International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.04, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

CRITICAL ANALYSIS ON BIG DATA HANDLING AS AN OPTIMIZATION PROBLEM

Anju Bala[1], Priti[2]

[1]Research Scholar, DCSA, M.D.U Rohtak

[2]Assistant Professor, DCSA, M.D.U Rohtak

ABSTRACT

In today's world, managing and information extraction from big data is compulsory for the business enhancement. Now a day, big data is available everywhere due to growth in technology and popularity of digitization. Big data isgenerated continuously at a high speed in a huge amount with lots of variety. These features i.e. volume, variety and velocity of data works as challenge to extract the information. The handling of big data can be done efficiently by using the optimization techniques. This paper discusses handling of big data as an optimization problem along with the details of big data and optimization algorithm. This study concludes that big data can be handled easily by using optimization algorithms.

***Keywords: Big Data, Optimization, Velocity, Variety, Volume*.**

I INTRODUCTION

Big data is increasing enormously now a days. The amount of data is exploding at an extraordinary rate as a result of developments in different Web technologies, social media,mobileandsensing devices and satellite communications[1][2]. As the user requirements/demands are increasing, more and more data is contributed towards data flood[3].In 2011, McKinsey Global Institute defined big data as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse"Knowledge discovery and better decision making is possible if information is fetched efficiently from this large collection of data[4].

But knowledge discovery from this large amount of data faces a number of challenges[5][6]. Proficiency to extract value from Big Data depends on data analytics [7].Machine learning is fundamental components in data analytics[8]. All of Machine learning algorithms are based on certain assumptions. Some of which are listed here:

(i) Any algorithm can learn better with more data. If learning is better, better will be results.
(ii) Entire data set can fit in the memory.

Big data don't follow any of the assumption, so fetching the information from these large data sets has become a tedious job which require special tools, techniques and procedures embedded with already existing techniques[9]. Big data is characterized by 5 V's illustrated in figure 1.

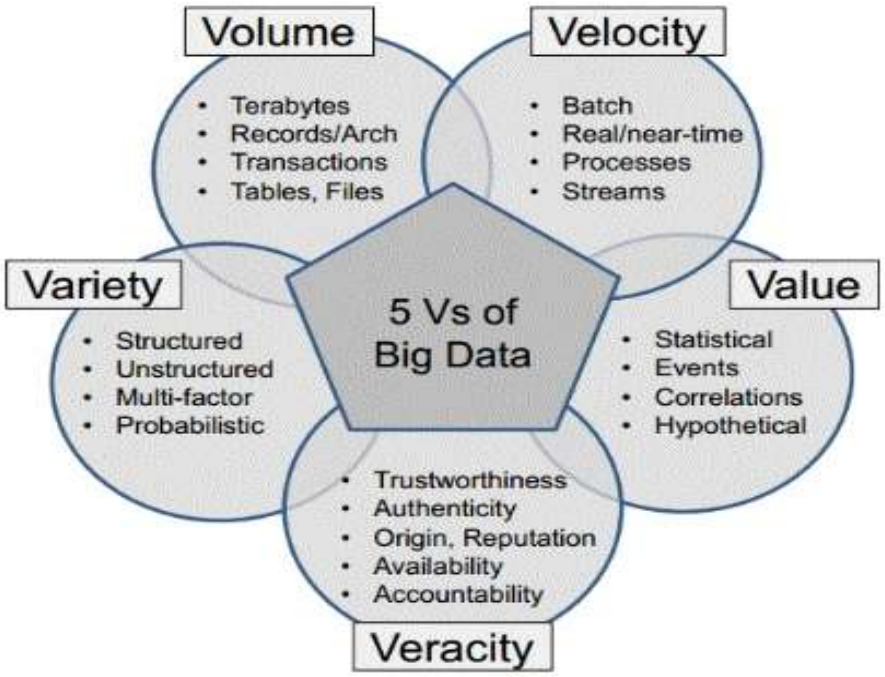

Figure 1: 5V's of Big Data

Machine learning algorithms faces a lot of challenges allied with characteristic of Big Data. Various authors gives different algorithms to handle the big data. Different authors gives different solutions to various issues faced to handle big data are given in table 1.

Table 1:showing the challenges faced by different already existing Algorithms:

Sr. No.	Dimensions of Big Data	Challenges	Solutions
1.	**Volume:** amount, size and scale of data.	(i) ProcessingPerformance (ii) Curse of Modularity (iii) Class Imbalance (iv) Curse of Dimensionality (v) Feature Engineering (vi) Non Linearity (vii)Variance and Biases	(i)RDD(Resilient distributed datasets) are used for in memory computations. (ii)K-means is used to remove curse of modularity. (iii)Regularization is used to remove generalization which is having two components VARIANCE & BIASES.
2.	**Variety:** Data types, sources and what it represents	(i) Data Locality (ii) Data Heterogeneity (iii) Dirty & Noisy data	(i)Map reduce approach is used to bring computation to data not data to the computation, (ii)Signal are to be extracted from noisy data directly.
3.	**Velocity:** Rate at which data is produced and analysed.	(i) Data Availability (ii) Real time processing/streaming (iii) Concept of drift	(i)Incremental Learning (ii)Twitter's storm and Yahoo's S4 must be merged with machine learning algorithms for solving streaming problems'.
4.	**Veracity:**Reliability of data forming the datasets.	(i) Data Provenance (ii) Data Uncertainty (iii) Dirty & Noisy data.	(i)(RAMP) Reduce and Map Reduce Provenance. (ii)Data Cleaning.

Despite the growth in these technologies and algorithms to handle big data, the big data is not handled efficiently till now. The next section discuss the optimization techniques.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.04, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

2. OPTIMIZATION PROBLEMS

A number of optimization techniques exists till now for solving the class of optimization problems[10].Optimization is a skill of selecting the best alternativeamong a given set of options. Optimization problems arise in various disciplines such as engineering design, agricultural services, manufacturing system, economics, scheduling and production planning,location and distribution management, Internet routing, etc.

2.2.1 Types of Optimization Problems: Optimization problems can be of different type such as multi objective optimization, multimodal optimization and combinatorial optimization.

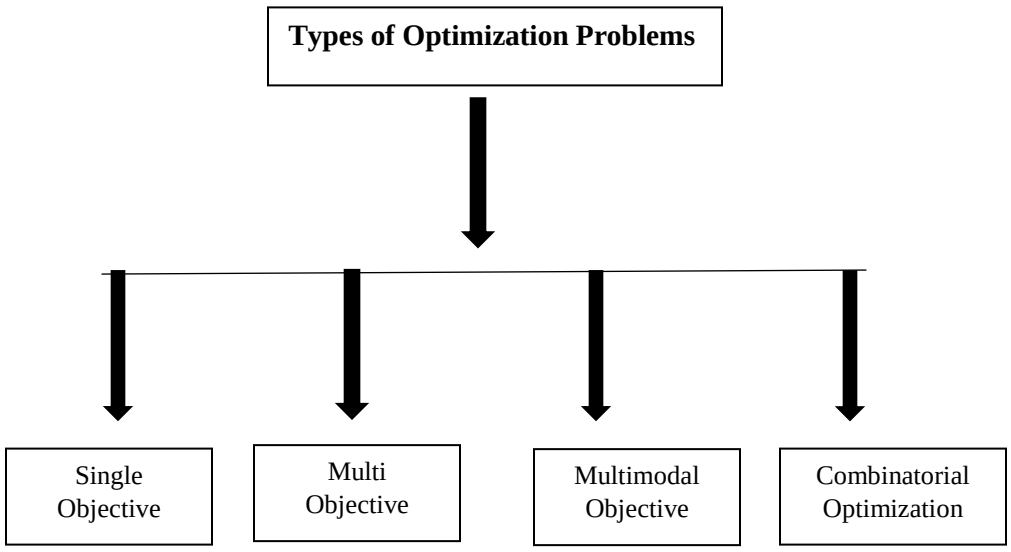

Figure 2: Type of Optimization Problems

The single objective problems either minimize or maximize a single variable according to the needs. The multi-objective algorithm is used to optimize a set of variable where in betterment of one may lead to diminish the other. The multimodal problems has several local solution but the main problem is to find the global solution. A Combinatorial Optimization Problem (COP) can be stated as a finite set of possible solution from which we look for the best one minimum or maximum[11][12].

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.04, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

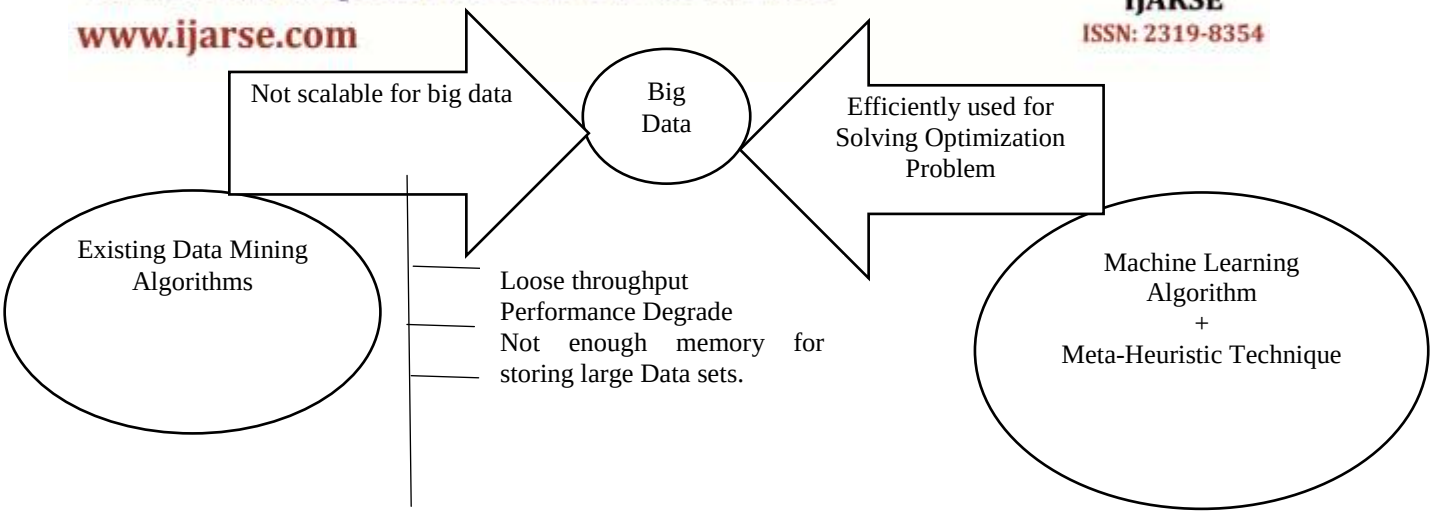

Figure 3: Big Data Mining

These real-life COP (Combinatorial Optimization Problems) are frequently characterized by their large-scale sizes. Big Data is also having one of its characteristic as large scale data, so it is considered as one of the Optimization Problems. Exact algorithms are not enough to provide good quality solutions within reasonable amount of time as shown in figure 3.

3. BIG DATA AS AN OPTIMIZATION PROBLEM

Big Data is also one of the optimization Problem due to following features:

(i) **Curse of Dimensionality:** Due to increase in amount of data tremendously, number of dimensions of data are also increasing. This is known as curse of Dimensionality. This curse of Dimensionality is making the search space to run exponentially. Optimization is a field of problems which require special techniques to find optimal solution from this large search space. Big data is considered as an Optimization Problem.

(ii) **Decision Making:** A no of big data mining algorithms are used for fetching valuable and timely information from these large data sets. Already existing techniques drop their performance and throughput when applied on big data. An Objective function is used as criteria for decision making online from the stream of data.

(iii) **Applicability of approximation:** There exist no efficient and exact algorithms for solving Optimization Problem. Optimization problems can provide accurate result if solved with the help of heuristic. A class of heuristic techniques exists which helps in big data mining, because they help in overcoming the problem of local minima.

Big Data is also considered as an Optimization problem due to the features described above.

4. OPTIMIZATION ALGORITHMS

Optimization algorithms are general-purpose heuristic method designed to guide underlying problem specific heuristictowards promising regions of the search space. This approach is proved to be more successful in obtaining near optimal solution. Optimization algorithms are categorized as deterministic and stochastic algorithms which are further categorized in to different types shown in figure 4.

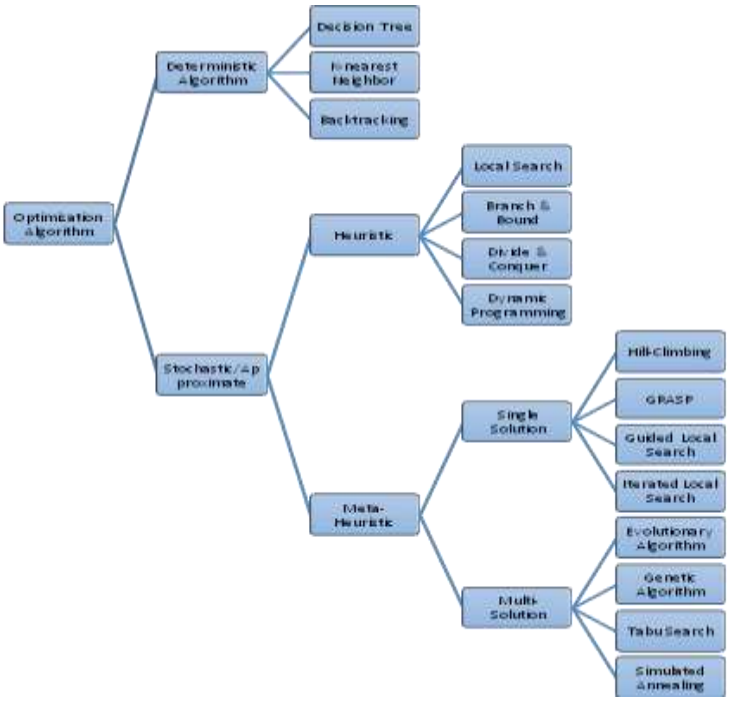

Figure 4: Optimization techniques

In software engineering, meta heuristic assigns a computational strategy that advances an issue by iteratively endeavouring to enhance an applicant arrangement as to a given measure of quality[13]. Meta heuristics make few or no suppositions about the issue being improved and can seek substantial spaces of applicant arrangements. In any case, Meta heuristics don't ensure an ideal arrangement is ever found. Numerous Meta heuristics actualize some type of stochastic improvement. The convergence speed is a suitable measure of performance when there is an algorithm that is proven to converge towed optimal solution, and in each following step finds a solution that is better or equally good as the previous solution. However, this is not the case for other existing algorithms.

5. CONCLUSION

This paper describes the big data with its different features. Then the paper designs big data as an optimization problem by correlating the properties of big data and the optimization algorithm. The paper also describes different

optimization algorithms. These algorithms can be sued to handle the data in an efficient way. In future big data can be handled efficiently by using different optimization algorithms.

REFERENCES

[1] Kuchipudi Sravanthi,Tatireddy Subba Reddy,Applications of Big data in Various Fields, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015, 4629-4632.

[2]Mohamed N, Al-Jaroodi J. Real-time big data analytics: applications and challenges. In: High performance computing & simulation (HPCS), 2014 international conference; 2014. p. 305–10.

[3] Xiaolong Jin, Benjamin W. Wah, Xueqi Cheng, Yuanzhuo Wang, Significance and Challenges of Big Data Research, Big Data Research, Volume 2, Issue 2, June 2015, Pages 59-64, ISSN 2214-5796,

[4] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107, 2014.

[5]Feng Chen,1,2 Pan Deng,1 JiafuWan,3 Daqiang Zhang,4 Athanasios V. Vasilakos,5 and Xiaohui Rong, "Data Mining for the Internet of Things:Literature Review and Challenges," International Journal of Distributed Sensor Networks,Volume 2015, Article ID 431047, 14 pages.

[6]Sin K, Muthu L (2015) Application of big data in education data mining and learning analytics—a literature review. J Soft Comput 5(4):1035–1049.

[7] S. A. Hossain, "Big Data Analytics in Education: Prospects and Challenges," in 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), Noida, India, 2-4 Sept. 2015.

[8]B. Daniel, "Big Data and analytics in higher education: Opportunities and challenges", *British Journal of Educational Technology*, vol. 46, no. 5, pp. 1-17, 2014.

[9] Archenaa, J., and Anita, M., A survey of big data analytics in healthcare and government. *Procedia Comput. Sci.* 50:408–413, 2015.

[10] Beheshti, Z. and Shamsuddin, S. M. H. "A review of population-based meta-heuristic algorithms". International Journal of Advances in Soft Computing & Its Applications, 2013;5(1):1-35.

[11] Iztok Fister Jr., Xin-She Yang, IztokFister, Janez Brest, DusanFister, "A Brief Review of Nature-Inspired Algorithms for Optimization ", ELEKTROTEHNISKI VESTNIK, 80(3): 1–7, 2013

[12] Shi, Y., Eberhart, R. C., "Empirical study of particle swarm optimization", Proceedings of IEEE Congress on Evolutionary Computation, (1999), pp. 1945–1950.

[13] Bratton, D. and Kennedy, J., "Defining a standard for particle swarm optimization", Proceedings of the 2007 IEEE Swarm Intelligence Symposium, (2007), pp. 120–127.