Journal of Engineering and Applied Sciences 13 (11): 4080-4092, 2018
ISSN: 1816-949X

Big Data Clustering Using Grid Computing and Bionic Algorithms Based an Entropic Optimization Technique

Saad M. Darwish, Adel A. El-Zoghabi and Moustafa F. Ashry
Institute of Graduate Studies and Research, Department of Information Technology,
University of Alexandria, Alexandria, Egypt

Abstract: More effective marketing, along with new revenue opportunities, enhanced customer service, improved operational efficiency, competitive advantages over peer organizations and huge business benefits are the outcome of the analytical findings. The organizations performance is raised to the maximum using big data which transforms the tremendous amounts of data into knowledge. Performance and utilization of the grid computing are basically dependent on a complex and excessively dynamic way of optimally balancing the load between the available nodes. This study introduces a framework for big data clustering which utilizes grid technology and bionic based algorithms. Analysis of Genetic agorithm, ant colony optimization and particle swarm optimization are implemented regarding to their solutions, issues and improvements concerning load balancing in computational grid. Consequently, a significant system utilization improvement was attained.

Key words: Grid computing, big data, bionic algorithm, load balancing, fault tolerance, significant, concerning

INTRODUCTION

Recently, the ongoing expansion of big, complex, mixed and multi-dimensional data has emphasized the capacity of the existing data management software and hardware infrastructure. Currently, many different sorts of data sources such as sensors, scientific instruments and the Internet, are found contributing to the data booming. Huge challenges that require straight away attention from academia, research and industry are faced due to the relatively tardier evolution of new and efficient data models to process complex and large-scale data (Sharma *et al*., 2014). As traditional data models which are basically relational in nature, can't handle the today's data needs a new technology in data science is produced introduces a fast development of a broad range of non-relational data models which currently are popularly known as NoSQL and/or big data models (Sharma *et al*., 2015). Big data analytics is defined asthe process of inspecting large data groups including miscellaneous data types to reveal concealed patterns, unknown correlations, market trends, customer preferences and other useful busines information. Now a days big data is a reality: The size, heterogeneity and speed of data penetrating any organization continue to attain unexpected levels. This remarkable extension requires not only understanding big data in order to decipher the information that counts indeed but also comprehending the potentialities of big data analytics. Nevertheless, many challenges exist in dealing with big data such as storage, transfer, management and manipulation of big data (Sharma *et al*., 2015; Bajaber *et al*., 2016). High-performan ceanalytics are compulsory to process that huge data in order to determine what's important and what isn't. The usage of high-performance data mining, predictive analytics, text mining, forecasting and optimization on big data gives the possibility to perpetually reach innovation and make optimum decisions. Big data is featured by wide data sets with heterogeneous data types which can be ranked asstructured, semi-structured or unstructured (Chandhini and Megana, 2013). Structured data fits appropriately into tables with neatformat, resulting in relatively smooth manage and process. Structured data is characterized by being easily entered, stored, queried and analyzed. This implies to manufacturing data stored in relational databases and data from manufacturing execution systems and enterprise systems. On the other hand, images, text, machine log files, human operator generated shift reports and manufacturing social collaboration platform texts are considered unstructured data that may be in a raw format and necessitates decoding before data values can be extracted. Semi-structured data is defined as a sort of structured data that does not adapt to the formal structure of data models associated with relational databases or any other

Corresponding Author: Saad M. Darwish, Institute of Graduate Studies and Research, Department of Information Technology, University of Alexandria, Alexandria, Egypt

data table forms, yet, contains tags or other markers to separate semantic elements and executes hierarchies of records and fields within the data. In manufacturing, the power of big data revolution stems from the possibility to merge and correlate these data set types to create business value through newfound insights. Another benefit of big data technology is it helps manufacturers to sum up and centralize different types of data in both cost effective and scalable way. There are four approaches to analytics and each falls within the reactive or proactive category (Toga and Dinov, 2015).

Reactive business intelligence: In the reactive category, Business Intelligence (BI) provides standard business reports, ad hoc reports, OLAP and even alerts and notifications based on analytics. This ad hoc analysis addresses at the static past which is aimed to a limited number of situations.

Reactive big data BI: By reporting pulls from huge data sets it is said that this is performing big data BI. But decisions based on these two methods are still reactionary.

Proactive big analytics: Implementing forward-looking, proactive decisions needs proactive big analytics such as optimization, predictive modeling, text mining, forecasting and statistical analysis. They enable the identification of trends, spot weaknesses or the determination of conditions for making decisions about the future. But even if it's proactive, big analytics cannot be executed on big data because traditional storage environments and processing times cannot be maintained.

Proactive big data analytics: The usage of big data analytics enables the extraction of only the relevant information from terabytes, petabytes and exabytes and the analysis of it to transform business decisions for the future. Becoming proactive with big data analytics isn't a one-time effort; But it implies more of a culture change a new approach of making profit freeing analysts and decisionmakers to face the future with solid knowledge and insight. Resources of data are from web applications such as Google and Facebook which produce a huge amount of data due to the big number of customers that they have. Big data can also be gathered from areas such as science, finance, communication and business. sources of big data can be divided into human-generated data and machine-generated data. Companies such as Google, Horton works and Amazon tried to offer solutions for big data by using MapReduce and Hadoop (Lee *et al*., 2015; Szabo *et al*., 2014). Map/Reduce is a software in distributed computing environment which relies on two functions called map and reduce. The functions are designed to work with a list of inputs. The map function creates an output for eachitem in the list while the reduce function results in a single output for the entire list. Hadoop is a software library framework for developing highly scalable distributed computing applications. The Hadoop framework deals with the processing details in such a way that enables developers to focus on application logic. Yet the use of MapReduce and Hadoop is still in the early stages in manipulating bigdata (Singh and Reddy, 2004; Bajaber *et al*., 2016). Classification and clustering of big data is significant to reveal useful information and knowledge. However, the algorithms behave poorly in reference to the computation time when the size of the data becomes large and even are impractical without modification when the data exceeds the capacity of memory. Hardware constraints occur when processing big data such as the limitation in storage capacity and the processing speed (Ku-Mahamud, 2013). Grid technology has come up as a new epoch of wide distributed computing with high-performance orientation. Grid resource management is known as the process of distinguishing requirements, corresponding resources to applications, allocating those resources, scheduling and observing grid resources continually in such a way to run grid applications a sefficiently as possible (Berman *et al*., 2003). Resource scheduling of the tasks could be solved by focus on the scheduling of a set of independent tasks (Elavarasi *et al*., 2011; Singh and Chana, 2016). Other works have been based on scheduling technologies using economic/market-based models (Sharma *et al*., 2010). There are two different types of job scheduling known to be NP-complete, the use of non-heuristicsis an effective method that practically matches its difficulty and on the other hand, single heuristic approaches for the problem include local search, simulated annealing and tabu search are produced (Sharma *et al*., 2010). In recent years, some new-type bionic algorithms are become hot research topics suchas Genetic Algorithm (GA), Particle Swarm Optimization (PSO) algorithm, Ant Colony Optimization (ACO) algorithm, Bees Algorithm (BA), Artificial Fish Swarm Algorithm (AFSA) (Sharma *et al*., 2015). In the ant colony algorithm, each job submitted issues an ant which searches through the network finding the best node to deliver the job to it. Ants catch information from each node they pass through as a pheromone in each node. This eventually helps other ants to find their paths more efficiently. In the particle swarm algorithm each node in the network individually acts as a particle which sends or receives jobs from its neighbors to optimize its load locally, resulting in a partially global optimizationof the

load within the whole network given time constraint limitations (Li *et al*., 2015; Rathore and Chana, 2014). An infrastructure which supports he scalability, distribution and management of data are necessaryfor the application of big data solutions. Hence, this research proposes grid methodology to surmount the hardware limitation such as storage space, processing power and memory size. For algorithms calability, bionic-based entropic clustering algorithm is proposed.

MATERIALS AND METHODS

Big data applications: Cloudera (2017), since, the firms that sponsored this report are all good examples of software andhardware vendors that offer tools, platforms and professional services for managing big data, let's take a brief look at the product portfolio of each. The sponsors form a representative sample of the vendor community, but their suggestions include different approaches to big data management. Cloudera is aleader in enterprise analytic data management powered by Apache Hadoop and the top contributor to the Hadoop open source community. Cloudera's 100% open source distribution, CDH is comprehensive and widely deployed and it includes: Cloudera impala for interactive SQL of data in HDFS and HBase through popular BI tools, Cloudera search to enable non-technical users tointuitively explore Hadoop data, plus enterprise capabilities such as Sentry for fine-grained, role-based access control and native high availability for extreme fault tolerance. Cloudera enterprise, available as a commercial subscription, couples that platform with Cloudera's support and a suite of system and data management software built for the enterprise, including: Cloudera manager to simplify and reduce the cost of Hadoop configuration, deployment, upgrades and administration; Cloudera navigator for audit and access control of Hadoop data and optional support for impala, search and Hbase. Cloudera also offers consulting services and a wide range of Hadoop training and certification programs. More than 700 partners across hardware, software and services have teamed with Cloudera to ensure maximum integration with customer's existing investments. Leading companies in every industry run Cloudera in production, Including over 65% of the Fortune 500 in finance, telecommunications, retail, internet, insurance, energy, healthcare, biopharmaceuticals, networking and media, plus three of the biggest intelligence agencies and two of the leading three defense agencies.

Quest Diagnostics (2000) for years, Dell Software has been acquiring and building software tools (including the acquisition of Quest Software and its database development and administration tool, toad) aiming to assemble a comprehensive set of IT administration tools for securing and managing networks, applications, systems, endpoints, devices and data. Within that portfolio, Dell Software now offers a range of tools specifically for data management with a focus on big data and analytics. Forinstance, toad data point provides IT users with data provisioning and administrative functions formost traditional databases and packaged applications, in addition to novel big data platforms such as Hadoop, MongoDB, Cassandra, SimpleDB and Azure. Toad business intelligence Suite combines toad data point, toad decision point for data visualizations and a server component called toad intelligence central for collaboration to provide a totally integrated self-service BI solution that work salongside corporate BI systems. Shareplex supports Oracle-to-Oracle replication, besides Oracle-to-Hadoop replication. Kitenga analytics suite is a big data analytics tool that combines search, visualization and analytics into a platform that enables rapid transformation of diverse unstructured data into actionable insights. Dell Boomi is a data integration service that is based on cloud andconnects any combination of hosted and on-premises applications. Combine all the new Dell Software capabilities with Dell's traditional Hadoop and hardware options and Dell is well positioned to provide a true end-to-end big data analytics solution.

Oracle Corporation (2016) has recently made a substantial investment in new products and platforms engineered to provide all phases of big data management especially scalability and real-time operation specifically for advanced analytics with big data. For example, the Oracle big data appliance integrates and optimizes all the hardware and software components needed to build comprehensive analytic applications. The big data appliance in cludes Cloudera's distribution of Hadoop, Cloudera manager, a distribution of R, Oracle Linux, Oracle NoSQL Database and the Oracle HotSpot JavaVirtual machine. Oracle big data connectors provides a gateway from Hadoop and NoSQL Databaseto Oracle Database 11g and 12c. As another example, the Oracle exalytics in-memory machine provides high-performance in-memory analytics as required of growing practices such as business performance management, operational BI and working analytics. Other platforms conducive to high-performance big data management and analytics include the Oracle exadata database machine and the Oracle exalogic elastic cloud.

Pentaho Hitachi (2017), Inc., is famous for its unified, open, embeddable, pluggable business analytics platform which strictly binds both Data Integration (DI) and

Business Intelligence (BI). As we movedinto the age of big data analytics, Pentaho evolved Pentaho Data Integration (PDI) (an enterprise class, graphical ETL tool) to include support for multiple layers and types of Hadoop, NoSQL and analytic appliance environments and added its visual MapReduce tool that excludes the requirement for the complex coding normally required. Pentaho evolved its business analytics suite as well. Pentaho business analytics now also includes visualization tools, data mining and predictive algorithms, plusan analytic modeling workbench in addition to its traditional BI reporting and dashboard tools. Pentaho supports the entire big data analytics process. Pentaho makes this unified approach a practical reality by generating data integration logic and predictive models in Hadoop-friendly Java. And itsengine can be run directly in the Hadoop cluster without producing code. This means that Pentaho-based solutions are easy to embed in Hadoop and a new adaptive big data layer from Pentaho ensures portability of Pentaho solutions across Hadoop distributions from Cloudera, Hortonworks, MapR and Intel in addition to Cassandra, MongoDB and Splunk.

SAPSE. (2016) provides a comprehensive set of solutions for big data, including analytic applications, rapid deployment solutions, BI and advanced analytic tools, analytic databases, data warehousingsolutions and information management tools. Furthermore, SAP enables its customers to include Hadoop into their existing BI, advanced analytic and data warehousing environments in various ways allowing customers to customize Hadoop to their requirements. Many customers are deploying Hadoop alongside SAP HANA, an in-memory database used for real-time analytics and other applications. Customers can use SAP data services to search and load data from HDFS or Hive into SAP HANA or SAP Sybase IQ or they can use SAP HANA smart data access to push queries into Hive Hadoop (orother data sources) and achieve data virtualization with virtual tables in SAP HANA. To accommodatethe extremes of big data management, version 16 of SAP Sybase IQ software recently achieved a guinness world record for loading and indexing big data, achieving an audited result of 34.3 Terabytes Per hour. Streaming and CEP are ably served by SAP Sybase ESP. Furthermore, SAP business objects BI, SAP visual intelligence and SAP predictive analysis users can query Hive environments, giving business analysts the possibility to explore Hadoop data directly. New in-memory spatial and enhancednatural language features in SAP HANA allow organizations to reveal richer and more significant signals from business and geospatial data. Finally, the completion of the integration of Sybase data management with SAP HANA will further transform customer's end-to-end data management landscape.

SAS Institute Inc. (2016) is a leader in analytic solutions that provide business performance improvements. SAS deals with the big data challenge holistically and promotes a lifecycle methodology that orchestrates data management, preparation and exploration to model development and governance. Recently, SAS launched its high-performance analytics product that supports the Hadoop approach to data filing, query processing and management. For example, SAS visual analytics provides access to the shared memory of the Hadoop cluster for users to provide an in-memory exploratory analytics platform, allowing users to visualize and assess huge quantity of structured data and text in the Hadoop ecosystem. SAS high-performance analytics is an in-memory suite of analytic solutions that empowers analysts to build predictive models (including those using text) and offer insights in minutes without moving data outside the Hadoop cluster. From a data management perspective, SAS providesaccess to Hadoop (via. HiveQL), Oracle Exadata, SAP HANA, Teradata, Vertica, and IBM PureDatafor Analytics (formerly known as Netezza) using SAS/ACCESS software. Users can also apply existing Map Reduce, Pig or Hive code from within the SAS environment, supporting Cloudera and similar systems. SAS Data Federation Server provides an intuitive, graphical interface to integrate and transform data to and from Hadoop and other sources. Finally, the new SAS event stream processing engine implements a kind of complex event processing to manage streaming big data and execute related high-volume, real-time tasks like risk management and anti-fraud analytics.

RESULTS AND DISCUSSION

Framework for big data with grid technology: Grid technology is used for data storage and data processing while bionic-based algorithm isutilized for data clustering in the proposed big data framework manipulation as depicted in Fig. 1.

The flow chart starts with various data sources sending data to the databases. These databases construct the data intensive computing capable of handling high volume data flows (Almuttairi, 2015). Due to the requirements of data intensive applications, data storage elements are necessary to allow to scale tolarge capacities with features like reliability and availability, flexibility, manageability and security (Sharma *et al*.,

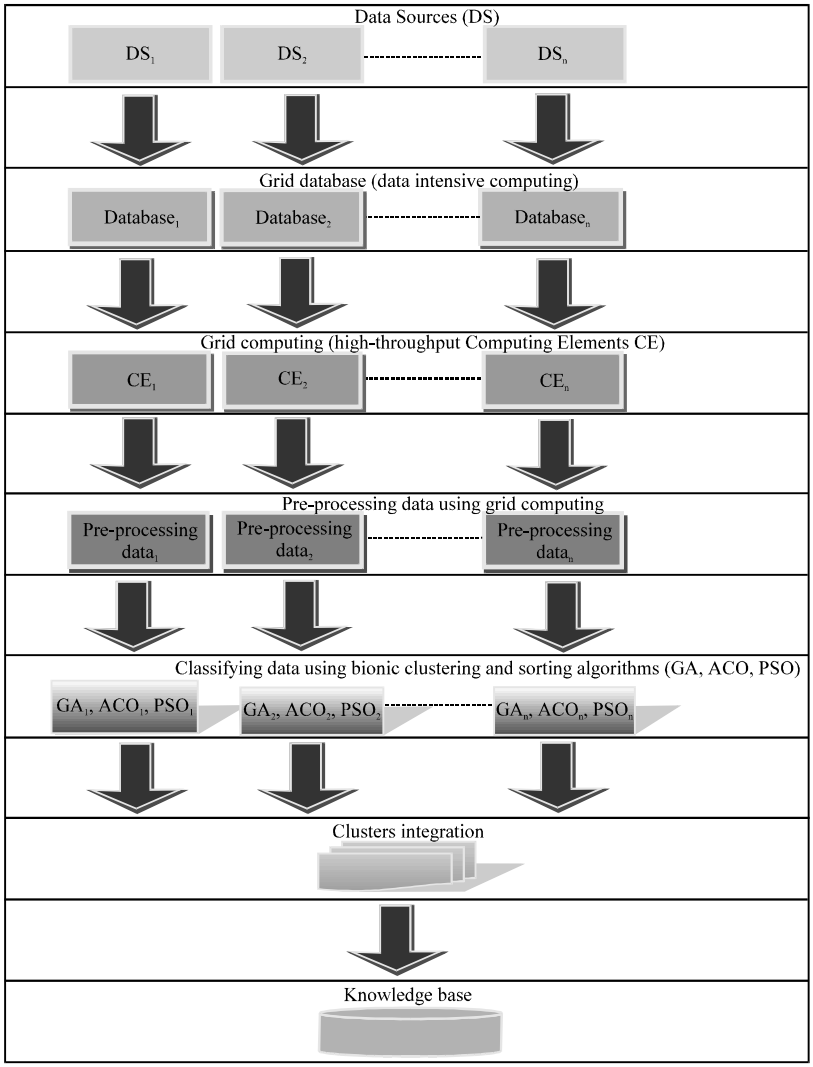

Fig. 1: The proposed framework for big data manipulation

2010; Wang *et al*., 2017). An open source projects called Hadoop distributed file system and Kosmos file system could be employed to address the requirements (Singh and Reddy, 2014). The following layer is high-throughput computingagents having the advantage of using unused processor cycles in the grid computing which can execute independent tasks (Chandhini and Megana, 2013; Wang *et al*., 2017). Thus, complicated process can be split into multiple tasks and process them by the computing elements in the grid computing environment (Sharma *et al*., 2010).

The divide and conquer method will load the memory with as much data as possible, cluster it and save some sort of module (cluster representation) of the data for future use. In other word, each computing element will be loaded with data from data intensive layer according to its memory size. For additional complex grid technology scenario, the job could be split into multiple tasks and distributed on more computing element using sophisticated resource management system or job scheduling suchas co-scheduling of parallel job proposed by Darwish *et al*. (2015a, b).

The preprocessing data stage performs data cleaning, data representation and data scaling to address outliers, missing values, in consistent values and duplicate data. Techniques such as aggregation, sampling, dimensionality reduction, feature creation, discretization and binarization and variable transformation can be applied (Darwish *et al*., 2015a, b). Clustering process can be executed once the data preprocessing is completed. Bionic-based clustering algorithm is proposed as it provides a powerful nature-inspired heuristics for solving

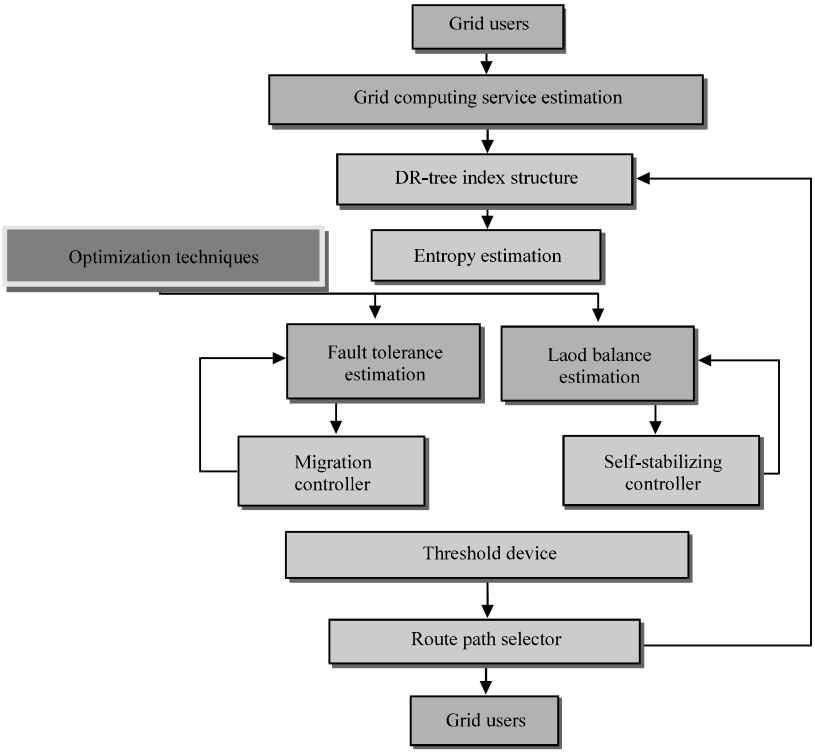

Fig. 2: The complete adaptive proposed algorithm

the clustering problems and this can be used in each computing element on different data set as proposed by Darwish *et al*. (2015a, b). The final module in the framework will integrate the clustersfrom all computing elements and save them in knowledge base.

Bionic based entropic clustering algorithm: The optimization procedure relies on the complete adaptive proposed method which is shown in Fig. 2. It starts with estimating Grid Computing Service (GCS) parameters then mapping this gridstructure into DR-tree index structure enhanced by the entropy method to minimize the completiontime of the decision maker. At the end, we use threshold device to distinguish the route path depending on two parameters load balance and fault tolerance controller. The fault tolerance is improved by using migration controller while self-stabilizing controller is used to improve the load balance in cumulative condition way and this method depends on the suggested method introduced by Darwish *et al*. (2015a, b).

The optimum solution is reached with applying three different optimization techniques on the herein proposed system, namely: genetic algorithm, ant colony optimization and particle swarm optimization. Computing jobs along with their hardware requirements are submitted by every user to the GCS. The GCS then replies to the user by sending the results after finishing the processing of the jobs (Darwish *et al*., 2015a, b).

At the first step GCS estimation will analyze the network parameters by determining the three-level top-down view of the grid computing model depending on the method produced by (El-Zoghdy, 2011) as shown in Fig. 3. Level 0: Local Grid Manager (LGM), the network is subdivided into geographical areas where any LGM manages a group of Site Managers (SMs). Level 1: Site Manager (SM), every SM is assigned the management of processing elements (computers or processors) cluster which is dynamically configured (i.e., processing elements may join or leave the cluster at any time). Level 2: Processing Elements (PE) any public or private PC or workstation can register within any SM to jointhe grid system and offer its computing resources to be exploited by the grid users. Right after beingadhered to the grid, a computing element triggers the GCS system which will reply to the SM with some information about its resources like CPU speed.

First the tree model of grid computing nodes is converted into DR-tree (Darwish *et al*., 2015a, b) for the similarity features then addressing with the new tree with the DR-tree conditions. Entropy Darwish *et al*. (2015a, b) can be defined as the average self-information that is the

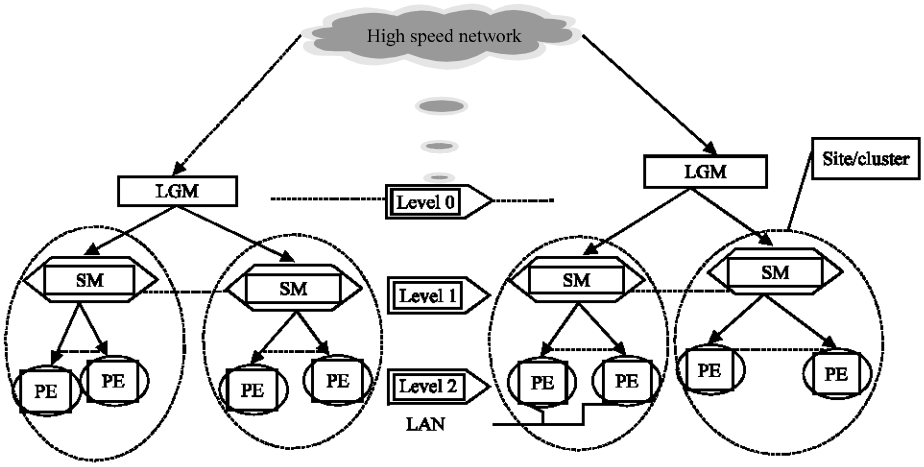

Fig. 3: Grid computing model structure

mean (expected or average) amount of information for anoccurrence of an event x_i. In the context of message coding, entropy can be represented as the minimum bound on the bits average number for each input value. The function H has the following lower and the upper limits:

$$0 = H(1, 0, 0, ..., 0) \le H(p_1, p_2, p_n) \le H\left(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right) = \log n \quad (1)$$

Searching the pool of domain blocks is considered time consuming as good approximations are obtained when many domain blocks are allowed. This method consists of omitting the domain block having high entropy from the domain pool and hence all the unnecessary domains will disappear from the pool to reach a higher production domain pool. This way will reduce the overhead of the network by decrease the number of searching nodes and then improve the performance of the grid computing networks. During GCS estimation, the grid manager will trigger Algorithm 1 by picking up some parameter ε.

Algorithm 1; Bionic based entropic clustering algorithm:

1. Initialization choose parameter ε
2. Divide the input grid into n: Virtual_Node
3. For (j = 1; j. n; j ++) {
4. H = entropy (Virtual_Node)
5. If (H.ε)
6. Void onSplit (n:Virtual_Node)
7. If n is Virtual_ Root then
8. Initiate new Virtual_ Root = create n.Virtual_Node
9. Calculate n.virtual_ father = initiate new Virtual_Root
10. End if
11. Calculate m where m = select n.virtual_children
12. Initiate then calculate new Virtual_Node = create m.Virtual_Node
13. Divide D = n.virtual_children
14. Create new Virtual_Node.virtual_children = D
15. Create new Virtual_ Node.virtual_father then put n.vrtual_father into it
}

Optimization techniques: A load balancing algorithm should optimize the employment of available resources either in the grid such as computational or data resources, in addition to time or cost related to these resources, etc. The grid environment produces a dynamic search space and consequently this represents a partial optimal solution that enhances the performance. The proposed approaches optimize the parameter ε by combining genetic algorithm, ant colony and particle swarm optimization following the initialization process in the above mentioned Algorithm 1. This leads to the achievement of the optimum resource utilization. The details of the last three optimizing methods were discussed earlier by Li *et al.* (2015).

Genetic Algorithm (GA): GA is known as part of the group of Evolutionary Algorithms (EA). The evolutionary algorithms uses the three key principles of the natural evolution: natural selection, reproduction and species diversity, preserving the differences of each generation with the previous, respectively. Genetic algorithm usually deals with a group of individuals, providing possible solutions of the task. Giving an individual assessment corresponding to the desired solution, the selection principle is implemented by using a criterion. The matched individuals represent the next generation. The huge diversity of problems in the engineering environment, equally as in other fields, necessitates the application of algorithms from different type, with multiple characteristics and settings. Algorithm 2 demonstrates the GA procedure used for select optimum parameter ε.

Algorithm 2; Genetic algorithm:

1. Initialize population of individual grid user jobs to both longest and smallest to fastest processor with 60 random schedules
2. Evaluate the fitness of all individuals
3. While termination condition not met do
4. Select fitter individuals for reproduction at minimum execution time
5. Crossover between individuals by two-point crossover
6. Mutate individuals by simple swap operator
7. Evaluate the fitness of the modified individuals that keeping relevant fitness
8. Generate a new population
9. End while
10. Initiate Authorized task for sequence Si For each processor Pi
11. Concatenate sequences Si. A permutation sequence of tasks will be assigned to processors

Ant Colony Optimization (ACO) algorithm: Ant Colony Optimization (ACO) is designated as an analytical way for solving optimization problems according to the population met. ACO provides satisfying solutions to the optimization problems by rearranging work among the nodes by using of ants. The ants navigate the grid network leaving the pheromones on the path. After reaching the destination, the ants update the pheromones tables. The pheromone trail laying and tracing the behavior of real ants depicts the main source of ACO. The ants move from node to node, so as to examine the information defined by the pheromones values and hence incrementally build the resultant solution. Algorithm 3 depicts ant colony algorithm procedure used for select optimum parameter ε.

Algorithm 3; Ant colony optimization algorithm:

```
Input: Parameters for the ACO
Output: Optimal solution to the problem
Begin
    Initialize the pheromone
    While stopping criterion not satisfied do
        Position each ant in a starting node
        Repeat
        For each ant do
        Chose next node
        by applying the state transition rate
        End For
        Until every ant has built a solution
        Update the pheromone
    End While
End
```

Particle Swarm Optimization (PSO) algorithm: Due to its simplicity and ability to successfully facing these problems, particle swarm optimization technique is also applied in many optimization and search problems. PSO optimizes an objective function by constantly enhancing a swarm of solution vectors, known as particles, with the usage of special memory management technique. Each particle is altered by attributing the memory of individual swarm's best information.The swarm regularly improves its best observed solution and converges to an optimum with the help of the aggregated intelligence of these particles. Every element ranges from 0-1 and conversely. Additionally, each particle processes a velocity vector with dimension D whose element's range is [-Vmax,Vmax]. Velocities are interpreted according to probabilities that a bit will be in one state or the other. At the beginning of the algorithm, a group of particles with their velocity vectors are developed randomly. Then in some phase the algorithm's aim isto obtain the optimal or near optimal solutions referring to its predefined fitness function. Two best positions, pbest and nbest are used to modify the velocity vector in each step and these are used to update the particles position. Pbest and nbest are D-Dimensional, having the elements consisting of 0 and 1 evenly like particles position and depicts the algorithm's memory. The personal best position, pbest is the best position the particle has reached whereas nbest is the best position reached by the particle and its neighbors, since, the first time step. When all of the population scale of the swarm turns to be the neighbor of a particle, nbest is titled global best (star neighborhood topology) and in case the smaller neighborhoods are set for each particle (e.g., ring neighborhood topology), then nbest can be named local. Algorithm 4 shows particle swarm optimization algorithm procedure exploited to select optimum parameter ε.

Algorithm 4; Partical swarm optimization algorithm:

1. Randomly particle and position will store in a created and initialized m x n matrix
   Where
   'm' = denotes a number of node
   'n' = denotes a number of job
2. Calculate the estimated time to complete values of each node using Range Based Matrix
3. Calculate the Fitness of each Particle and search for the Pbest and Gbest
4. X_k is estimated and in case its value is greater than the fitness value of pbestk, pbestk is replaced with X_k. Where X_k is denotes the updated position of the particle
5. Replace nbest with pbest because Fitness value of nbest is smaller than pbest
6. Until to reach the max Velocity
7. If it is satisfying the maximum velocity
8. Stop
9. End
10. If the maximum velocity is not satisfied
11. Update the Position Matrix
12. Update the Velocity Matrix
13. End

Migration controller of the fault tolerance: Reinsertion policy and replication policy are employed to assess DR-tree insertion processes which uses internal virtual nodes. At the moment when the crash of one non leaf physical node was produced for every DR-tree, the system restoration cost in relation to the number of messages and stabilization time is computed referring to both the reinsertion and replication policies.

Stabilization time: The reinsertion mechanism is accountable for balancing the system in a set ofcycles in proportion to both, the tip of the participation graph and the level of the burst physical node. The stabilization time is proportional to $\log_m$ (N) because it is the time of the longest reinsertion.

The message cost of the restoration phase: Expressed by the number of messages needed to balance the system from a non leaf physical node burst. Due to the variation in the magnitude of the costs, thenumber of message distribution is much skewed with the reinsertion policy, which causes a highstandard deviation.

Self-stabilizing controller of the load balance: To calculate the mean job response time, one LGM scenario is assumed as a simplified grid model. In this scenario, the time elapsed by a job in the processing elements is monitored. The jobs arrive sequentially from clients to the LGM with the premise of a time-invariant poisson process. In addition to being independent, the inter-arrival times are also identically and exponentially distributed with the arrival rate λ jobs/sec. Simultaneous arrivals are excluded. Each of PE in thenon-static site pool will be represented by an M/M/1 queue. Jobs that arrive to the LGM will be naturally allocated on the sites specified by that LGM with a routing probability, $PrS_i = SPC_i/LPC$.

Conforming to the Load Balancing Policy (LBP) if i is the site number $\lambda_i = \lambda \times PrS_i = \lambda \times SPC_i/LPC$. Comparably, the site i arrivals will be also naturally distributed on the PEs managed by that site with a routing probability PrE_{ji}/SPC_i depending to the LBP where j is the PE number and i is denoted as the site number $\lambda_{ij} = \lambda_i \times PrE_{ij} = \lambda_i \times PEC_{ij}/SPC_i$. As the simulation of the arrivals to LGM is to follow a poisson process, the arrivals to the PEs will also follow a poisson process. Let us consider that the service times at the jth PE in the ith SM are exponentially shared with fixed service rate μ_{ij} jobs/sec. They will depict the PE's Processing Capacity (PEC) in our load balancing policy. The service controlis first come first serviced. In order to get the expected mean job response time, E [T_g] is assumed todenote the mean time spent by a job at the grid to the arrival rate λ and E [N_g] denotes the total number of jobs in the system. So, the mean elapsed time by a job at the grid is stated by Eq. 2 as follows:

$$E\left[N_g\right] = \lambda \times E\left[T_g\right] \quad (2)$$

E [N_g] can be calculated from the summation of the mean number of jobs in every PE at all grid sites, so, $E\left[N_g\right] = \sum_{i=1}^{m}\sum_{j=1}^{n} E\left[N_{PE}^{ij}\right]$ where i = 1, 2, .., m is the number of site managers handled by a LGM, j = 1, 2, ..., n is the number of processing elements handled by a SM and is the mean number of jobs in a processing element number j at site number i. Because every PE is represented as an M/M/1 queue $E[N^{ij}_{PE}] = p^{ij}/1\text{-}p_{ij}$ where $p_{ij} = \lambda_{ij}/\mu_{ij}$, $\mu_{ij} = PEC_{ij}$ for PE number j at site number i. Referring to Eq. 2, the expected mean job response time is given by:

$$E\left[T_g\right] = \frac{1}{\lambda} \times E\left[N_g\right] = \frac{1}{\lambda} \times \sum_{i=1}^{m}\sum_{j=1}^{n} E\left[N_{PE}^{ij}\right] \quad (3)$$

Note that the stability condition for PE_{ij} is p<1. Algorithm 5 will be executed to calculate the traffic intensity p_{ij} and hence, the expected mean job response time:

Algorithm 5; Traffic intensity:

1. Obtain λ, μ where λ: is the external job arrival rate from grid clients to the LGM μ : is the LGM processing capacity
2. Calculate p = λ/μ is the system traffic intensity. For the system to be stable ñ must be less than 1
3. For i= 1 to m
4. Calculate λ_I, μ_i where λ_i is the job arrival rate from the LGM to the ith SM which is controlled by that LGM, μ_i is the ith SM processing capacity
5. Calculate the traffic intensity of the ith SM, $p_i = \lambda_i/\mu_i$
6. For j = 1 to n
7. Calculate λ_{ij}, μ_{ij} where, λ_{ij} is the job arrival rate from the ith SM to the jth PE controlled by that SM, μ_{ij} is the jth PE processing capacity which is controlled by the ith SM
8. Calculate the traffic intensity of the jth PE which is controlled by ith SM, $p_{ij} = \lambda_{ij}/\mu_{ij}$
9. Calculate the expected mean job response time, E [T_g]
10. End
11. End

Threshold device: A novel 2-D figure of merit is employed to test the network performance and a detailed discussion produced by Darwish *et al.* (2015a, b). An example of the 2-D figure of merit is shown in Fig. 4.

Cluster integration: An approach called cluster ensembles proposed by Strehl and Ghosh (2002) can be used to combine multiple portions of a group of objects to form a single consolidated cluster. Three effective and efficient techniques to obtain high-quality combiners which are known as consensus functions will be applied. The first combiner focuses on the similarity measurement in the partitions and then re-clusters the objects. The second combiner relies on hyper graph partitioning and the third technique collapses groups of cluster into meta-cluster which then competes for each object to obtain the combined clustering. The combiner examines only the cluster label but not the original features. In other words the combiner works with the output from any algorithms that were used to obtain these clusters.

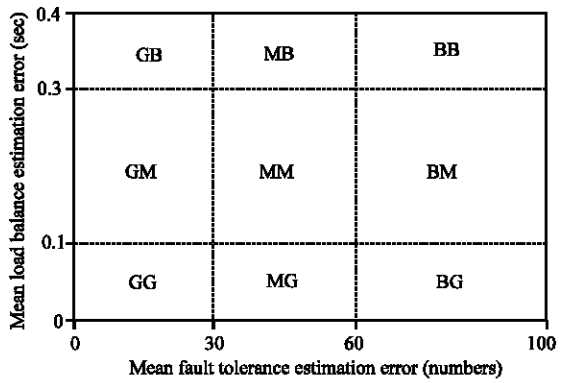

Fig. 4: The proposed 2-D figure of merit

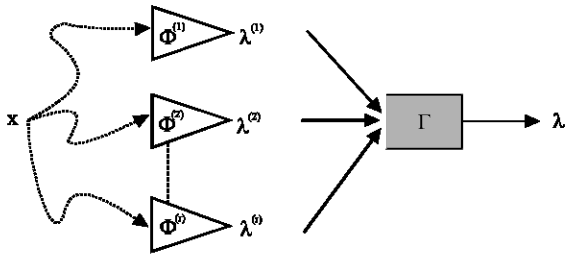

Fig. 5: Cluster ensemble model

Figure 5 shows the cluster ensemble model where X is specified as a group of features, ϕ represents the clustering algorithm, λ represents the cluster and Γ is the consensus function (combiner).

Clusters ensemble approach by Yang and Kamel (2006) where multi-ant colonies algorithm for clustering is suggested. The approach consists of two sections. The first section focuses on several independent and heterogeneous ant colonies where each uses ant-based clustering algorithm. In the second section, a queen ant agent (also called master) uses a hyper graph model which is proposed by Strehl and Ghosh (2002) to gather the output clusters from each ant colony. Each ant colony works simultaneously to produce clusters and send them to the queen ant agent. The queen ant agent merges the clusters to update and broadcast the similarity matrix and then the process is iterated. The approach which uses a queen ant agent to aggregate the clusters is very suitable to be adopted. The output clusters from ant-based clustering algorithm will be sent to the queen ant agent for combination process using aggregation with hyper graph model. Based on Strehl and Ghosh (2002) and Yang and Kamel (2006) implementation, the first step is to transform the output clusters label into a suitable hyper graph representation. A hyper graph consists of vertices and hyper edges. The regular graph edge connects exactly two vertices. The hyper graph is a stereotype of an edge in that it can bind any set of vertices. The queen ant agent will construct a new similarity matrix to combine the clusters depending on the similarity. The benefit of using queen ant as an agent is that the computing of the new similarity matrix will be done centrally by the queen ant agent rather than letting information exchanged locally by all the colonies. A simulation model is constructed using MATLAB simulator to assess the performance of grid computing system according to the proposed algorithm. This simulation model is consisting of one local grid manager which manages a number of site managers which in turn manages a number of processing elements (workstations or processors). All simulations are performed on a PC (Dual Core Processor, 2.3 GHz, 2 GB RAM) using Windows 7 Professional OS and the data-sets are available from http://strehl.com/. Finally, replication time and message cost have to be decreased in order to improve the estimation of fault tolerance and this will leads to an increase in probability of job completed. Whereas a decrease in mean job response time results in an improve in load balance estimation and this will lead to an augmentation in number of jobs/second. The improvement ratio (gain) can be calculated and viewed in Fig. 6 and from this figure, we can see that maximum improvement ratio is 98%. The second experimental results is shown in Fig. 6, here fault tolerance estimation varies with grid size and it can be shown that grid size for the proposed method surpasses that of the old algorithm stated by El-Zoghdy (2011) and the calculated values tested at various levels of entropy values. The threshold device opts for the best route path for the processing element of the all members of the grid then determine the best system to be stable according to load balance and fault tolerance policies. As this best selected route path is applied and feed back again in order to choose the best value for the entropy threshold that is used to enhance both load balance and fault tolerance, load balance is found to be unchanged which validates the stability condition of DR-tree for the system. But the experimental results demonstrate the improvement of fault tolerance with various entropy threshold values with improvement ratio for fault tolerance is 33%. The last experiment proves that the decrease in the entropy threshold under 80% of its values affects the stability of the system and leads to inaccurate output result this can be illustrated in Fig. 6 as the total final improvement ratios are 98% for load balance and 33% for fault tolerance which are very good enhancement ratios.

Finally, applying optimization algorithms to improve the performance of system utilization by using parameters as follows: No. of dimensions = 20, No. of particles = 64,

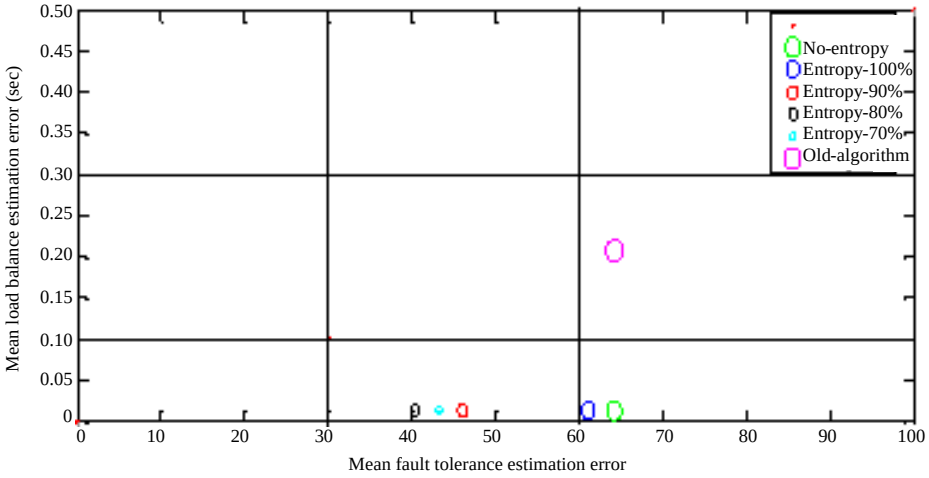

Fig. 6: Path selector with 2-D figure of merit with different entropy levels compared with the old algorithm

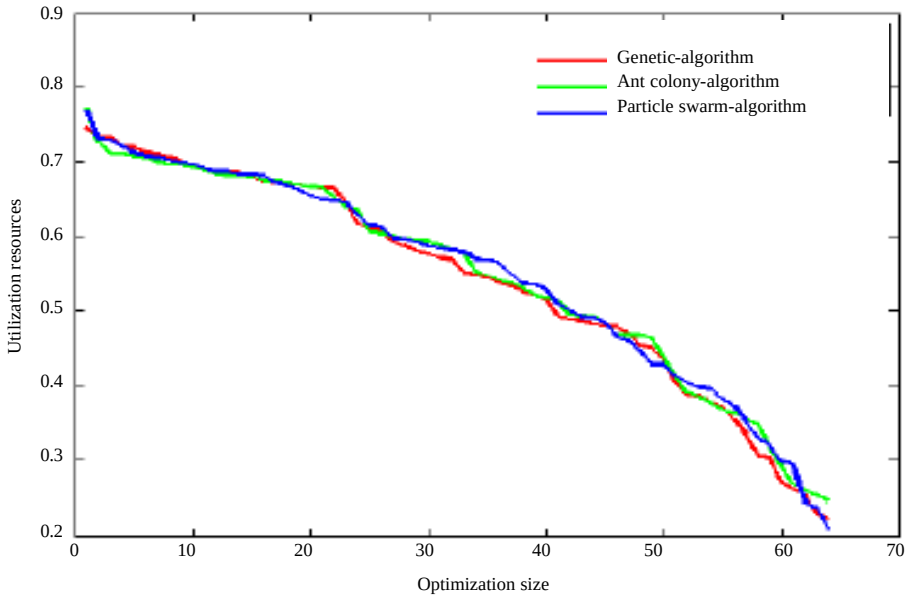

Fig. 7: System utilization

No. of iterations = 1000. As the three different bionic optimization algorithms are compared according to system utilization, the experimental study prove that they are nearly alike but PSO algorithm give the best performance and the system utilization decrease about 75% as illustrated in Fig. 7.

CONCLUSION

This study introduces a new adaptive methodology depending on advanced fractal transform that enhances the tree model structure of grid computing environment to enhance the network performance parameters affected by fault tolerance and load balance equally. First, an estimate of the fault tolerance and load balance for the network parameters have been calculated based on fractal transform. In this study, simulator of fault tolerant method for load balancing in grid environment is constructed. In this one local grid manager with 9 sites and each site have 1-3 processing elements and queue length of each computing element is from 1-50 and this was illustrated in a 2-D figure of merit. The grid computing routing protocol is ameliorated by improving fault tolerance with load balance estimation in a novel 2-D figure of merit. The improvement of the fault tolerance estimation is carried out by reducing replication time and message cost and this will result in an increase in the probability of job completed. On the other hand, reducing mean job

response time results in an enhancement of the load balance estimation and this in turn will induce an increase in number of jobs/second. Finally, the improvement ratios are 98% for load balance and 33% for fault tolerance which are very good enhancement ratios. Also when assimilating the system utilization by three various optimization algorithms GA, ACO and PSO they have nearly the same outcome but PSO yield to the best performance and decrease the system utilization to 75% which is high performance parameters. Further experimentation can be done by implementing our system on a real grid computing network and study its performance. A framework for the clustering of big data using grid computing and bionic algorithm has been proposed. The grid concept is to allow the storage of data in distributed databases across a wide geographical area while bionic-based algorithm is for the clustering of big data. Entropic bionic-based algorithm has many advantages to be used in big data mining because it has the potential to scale with the size of the data set, prior knowledge of the count of expected clusters is not needed and easy to integrate with clusters ensemble model. Big data analysis gives the chance for many research areas and one of the most important areas is the data security.

SUGGESTIONS

Big data analytics and the internet of things in manufacturing as an end-to-end platform is the critical base to enlarge the vision of smart manufacturing. This platform is scalable and available in various configurations using currently available industry-standard building blocks.

REFERENCES

Almuttairi, R.M., 2015. Taxonomy of optimization approaches of resource brokers in data grids. Intl. J. Comput. Sci. Inf. Technol., 7: 135-144.

Bajaber, F., R. Elshawi, O. Batarfi, A. Altalhi and A. Barnawi *et al*., 2016. Big data 2.0 processing systems: Taxonomy and open challenges. J. Grid Comput., 14: 379-405.

Berman, F., G. Fox and A.J. Hey, 2003. Grid Computing: Making the Global Infrastructure a Reality. In: Wiley Series in Communications Networking and Distributed Systems, Berman, F., G. Fox and A.J. Hey (Eds.). John Wiley & Sons, Hoboken, New Jersey, USA., ISBN:9780470853191, pp: 809-824.

Chandhini, C. and L.P. Megana, 2013. Grid computing-a next level challenge with big data. Intl. J. Sci. Eng. Res., 4: 1-5.

Cloudera, 2017. Be one with your data introducing SDX: Experience your data whenever, wherever and however you'd like. Cloudera, Palo Alto, California, USA. http://www.cloudera.com

Darwish, S.M., A. Adel and M.F. Ashry, 2015. An entropic optimization technique in heterogeneous grid computing using bionic algorithms. Intl. J. Comput. Sci. Inf. Technol., 7: 19-37.

Darwish, S.M., A.A. El-zoghabi and F.A. Moustafa, 2015. Improving fault tolerance and load balancing in heterogeneous grid computing using fractal transform. World Acad. Sci. Eng. Technol., 2: 1-1.

El-Zoghdy, S.F., 2011. A load balancing policy for heterogeneous computational grids. Intl. J. Adv. Comput. Sci. Appl., 2: 93-100.

Elavarasi, S.A., J. Akilandeswari and B. Sathiyabhama, 2011. A survey on partition clustering algorithms. Intl. J. Enterp. Comput. Bus. Syst., 1: 1-14.

Hitachi, 2017. Introducing Hitachi vantara. Hitachi, Chiyoda, Tokyo, Japan. http://www.pentahobigdata.com

Ku-Mahamud, K.R., 2013. Big data clustering using grid computing and ant-based algorithm. Proceedings of the 4th International Conference on Computing and Informatics ICOCI, August, 28-30, 2013, Universiti Utara Malaysia, Sarawak, Malaysia, pp: 6-14.

Lee, W.H., H.G. Jun and H.J. Kim, 2015. Hadoop mapreduce performance enhancement using in-node combiners. Intl. J. Comput. Sci. Inf. Technol., 7: 1-17.

Li, C.B., S.K. Li and Y.Q. Liu, 2015. Comparative study of typical bionic intelligent optimization algorithm. Intl. J. Manage. Sci. Eng. Res., 2: 17-27.

Oracle Corporation, 2016. Experience oracle cloud-get $300 in free cloud credits. Oracle Corporation, Redwood City, California, USA. http://www.oracle.com

Quest Diagnostics, 2000. Spend less time on it administration and more time on it innovation. Quest Diagnostics, Madison, New Jersey, USA. http://www.dellsoftware.com

Rathore, N. and I. Chana, 2014. Load balancing and job migration techniques in grid: A survey of recent trends. Wireless Pers. Commun., 79: 2089-2125.

SAPSE., 2016. Turn insight into action with SAP HANA. SAP SE, Walldorf, Germany. http://www.sap.com

SAS Institute Inc., 2016. The power to know. SAS Institute Inc., Cary, North Carolina.

Sharma, R., V.K. Soni, M.K. Mishra and P. Bhuyan, 2010. A survey of job scheduling and resource management in grid computing. World Acad. Sci. Eng. Technol., 64: 461-466.

Sharma, S., R. Shandilya, S. Patnaik and A. Mahapatra, 2016. Leading NoSQL models for handling big data: A brief review. Intl. J. Bus. Inf. Syst., 22: 1-25.

Sharma, S., U.S. Tim, J. Wong, S. Gadia and S. Sharma, 2014. A brief review on leading big data models. Data Sci. J., 13: 138-157.

Sharma, S., U.S. Tim, S. Gadia, J. Wong and R. Shandilya *et al*., 2015. Classification and comparison of NoSQL big data models. Intl. J. Big Data Intell., 2: 201-221.

Singh, D. and C.K. Reddy, 2014. A survey on platforms for big data analytics. J. Big Data, 1: 1-20.

Singh, S. and I. Chana, 2016. A survey on resource scheduling in cloud computing: Issues and challenges. J. Grid Comput., 14: 217-264.

Strehl, A. and J. Ghosh, 2002. Cluster ensembles a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res., 3: 583-617.

Szabo, C., Q.Z. Sheng, T. Kroeger, Y. Zhang and J. Yu, 2014. Science in the cloud: Allocation and execution of data-intensive scientific workflows. J. Grid Comput., 12: 245-264.

Toga, A.W. and I.D. Dinov, 2015. Sharing big biomedical data. J. Big Data, 2: 1-10.

Wang, S., K. Li, J. Mei, G. Xiao and K. Li, 2017. A reliability-aware task scheduling algorithm based on replication on heterogeneous computing systems. J. Grid Comput., 15: 23-39.

Yang, Y. and M.S. Kamel, 2006. An aggregated clustering approach using multi-ant colonies algorithms. Pattern Recognit., 39: 1278-1289.