BIG DATA OPTIMIZATION TECHNIQUES: AN EMPIRICAL STUDY

Issa M.S.Ali Dr. B.Mukunthan

Abstract— On account of the world being digitized speedily in which the magnitude of data is over owing from diverse sources in various formats, it is not potential for the classic system to compute and analyze this kind of huge data for which big data tools like Hadoop is used which is open-source software. It stores and computes data in a divided environment. Since a decade Big Data Application development has become increasingly paramount. Many organizations are relied on getting knowledge essence from a huge amount of data. However classic data technique demonstration includes reduced performance, accuracy, slow responsiveness and lack of scalability. To resolve the complicated Big Data problem, many of the work has been carried out. For that various types of technologies have been developed. This research paper focuses on the survey of recent optimization technologies and their Applications developed for Big Data. Its purpose is to help to choose the right collaboration of different Big Data technologies to approve to the requirements.

Keywords: Big Data, Bio Algorithms, Hadoop, Optimization, Scalability.

1. INTRODUCTION

Big data is a set of data groups that are so huge and complex so it is risky to manage the database using just one tool or conventional data processing APPs. Challenges comprise capture, duration, storage, research, participation, transport, analysis, and visualization. The trend to huge data-sets is due to the additional information derived from the analysis of one large set of relevant data, compared to smaller separate groups with the same total data volume from 2012 the size specified on the data sets suitable for processing in a plausible amount of time was subject to the exabyte measurement unit. Scientists often face many constraints due to big data sets in many areas, including meteorology, genetics, complex physical simulation, and biological and environmental research. Restrictions also affect Internet search, business technology, and finance. Data sets are increasing in volume on one side due to their influence in sensor wireless[1][2], sensor networks, transmitter frequency sensors, mobile information sensors, microphones, program registers, and cameras.

The global technological amplitude of storage information per person has its reach to multiples every 40 months of the 1980s and, starting in 2012, generates 2.5 quintillion bytes (2.5 x 1018) of data per day. The defiance for massive companies is to determine who should have large data initiatives spread across the organization.

2. WHAT IS BIG DATA?

Big data is usually described as relevant datasets for large data volumes to fail with these difficulties, new “large data” platforms are being developed, in research reports and years of lectures in 2001, Doug Lani, a Meta Group analyst (now known as Gartner), defined the challenges of data growth and opportunities. And diversity (diversity of data types and sources). Gartner along with many companies in the industry now and in 2012, updated his definition to read: "big data is the asset of large, high-speed, and/or high-impact”, TBDI is a term applied to large objects of data that vary in nature whether organized, unorganized or semi-structured, including external or internal sources of the organization, and are created with high speed with a turbulent model. It is fully compatible with traditional and organized data repositories and requires a robust complex ecosystem with a high rendering computing platform and analytical capabilities to capture, process, transform, discover and extract value and insights deep within a passable time frame."

---

• Issa M.S.Ali ,Research scholar, Department of CS, Sri Ramakrishna College of Arts & Science, Coimbatore, India, E-mail: issamohammed.cs@gmail.com. Tel.: 00916281597461.
• B. Mukunthan, Research Supervisor & Associate Professor, Department of Computer Science, School of Computing, Sri Ramakrishna College of Arts and Science, Coimbatore - 641006, Tamil nadu, India. Email: dr.mukunthan.bmk@gmail.com, https://orcid.org/0000-0001-8452-3164
A. Volume

The main feature that makes data "big" is the phenomenal volume and in these days, we are in the situation where every second we are generating a huge amount of data from Twitter messages, WhatsApp, Facebook, photos, sensor data and video clips that we produce and post it on social media every second. We are talking about zettabytes or brontobytes of data. The huge volume of the data needs space and different processing technologies than classic storage and processing space.

B. Variety

One of the reasons why we call data the huge data is the variety. We get the Big Data from a big variety of provenance and in general that data are three species:

1) Structured Data

The structured data is data that can be stored, accessed it and processed it in the form of a stable format. Structured data indicates to high-level data types of organization, like information in a relational database.

2) Semi-structured Data

The semi-structured data is data or information that has not resided in a relational database but it has some organizational characteristics that make it easy for us to analyze and store it in a relational database. Examples of this type: JSON and XML documents are semi-structured, NoSQL databases also as semi-structured.

3) Unstructured Data

If the data do not have structure or any organizational characteristics to classify it as semi-structured referred to as unstructured data. In other words, any data has unknown form or unknown structure is categorize as unstructured data. It often includes multimedia and text content.

C. Veracity

Veracity refers that the data being analyzed is of high quality and accurate, which are reflected in sound engineering, decision making based on analysis of these data. In contrast, data of low resolution and low quality contain a lot of data to be discarded or what we called noise. In order to get rid of this data, which is not valuable and because it is of large size and has high speed, we need to use advanced tools.

D. Velocity

Though, we call data large data, it must be generated very fast at the same time we should analyze it and process it fast also to get the information the faster we process your information we can make the right decision at the right time.

3. TECHNOLOGIES USED FOR BIG DATA

With the evolution of technology and multitudes of data fluxing in and out of companies daily, it has become difficult to manage data with old technologies and it is necessary to find new technologies capable of absorbing this large amount of data to be processed and stored in quality and efficiency. In this division, we will describe some of those tools related to this subject.

- units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.

A. Hadoop

Hadoop an open-source software platform that is designed to store and process data. The Hadoop framework outfits reliability and data traffic for applications. Hadoop includes a portion called MapReduce, in this portion applications are divided into many small portions of the work, each portion may be implemented or re-implemented on any node in a cluster. Besides, it provides a distributed file system that stores data on the contract of the account, providing a very high degree of width of the total movement across the block. Both map / reduce and the distributed file systems are designed to handle the problem of node failure automatically by the frame. Cementing apps to work with many freelance computers and bits of data. The Apache Hadoop platform is currently made up of the Hadoop kernel, MapReduce and the Hadoop distribution system (HDFS), as well as related projects including Apache Hive, Apache H-base, etc...

It is noteworthy that Hadoop has written in the programming language "Java", it is also one of the leading Apache projects built and used by a global community of contributors. Hadoop and its related projects (H-base, Hive, Zookeeper, etc...) have a lot of shareholders through global software companies. Although JavaScript code is the most popular, any programming language can be used with the "flow" process to perform the "map" and "reduce" parts of the system.
Hadoop was created by Doug Cutting and Mike Cavarilla in 2005. Doug, was working at Yahoo, he named his son's game the elephant the development process was primarily to support the distribution of the Nutch search engine project.

B. MapReduce

MapReduce is a programming pattern appropriate for the processing of big data. The MapReduce consists of two parts:

1. Map phase
2. Reduce phase

1. Map phase: This phase is processing the input data. In general, the input data is in the brew of files or directories and is stored in a Hadoop file system (HDFS). The mapper processes the data and makes many small blocks of data.

2. Reduce phase: In this phase, the input data is Map phase's output. The Reducer's job is to process the data which comes from the mapper. After processing, it generates a new collection of output that will be stored in the HDFS.

C. HBase

HBase is a column-oriented runs on top of HDFS. It is designed to handle huge data sets with very large tables making it a great option for storing multiple or strung structures of data. It is incorporated with Hadoop and works smoothly beside other access engines through YARN. Apache HBase supplies access random, in real-time to our data in Hadoop. It does not support a structured query language like SQL.

D. Hive

The Apache Hive™ data warehouse software aids us with querying and controlling large datasets actual fast in SQL-like interface to query data which in various databases and file systems that Merge with Hadoop. While initially developed by Facebook, Apache Hive is used and developed by other companies such as Netflix and the Financial Industry Regulatory Authority (FINRA).[3][4] Amazon maintains a software fork of Apache Hive included in Amazon Elastic MapReduce on Amazon Web Services.[5]

4. LITERATURE SURVEY

Ilango et al. [6] they Notice that the data is increasing rapidly and becomes difficult for clustering so they proposed for management of Large data by an artificial bee colony (ABC) algorithm-based clustering technique so they designed method for reduce the time and ameliorate the regularity and because of the parallelism in ABC it was the fastest in execution for all data type when they used Hadoop that include MapReduce the result was more effective exit than particle swarm optimization (PSO ) and the differential evolution ABC has few important control parameters such as population size, limit, and maximum cycle number. The main advantage [7]Cheng et al [8] suggested a swarm intelligence (SI) algorithm for handling a large amount of data, that data is a high dimensional one which ameliorates the regularity data processing. When it is executed they noted that empty loading is reduced compared with the existing but they face problems when they want to execute it in real work because of the way of the work in the port Manikandan and Kalpana [9] suggested a fish swarm optimization (FSW) algorithm for Choosing a feature in massive data because of the fact, large data comprises of data fluxes and data emphasizes no homogeneity by using dimension chosen techniques for those databases will be reduced. They noticed that the databases will need all values for one solution includes a feature selection problem and diversifies it into a form to be dealt with by the AFSO. After they did that the results come to prove the superiority of an algorithm Proposed named FSO with CART method with 7.91% and 7.31% Respectively. These results they compared with MI-CART and FSO with a random forest and compared with the MI-random forest.Elshebiny et al. [10] suggested the intelligent water drops (IWD) algorithm to manage large data Their research Develop meta-heuristic IWD algorithm to IWDC algorithm after development they compared with of other previous cloud scheduling algorithm; MAX-MIN, MIN-MIN, according to the comparison there was an improvement It also provided better results than some algorithms which it is simpler than them like PSO and C-PSOKune et al. [11] suggested for analytics of big data a genetic algorithm (GA) this algorithm Depends on family scheduling approach they proposed to use Rough Set theory for the make groups. this algorithm reduces turnaround time because of using parallel processingSchmidt et al. [12] Suggested an artificial immune system (AIS) algorithm which is based on huge data optimization technique. This sophisticated AIS inspired algorithm is to manage and categorize Internet traffic data. In this they also modified the original algorithm and made it more predictable. When they compared their results with the Naïve Bayes classifier their result was more accurateGandomi et al. [13] suggested a multi-objective genetic programming (GP) algorithm based on the approach for the modeling of complex engineering systems. Its technique can Choose the most significant variables, formulate it an undefined structure, and solution. They linked MOGP technique suggested in their paper which is an equivalent tool for handling large data, and it can be used to provide Built-in and accurate modelsGeorge and Parthiban [14] suggested the hybridized optimization algorithm based on the firefly algorithm and Group Search Optimizer (GSO) To perform data collection using FSO for the high-dimensional
data set. When they did performance comparison this technique showed that it is performed well with all terms by using different datasets. Saida et al. [15] suggested the cuckoo search optimization (CO) algorithm based on a large data analytics approach for clustering data. They applied the algorithm to four different data groups. It is common to use the cuckoo search algorithm for clustering approaches also it is easy for using. They plan to crossbreed it with other algorithms. Wang et al. [16] suggested an FSO algorithm-based to hybrid (FSOH) tactic for huge data optimization to concentrate on six problems. The comparisons show result shows that the HMOFA better than NSGA-II, JADE, MOFA, DECC-DG, and SHADE on all test problems. Pu et al. [17] suggested a hybrid biogeography based on optimization (BBO) algorithm of analysis and processing of large data. This algorithm using To train Receptors multilayer under challenge they used Four standard big datasets (iris, heart, balloon, and vehicle) were employed to make sure that HCBBO able to perform in training MLPs. after that they compare the HCBBO result with BBO, PSO and GA the result show that the HCBBO better than BBO, PSO, and GAWang et al. [18] suggested a PSO algorithm for large data optimization. They proposed an approach to represent large-scale remote sensing to develop learn online dictionary. The suggested algorithm improves the performance and accuracy of ODL algorithms. The results show that the precision of the representation by their suggested algorithm is higher than that of ODL. Mafarja and Mirjalili [19] suggested a simulated annealing (SA) algorithm for large data optimization technique, which uses the whale optimization algorithm (WOA). The suggested approach helps to design different feature choice methods to reduce manipulation by searching the most capable areas. Neeba and Koteeswaran [20] suggested a bacterial foraging optimization (BFO) algorithm to Categorize the informative and affective content already present in medical weblogs. they suggested BFISO to improve classification procedures. The extracted features using suggested BFISO achieved the best rating precision for KNN, CART, and NB.

**Table- 1: OPTIMIZATION SURVEY**

<table>
<thead>
<tr>
<th>Author and Co-authors</th>
<th>Algorithm</th>
<th>Objective</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ilango et al.</td>
<td>Artificial bee colony (ABC)</td>
<td>To allocate scarce resources to achieve the best possible effect and reduce execution time to optimize different sizes of the data set.</td>
<td>2018</td>
</tr>
<tr>
<td>Pu et al.</td>
<td>Biogeography based optimization (BBO)</td>
<td>To solve two of the hardest defy in development (MLP) are the local-minima problem and the problem of slow convergence under big data challenge</td>
<td>2018</td>
</tr>
<tr>
<td>Neeba and Koteeswaran</td>
<td>Bacterial foraging optimization (BFO)</td>
<td>To propose improved multi-phase optimization algorithm to classify the user’s opinions or his personal sentiments, opinions and the factual information</td>
<td>2017</td>
</tr>
<tr>
<td>Elshebiny et al.</td>
<td>Intelligent water drops (IWD)</td>
<td>To enhance unprecedented algorithm expansion the natural based IWD algorithm that develop the plan of performance on the cloud</td>
<td>2017</td>
</tr>
<tr>
<td>Mafarja and Mirjalili</td>
<td>Simulated annealing (SA)</td>
<td>Two hybridization models are used to design different feature selection techniques based on Whale Optimization Algorithm (WOA). In the first model, Simulated Annealing (SA) algorithm is embedded in WOA algorithm, while it is used to improve the best solution found after each iteration of WOA algorithm in the second model. The goal of using SA here is to enhance the exploitation by searching the most promising regions located by WOA algorithm.</td>
<td>2017</td>
</tr>
<tr>
<td>Schmidt</td>
<td>Artificial immune</td>
<td>To improve the AIS</td>
<td>2017</td>
</tr>
</tbody>
</table>

**BIO ALGORITHMS FOR BIG DATA OPTIMIZATION**

<table>
<thead>
<tr>
<th>Author and Co-authors</th>
<th>Algorithm</th>
<th>Objective</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ilango et al.</td>
<td>Artificial bee colony (ABC)</td>
<td>To allocate scarce resources to achieve the best possible effect and reduce execution time to optimize different sizes of the data set.</td>
<td>2018</td>
</tr>
<tr>
<td>Pu et al.</td>
<td>Biogeography based optimization (BBO)</td>
<td>To solve two of the hardest defy in development (MLP) are the local-minima problem and the problem of slow convergence under big data challenge</td>
<td>2018</td>
</tr>
<tr>
<td>Neeba and Koteeswaran</td>
<td>Bacterial foraging optimization (BFO)</td>
<td>To propose improved multi-phase optimization algorithm to classify the user’s opinions or his personal sentiments, opinions and the factual information</td>
<td>2017</td>
</tr>
<tr>
<td>Elshebiny et al.</td>
<td>Intelligent water drops (IWD)</td>
<td>To enhance unprecedented algorithm expansion the natural based IWD algorithm that develop the plan of performance on the cloud</td>
<td>2017</td>
</tr>
<tr>
<td>Mafarja and Mirjalili</td>
<td>Simulated annealing (SA)</td>
<td>Two hybridization models are used to design different feature selection techniques based on Whale Optimization Algorithm (WOA). In the first model, Simulated Annealing (SA) algorithm is embedded in WOA algorithm, while it is used to improve the best solution found after each iteration of WOA algorithm in the second model. The goal of using SA here is to enhance the exploitation by searching the most promising regions located by WOA algorithm.</td>
<td>2017</td>
</tr>
<tr>
<td>Schmidt</td>
<td>Artificial immune</td>
<td>To improve the AIS</td>
<td>2017</td>
</tr>
</tbody>
</table>
### BIO ALGORITHMS FOR BIG DATA OPTIMIZATION

<table>
<thead>
<tr>
<th>Author</th>
<th>Algorithm</th>
<th>Objective</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>et al.</td>
<td>system (AIS)</td>
<td>algorithm, while making its execution 50–60% faster. And also replace the classification accuracy of the Euclidian distance by the Manhattan distance for this application, giving 1–2% higher accuracy. That will be made the accuracy of the AIS comparable to that of a Naïve Bayes classifier in original AIS algorithm that uses the same data set.</td>
<td>7</td>
</tr>
<tr>
<td>Gandomi et al.</td>
<td>Genetic programming (GP)</td>
<td>To suggestion (MOGP) algorithm for complex civil engineering systems. It combines the power of parameter estimation for classical regression and the ability to choose the model structure for standard genetic programming, at the same time Improves the complexity and make it appropriate.</td>
<td>2016</td>
</tr>
<tr>
<td>Cheng et al.</td>
<td>Swarm intelligence (SI)</td>
<td>To discuss the relationship between big data analytics and swarm intelligence techniques and analyze the potential applications of the swarm intelligence in the big data analytics</td>
<td>2016</td>
</tr>
<tr>
<td>George and Parthiban</td>
<td>Group searcher optimization (GSO)</td>
<td>To improve the clustering performance by inclusion of hybridized optimization technique which will employ firefly algorithm with Group Search Optimizer (GSO).</td>
<td>2015</td>
</tr>
<tr>
<td>Kune et al.</td>
<td>Genetic algorithm (GA)</td>
<td>They proposed a Genetic Algorithm based scheduler for such Big Data Cloud where decoupled computational and data services are offered as services. The approach is based on evolutionary methods focussed on data dependencies, computational resources and effective utilization of bandwidth thus achieving higher throughputs.</td>
<td>2014</td>
</tr>
</tbody>
</table>

### RESULTS

In this proposed work we compared various algorithms such as ABC, IWD, BBO, FSOH, AIS, FSW, AS, BFO, SI, GP, GSO, PSO, GA and CO based on them Scalability, Storage, Fault Tolerance, Agility, Virtualization, Cost, Ease of Use, Type of Analytics, No SQL DBMS, Mechanism, Type of Data, Dimension of Data Management and Data Mining Technique.
in which, IWD, BBO, SA, GSO AND PFO algorithms are good in clustering and ABC, BBO, GSO, GA, and FSOH algorithms are good in Storage and IWD, AIS, SA, and GP algorithms are good in Fault Tolerance respectively. From the above study it is identified that some of the hybrid algorithms are to be designed based on the requirements of the problem to be addressed. This study assists to choose algorithms that are to be fused to generate an algorithm of generic and hybrid nature.

Figure 1. Analyzes Type Comparison.

Figure 2. Data Type Comparison

Figure 3. No SQL DBMS Comparison

Figure 4. Bio algorithm Comparison

Figure 5. Data Mining Technique Comparison

Table- II: Comparison of Bio-Inspired Algorithm for Big Data Analytics
6. CONCLUSION
In this survey paper, various optimization techniques have been presented using big data tools. It is also esteemed that the existing optimization technologies, mechanisms, and techniques of big data framework are brought into focus. The individuality of this paper is that this paper gives a summary of several methods and highlights most of the substantial outcomes of existing research which is presented concisely in the above tables. Most of the examined methods deal with utilization, fairness, and starvation difficulties and we note that a single optimization algorithm cannot resolve all optimization problems and hence different problems require different algorithms. This is evident from the underperformance of best optimization algorithms. In certain complex optimization algorithms, problems due to inability to adapt themselves to those problems. This survey will be helpful for the further improvement and development of Big Data Analytics. In the next phase, we will improve the algorithm that can solve most of complex optimization problems.

REFERENCES
[3] Use Case Study of Hive/Hadoop


AUTHORS PROFILE

Issa MOHAMMED SAEED Ali, received his Bachelor of Science in Computer Science from Taiz University -Yemen in 2013 and Master of Computer Science from Bharathiar University -India in the year 2018, and his research work focuses on Algorithms, Big Data Analytics, Data Mining.

Dr. B. Mukunthan pursued Bachelor of Science in Computer Science from Bharathiar University, India in 2004 and Master of Computer Applications from Bharathiar University in year 2007 and Ph.D from Anna University - Chennai in 2013. He is currently working as Research Supervisor & Associate Professor, Department of Computer Science, School of Computing, Sri Ramakrishna College of Arts and Science(Affiliated to Bharathiar University), Coimbatore since 2017. He is a member of IEEE & IEEE computer society since 2009, a life member of the MISTE since 2010. He has published more than 25 research papers in reputed international journals. He is also Microsoft Certified Solution Developer. His main research work focuses on Algorithms, Bioinformatics, Big Data Analytics, Data Mining, IOT and Neural Networks. He also invented a Novel and Efficient online Bioinformatics Tool and filed for patent. He has 12 years of teaching experience and 10 years of Research Experience.