A NOVEL APPROACH FOR PROCESSING BIG DATA

D. Krishna Madhuri

Assistant Professor, Department of Computer Science and Engineering, GRIET, India

ABSTRACT

Applications of machine learning are widely used in the real world with either supervised or unsupervised learning process. Recently emerged domain in the information technologies is Big Data which refers to data with characteristics such as volume, velocity and variety. The existing machine learning approaches cannot cope with Big Data. The processing of big data has to be done in an environment where distributed programming is supported. In such environment like Hadoop, a distributed file system like Hadoop Distributed File System (HDFS) is required to support scalable and efficient access to data. Distributed environments are often associated with cloud computing and data centres. Naturally such environments are equipped with GPUs (Graphical Processing Units) that support parallel processing. Thus the environment is suitable for processing huge amount of data in short span of time. In this paper we propose a framework that can have generic operations that support processing of big data. Our framework provides building blocks to support clustering of unstructured data which is in the form of documents. We proposed an algorithm that works in scheduling jobs of multiple users. We built a prototype application to demonstrate the proof of concept. The empirical results revealed that the proposed framework shows 95% accuracy when the results are compared with the ground truth.

KEYWORDS

Big data, Map Reduce, Hadoop, Distributed programming framework

1. INTRODUCTION

Big data, as the name indicates, is very voluminous data with other features such as velocity (streaming data) and variety (data in different formats such as structured, unstructured and semi-structured). When data is in the form of relational database, it is known as structured data. When data is in the form of documents of any kind, it is known as unstructured data. When data is in the form of XML, it is known as semi-structured data. The characteristics of big data are shown in Figure 1 where it can be understood that velocity refers to speed, volume refers to volume and variety refers to complexity. By processing big data it is possible to gain comprehensive business intelligence that is needed by enterprises in the real world for making strategic decisions and business growth.

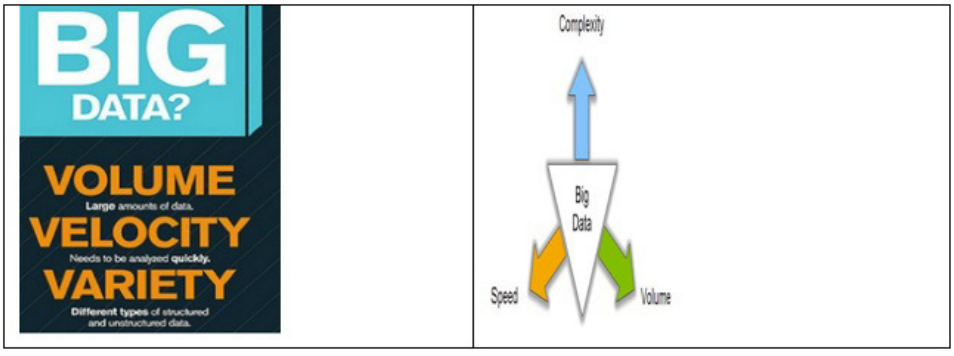

Figure 1: Shows Characteristics of Big Data

DOI: 10.5121/ijdms.2016.8502

Having understood what is big data it is essential to know why it needs to be mined. When big data is analyzed, it is possible that important and hidden trends or patterns can be obtained. When such data is not analyzed it may results in inaccurate business decisions. The ensuing sub section throws light into the need for processing big data.

1.1 Need for Big Data Mining

Big data is the complete data of business with all details. When such data is processed in a distributed environment, it is possible that it produces comprehensive business intelligence. If partial data is processed, it may result in inaccurate business intelligence that cannot be used for making expert decisions.

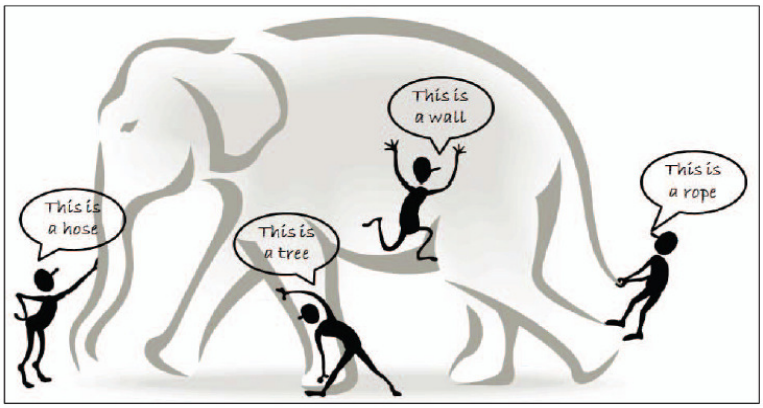

Figure 2 : Limited View of Users Provided Biased Conclusions

As shown in Figure 2, it is evident that people who are analyzing data were not able to produce correct output. There is elephant over there and people has seen a part of it and understood differently. For instance the leg of the animal is understood like a tree. Probably it is the act of blind person who cannot see the complete picture but can touch and say what it is. Here it is very obvious that biased conclusions are provided. These conclusions are wrong and they cannot help in making well informed decisions. Thus it is understood that when whole data (complete data) is considered, it can produce intelligence for making good decisions.

1.2 Big Data Evolution

Right from 1968 there has been evolution of big data. It has not happened in a year or two. It is the continuous improvement in data analytics over a period of time. In 1968 Online Transaction Processing (OLTP) was explored with day to day transactions stored in database and processed. In 1983, data warehousing technology came into existence. This has helped to have historical data to be obtained from OLTP data in order to use it for data mining. Thus historical data can be processed in order to make business intelligence out of it. This phenomenon was named as Online Analytical Processing (OLAP).

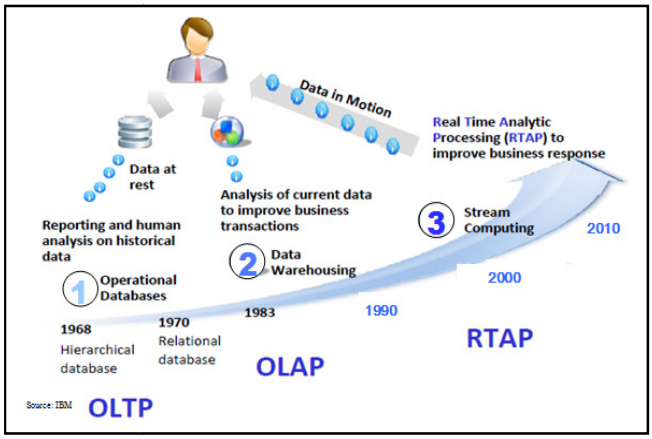

Figure 3: Evolution of Big Data

As shown in Figure 3, it is evident that the evolution is from OLTP to OLAP to RTAP. Real Time Analytical Processing (RTAP) is the subject pertaining to stream processing and big data analytics. The processing of data at rest, historical data and data in motion were the improvements from 1968 to 2010. Finally it resulted in big data and its real time processing for business analysis.

1.3 Map Reduce Programming

Map Reduce programming is a new programming approach based on object oriented programming using Java programming language. It is the process of writing program with two parts such as Map and Reduce. Map takes care of processing big data while reduce takes care of combining Map results provided by thousands of worker nodes in distributed environment. This kind of programming is supported by cloud computing, data centres and the presence of modern processors such as Graphical Processing Units (GPUs). The power of parallel processing is leveraged with big data in such environments.

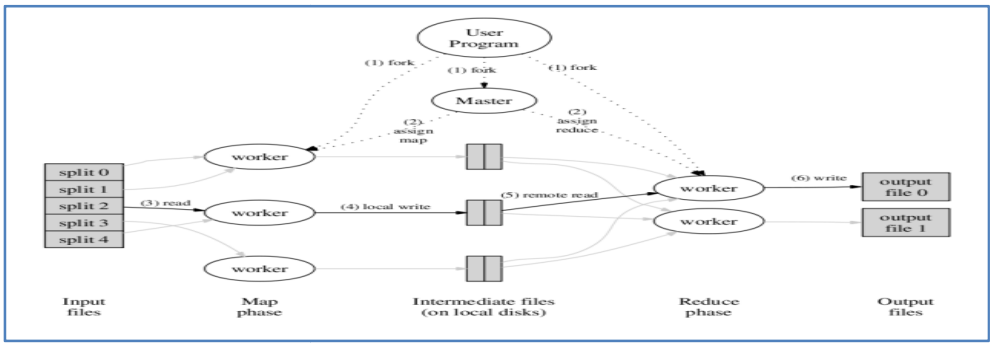

Figure 4:Map Reduce programming paradigm

The architecture shown in Figure 1 is related to Hadoop which is one of the distributed programming frameworks. The Map Reduce programming supported by Hadoop is illustrated in the figure. First of all the input files are divided into multiple parts and each part is given a mapper present in worker nodes. The mapper produces its output. Then the intermediate files are taken by other set of worker nodes for performing Reduce task. Once the reduce task is

completed, it is possible to have output files generated. Both input and output files are stored in distributed file system Hadoop.

In the existing systems, machine learning algorithms were not optimized for big data processing. In this paper we proposed a framework that supports generic machine learning process that is processing of big data. Stated differently, the proposed system takes big data (documents) as input and produces clusters. There are many operations involved that are related to processing documents. The operations include identification of keywords, finding feature space, TF/IDF and so on. Finally the resultant clusters. The remainder of the paper is structured as follows. Section II provides review of literature. Section III presents the proposed system in detail. Section IV presents experimental results while section V concludes the paper.

2. RELATED WORKS

Machine learning is the process of providing intelligence to programs so as to help them learn. The learning process may be of two types known as supervised learning and unsupervised learning. Various methods of machine learning can be found in [1], [2], [3], [4], [5] and [6]. These techniques are used in the real world. However, in the area of vision and natural language processing there is still possibility for further research. Moreover, the existing machine learning algorithms are not optimized for big data processing. The algorithms are to be developed keeping the distributed programming environment like Hadoop and new programming paradigm such as Map Reduce.

As explored in [7] Hadoop is a distributed programming framework that is used to process huge amount of data. A distributed file system is associated with Hadoop to support scalable and available processing of data. Hadoop can be compared with some in-memory products. However, Hadoop provides superior performance than its in-memory counterparts discussed in [8] and [9]. There are other systems that are powerful and versatile with low level programming interfaces [10], [11]. The problem with them is that they are specific and cannot provide general high level programming interface, scheduling and other needed mechanisms.

Making models and working with models is found in Pregel [12] which is graph-centric platform. It supports partitioning of models with in-built scheduling and mechanisms for consistency. However there is no realization of its widespread usage. Lohr [13] opined that big data bring new possibilities with machine learning algorithms. Moreover the big data is a wealth for making strategies in business. For instance Google is using such data in order to drive its business. Chen et al. [14] explored business analytics and intelligence on big data. Management information systems are improved with big data science. Jacobs [15] opined that organizations need to be careful about the pathologies of big data and carefully consider what exactly big data is. Chen et al. [16] explored the possibilities of big data in future with a good survey of articles on big data. They opined that big data can add big value to businesses when harnessed properly. Labrinidis and Jagadish [17] investigated opportunities with big data such as ability to obtain comprehensive business intelligence. They also found many challenges such as environment, algorithms, dealing with different kinds of data and so on. Kraska [18] discussed about the increasing need of big data and its processing which looks like finding a needle in haystack.

Agrawal et al. [19] explored the current state of the art on big data besides future opportunities. They focused on scalable DBMS that can help in processing big data. Snijders et al. [20] opined that knowledge gaps are filed with big data processing. Herodotou et al. [21] explored on big data and found that it can help in agility and depth in information processing and obtaining business intelligence. Cuzzocrea et al. [22] focused on big data and explored how big data revolution can work with multi-dimensional data for business intelligence. Chen and Zhang [23] studied big data

process in the context of data-intensive applications to find methodologies that can help in processing big data. Katal et al. [24] discussed on the tools already available for processing big data. The tools like Hadoop and other related products help in big data analytics. Bizer et al. [25] understood that four perspectives such as defining the problem, searching (processing), transforming, and entity resolution are important to obtain meaningful information. In this paper we proposed a generic framework that supports big data processing with machine learning algorithms. It also supports scheduling jobs in multi-user environments.

3. PROPOSED FRAMEWORK FOR DISTRIBUTED MACHINE LEARNING

We proposed a machine learning framework which is generic in nature with high level programming interface. It is meant for processing big data in multi-user environment. It provides common operations need to process unstructured data. Besides it supports algorithm for scheduling multiple jobs of users concurrently in a distributed environment. In fact the framework can support any machine learning algorithm for clustering documents. It supports information retrieval and natural language processing with machine learning to understand big data and perform clustering of documents.

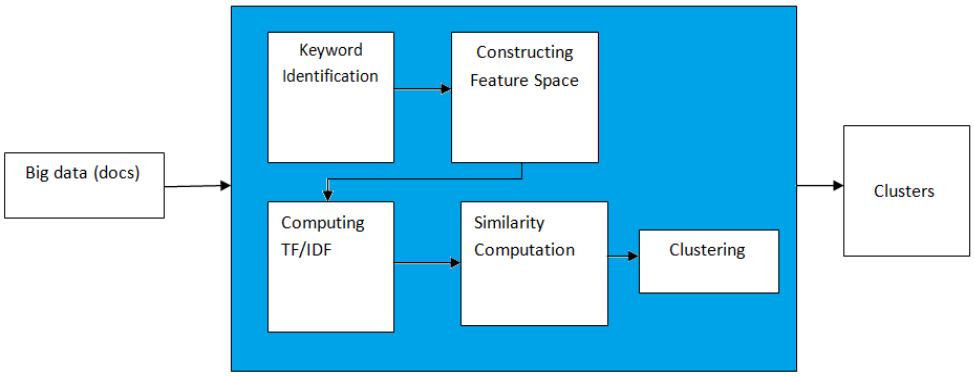

Figure 5: Generic Framework for Machine Learning on Big Data

The framework shown in Figure 5 is generic in nature and works with any set of documents that constitute big data. The framework takes big data as input and produces clusters. The operations involved in the framework include keyword identification, construction of feature space, computing TF/IDF, similarity computation and clustering. Keywords are the important words identified. The feature space is constructed in order to process the data. Afterwards, TF/IDF measure is used to find statistics based on term frequency. This will help in finding similarity between documents in order to make well-informed decisions on clustering. There are many distance measures that can be used for finding similarities between documents. They are as follows.

Jaccard Function

$$sim(d1, d2) = j(d1, d2) = f1 \cap f2 / f1 \cup f2$$

Cosine Function

$$sim(d1,d2) = C(d1,d2) = f1.f2 / ||f1||.||f2||$$

Euclidean Distance

$$sim(d1, d2) = Ec(d1, d2) = \sqrt{(f1 - f2).(f1 - f2)}$$

Extended Jaccard Function

$$sim(d1, d2) = EJ(d1, d2) \frac{f1.f2}{f1.f1 + f2.f2 - f1.f2}$$

Dice Function

$$sim(d1, d2) = D(d1, d2) = \frac{2f1.f2}{f1.f1 + f2.f2}$$

Listing 1: Shows Different Functions that are used as Similarity Measures

The listing 1 shows many similarity measures such as Jaccard function, cosine function, Euclidean Distance, Extended Jacccard Function and Dice function. All the functions are capable of supporting the similarity computation in big data processing. Especially in this paper they are used for finding similarity between two documents. The similarity measure results in a value between 0.0 to 1.0. The more this value is the more the similarity is between any two documents. Here is the proposed algorithm for multi-user job scheduling.

```
Algorithm: Multi-user job scheduling algorithm
Inputs: Jobs
Outputs: Scheduled jobs

01  Initialize job j
02  Initialize jobs vector J
03  Initialize node n
04  Initialize t1 to 15 milliseconds
05  Initialize t2 to 15 milliseconds
06  IF node n is busy THEN
07     Set j.wait=t1
08  END IF
09  IF n has free map slot THEN
10     Populate jobs to J
11     For j in J
12        IF j.wait=0 an n has still free map slot THEN
13           Assign j to n
14        END IF
15        IF j.wait =0 and n has no free map slot THEN
16           j.wait=t2;
17        END IF
18     END FOR
19  END IF
```

Algorithm 1: Multi-User Job Scheduling Algorithm

The algorithm is responsible to schedule jobs in order to improve the productivity of big data processing. The distributed machine learning environment takes number of user jobs and schedules them properly in such a way that they are processed in an optimized fashion. The waiting time concept is used to ensure that the jobs are given their turn and processing of big data is carried out in efficient manner.

4. EXPERIMENTAL RESULTS

We built a custom simulator (prototype application) that simulates distributed programming framework such as Hadoop with Map and Reduce functionalities. It supports multiple nodes in the processing and multiple jobs provided by many users simultaneously. The framework is tested with the application in terms of processing Big Data (documents) which is in unstructured format. The results are observed in terms of number of concurrent users and the performance of the proposed framework in accurate clustering of documents. The clustered documents are evaluated with ground truth.

No. of Users	10	20	30	40	50	60	70	80	90	100
Time (sec)	1	2.5	5	7.5	10	12.5	13.5	14	16	17

Table 1: Shows Performance in Terms of Workload VS. Time Taken

As shown in Table 1, there is relation between number of users and the time taken to process big data. In the simulated environment multiple users and their jobs are considered to measure performance in terms of time taken. The results revealed that when number of users increase, it also caused more time to be taken.

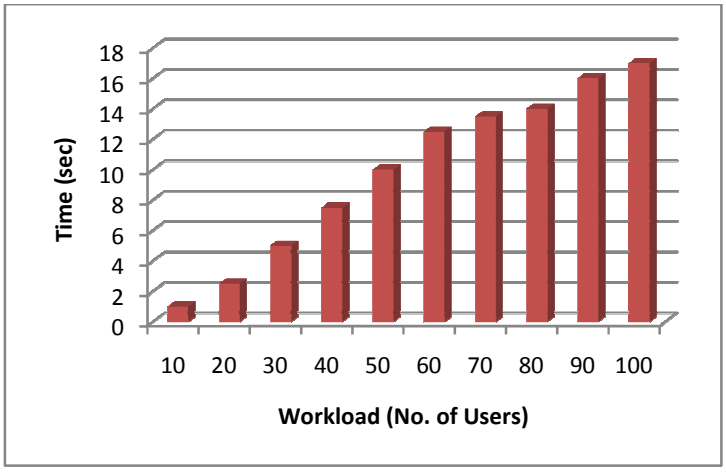

Figure 6: Clustering Performance in Multi-User Environment

As shown in Figure 6, it is evident that number of users is represented by vertical axis while the horizontal axis represents time taken in order to perform clustering of big data. The trend in the results reveals a fact that when number of users is increased the time taken for processing big data is also increased proportionately. When number of users is 100 the time taken is 17 seconds while the time taken for 10 users is 1 second.

5. EVALUATION

Human experts are involved in evaluating the work of the proposed framework. The clustering of big is done manually by spending time on a portion of big data in order to establish ground truth. The ground truth is then compared with the performance of the system. Thus the proposed framework with underlying algorithm is evaluated.

True Positives	False Positives
95	5

Table 1: Shows Results of Evaluation

Figure 7: Evaluation Results

As shown in Figure 7, the proposed framework performance is presented in the form of true positives and false positives. The system has showed 95% true positives and 5% false positives. This is carried out based on the ground truth values provided by human experts. The results revealed that the proposed system can be used effectively for processing big data. However, it needs further refinement in order to make it more useful and add value to enterprises.

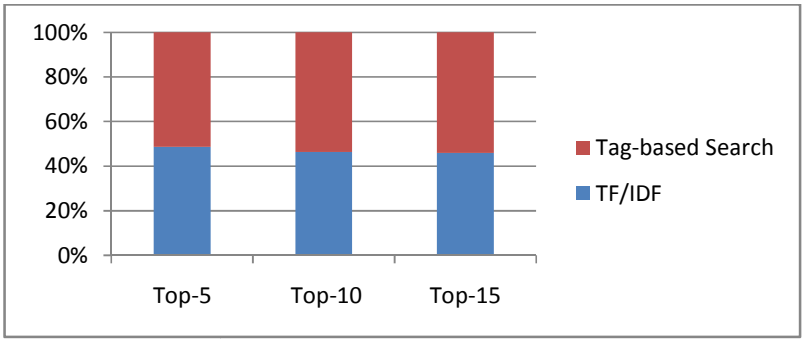

Figure 8: TF/IDF and Tab Based Search

The results shown in Figure 8 reveal that there is performance difference between tag-based search and TF/IDF based search. The top-n values are represented by horizontal axis while the percentage performance is shown in vertical axis. There is slight difference between the two kinds of searches in performance.

6. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a framework that supports distributed programming. The framework supports processing of big data with its pre-defined building blocks. It supports machine learning approach with natural language processing in order to perform clustering of given big data (documents). It makes use of Map Reduce programming paradigm in a simulated environment. The framework has provision for generic functionalities that can be reused by cloud users in the real world. Unstructured data is subjected to machine learning in order to obtain intelligence which is used to determine clusters. Recently emerged domain in the information technologies is Big Data which refers to data with characteristics such as volume, velocity and variety. The existing machine learning approaches cannot cope with Big Data. In this paper our framework with underlying algorithm supports scheduling jobs of multiple users concurrently. We built a custom simulator that demonstrates processing of big data in distributed environment. Our empirical results revealed that the performance of proposed framework is high in terms of accuracy when compared with that of ground truth. This research can be extended further to improve the framework to support different kinds of big data besides supporting multiple data mining operations on big data.

REFERENCES

[1] J. K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin, "Parallel coordinate descent for l1-regularized loss minimization," in Proc. Int. Conf. Mach. Learn., 2011, pp. 321–328.

[2] A. Agarwal and J. C. Duchi, "Distributed delayed stochastic optimization," in Proc. Adv. Neural Inf. Process. Syst., 2011, pp. 873–881.

[3] M. Zinkevich, J. Langford, and A. J. Smola, "Slow learners are fast," in Proc. Adv. Neural Inf. Process. Syst., 2009, pp. 2331–2339.

[4] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng, "Large scale distributed deep networks," in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1232–1240.

[5] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," J. Mach. Learn. Res., vol. 14, pp. 1303–1347, 2013.

[6] S. A. Williamson, A. Dubey, and E. P. Xing, "Parallel Markov chain Monte Carlo for nonparametric mixture models," in Proc.Int. Conf. Mach. Learn., 2013, pp. 98–106.

[7] T. White, Hadoop: The Definitive Guide. Sebastopol, CA, USA: O'Reilly Media, 2012

[8] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein, "Distributed GraphLab: A framework for machine learning and data mining in the cloud," in Proc. VLDB Endowment, vol. 5, pp. 716–727, 2012.

[9] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in Proc. 2nd USENIX Conf. Hot Topics Cloud Comput., 2010, p. 10.

[10] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su, "Scaling distributed machine learning with the parameter server," in Proc. 11th USENIX Conf. Operating Syst. Des. Implementation, 2014, pp. 583–598.

[11] R. Power and J. Li, "Piccolo: Building fast, distributed programs with partitioned tables," in Proc. USENIX Conf. Operating Syst. Des. Implementation, article 10, 2010, pp. 1–14.

[12] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski, "Pregel: A system for large-scale graph processing," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2010, pp. 135–146.

[13] STEVE LOHR. (2012). The Age of Big Data. Big Data's Impact in the World, p1-5.

[14] Hsinchun Chen,Roger H. L. Chiang and Veda C. Storey. (2012). BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT. MIS Quarterly. 36 (4), p1165-1188.

[15] Adam Jacobs. (2009). The Pathologies of Big Data. communicati ons of the acm. 52 (8), p1-9.

[16] Min Chen ,Shiwen Mao and Yunhao Liu. (2014). Big Data: A Survey.Springer Science+Business Media, p1-39.

[17] Alexandros Labrinidis and H. V. Jagadish. (2012). Challenges and Opportunities with Big Data. Proceedings of the VLDB Endowment. 5 (12), p1-2.

[18] Tim Kraska. (2013). Finding the Needle in the Big Data Systems Haystack. IEEE Computer Society, p1-3.

[19] Divyakant Agrawal,Sudipto Das and Amr El Abbadi. (2011). Big Data and Cloud Computing: Current State and Future Opportunities. ACM, p1-4.
[20] Chris Snijders, Uwe Matzat and Ulf-Dietrich Reips. (2012). "Big Data": Big Gaps of Knowledge in the Field of Internet Science. International Journal of Internet Science. 7 (1), p1-5.
[21] Herodotos Herodotou, Harold Lim, Gang Luo, Nedyalko Borisov, Liang Dong, Fatma Bilgen Cetin and Shivnath Babu. (2011). Starfish: A Selftuning System for Big Data Analytics. Biennial Conference on Innovative Data Systems Research, p1-12.
[22] Alfredo Cuzzocrea,Il-Yeol Song and Karen C. Davis. (2011). Analytics over Large-Scale Multidimensional Data: The Big Data Revolution.ACM, p1-3.
[23] C.L. Philip Chen and Chun-Yang Zhang. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, p1-34.
[24] Avita Katal,Mohammad Wazid and R H Goudar. (2013). Big Data: Issues, Challenges, Tools and Good Practices. IEEE. . (.), p1-6.
[25] Christian Bizer, Peter Boncz, Michael L. Brodie and Orri Erling. (2011). The Meaningful Use of Big Data: Four Perspectives – Four Challenges.SIGMOD Record. 40 (4), p1-5.

AUTHORS

D.Krishna Madhuri Received her B.Tech degree in computer Science and engineering from Swarnandhra college of Engineering & Technology, Narsapur, Andhra Pradesh in 2009,and M.Tech Degree in Computer Science and Engineering from Sridevi Women's Engineering college Hyderabad, Telangana in 2012. She is currently pursuing her Ph.D degree from Sri Satyasai University, Sehore, Madhya Pradesh. Currently she is working as Assistant Professor in the department of Computer Science and Engineering of Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana India. Herresearch interest includes Data Bases, Data Mining and Big Data.