436, pp. 219-233.

[10] Pankti Doshi, Vijay Raisinghani. K-QTPT: A Dynamic Query Optimization Approach for Autonomous Distributed Database Systems. *Advances in Computing, Communication, and Control*.2013, 361, pp. 1-13.

[11] Yahya Slimani, Faiza Najjar, Najla Mami. An Adaptive Cost Model for Distributed Query Optimization on the Grid. On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops.2004,292,pp.79-87.

[12] Wenjiao Ban, Jiming Lin, Jichao Tong, Shiwen Li.Query Optimization of Distributed Database Based on Parallel Genetic Algorithm and Max-Min Ant System.*8th International Symposium on Computational Intelligence and Design (ISCID)*.IEEE, 2015.

[13] Umut Tosun.Distributed database design using evolutionary algorithms. *Journal of Communications and Networks*. 2014 August, 16(4), pp. 430-435.

[14] Xiaofei Zhang, Lei Chen, Min Wang. Efficient Parallel Processing of Distance Join Queries Over Distributed Graphs. *IEEE Transactions on Knowledge and Data Engineering*.2015 March, 27(3), pp. 740-754.

[15] Vikash Mishra, Vikram Singh. Generating Query plans for Distributed Query Processing using Teacher-learner Based Optimization.*11th International Multi-Conference on Information Processing*.2015, pp. 281-290, Elsevier.

[16] S Venkata Lakshmi, Valli Kumari Vatsavayi. Query optimization using clustering and Genetic Algorithm for Distributed Databases. *International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 2016.

[17] S Venkata Lakshmi, Valli Kumari Vatsavayi. Query Plan Generation in DDS Using Non-Dominant Based Teacher-Learner Optimization (ND-TLBO) Algorithm. *International Journal of Soft Computing*. Medwell Journals, 2016, 11(3), pp.145-154.