

HEB

**RESEARCH IN BIG DATA**

CASS

Giriraj Parihar,


Department of Computer Science,

S.S. Jain Subodh P.G. Autonomous College, Jaipur

Email- [parihargiriraj@live.com](mailto:parihargiriraj@live.com)**ABSTRACT:**

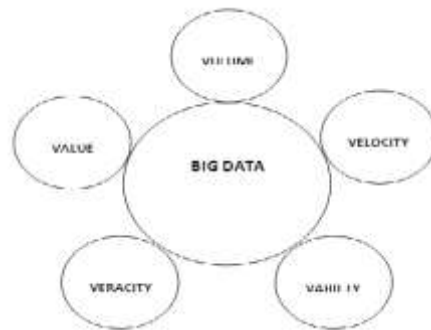
An enormous storehouse of terabytes of information is produced each day from present day frameworks and advanced advances, for example, distributed computing. Enormous information alludes to datasets that are huge, yet in addition high in assortment and speed, which makes them hard to deal with utilizing conventional devices and strategies. Huge information is a term which embodies the improvement and accessibility of information in each of the three configurations like structure, unstructured and semi positions. Structure information is situated in a fixed field of a record or document and it is available in the social information bases and spreadsheets though an unstructured information record incorporates content and sight and sound substance. Investigation of these monstrous information requires a great deal of endeavors at various dimensions to remove learning for basic leadership. So they should almost certainly increase important bits of knowledge from such shifted and quickly evolving information, running from day by day exchanges to client associations and interpersonal organization information. Such esteem can be given utilizing enormous information investigation, which is the use of cutting edge examination systems on huge information. In this manner, huge information investigation is a momentum territory of innovative work. These days, enormous information has turned out to be remarkable and favored research zones in the field of software engineering. Many open research issues are accessible in enormous information and great arrangements additionally been proposed by the scientists despite the fact that there is a requirement for improvement of numerous new strategies and calculations for huge information investigation so as to get ideal arrangements.

**Keyword:** Big data analytics; data mining; analytics; Hadoop; Massive data; Structured data; Unstructured Data; decision making.

Access this Article Online	Quick Response Code: 
Website: <a href="http://heb-nic.in/cass-studies">http://heb-nic.in/cass-studies</a>	
Received on 14/04/2019	
Accepted on 17/04/2019 © HEB All rights reserved	

## INTRODUCTION

Enormous information is related with huge informational collections and the size is over the adaptability of basic database programming devices to catch, store, handle and assess [1][2]. It furnishes transformative leaps forward in numerous fields with gathering of substantial datasets. Huge information investigation is basic for experts, analysts and businessmen to settle on better choices that were already not achieved. Figure 1 clarifies the structure of enormous information which contains five measurements to be specific volume, speed, assortment, esteem and veracity [2][3]. Volume alludes to the tremendous measure of information that are being created each day. Speed characterizes the constant entry of information streams from this valuable data's are gotten. Assortment gives data about the sorts of information, for example, organized, unstructured, semi organized and so on. Esteem is fundamental to get the monetary estimation of various information which fluctuates essentially. The fifth V alludes to veracity that incorporates accessibility and responsibility. Moreover enormous information has upgraded improved through-put, network and figuring rate of advanced gadgets which has affixed the recovery, procedure and creation of the information.



**Structure of Big Data**

Enormous information has three sorts of learning disclosure; they are oddity revelation, class disclosure and affiliation disclosure. Oddity disclosure is utilized to locate another, uncommon one, already unfamiliar and obscure from a billion or trillion articles or occasions [2]. Class disclosure finds new classes of articles and conduct and affiliation revelation is utilized to locate a strange co-happening affiliation. A portion of these extraction strategies for acquiring accommodating data was talked about by Gandomi and Haider [2]. Anyway precise definition for huge information isn't characterized and there is a trust that it is issue explicit. This will help us in acquiring improved basic leadership, knowledge revelation and streamlining while at the same time being creative and savvy.

Data Types	Data Sizes	Characteristics	Tools	Analytical methods	Examples
Small Data	Mega bytes	Hundred-thousand records	Personal computers, excel, R	Simple statistics	Sales records, customer Database for small companies

<b>Large data</b>	Giga bytes, bytes	billions of records - structured data	RDBMS, Data warehouses	Advanced statistics, Data mining, business intelligence	customer Database for big companies
<b>Big data</b>	Giga bytes, bytes	hundreds of millions of records - distributed and unstructured	Cloud, Data centre, NoSQL, Hadoop	Map reduce, Distributed file systems	Customer interaction- social network, mobile, multimedia

**Table 1: Comparative Study on Type of Data**

## NEED OF BIG DATA

The enormous volume of information couldn't be speedily prepared by customary database procedures and instruments and it fundamentally engaged and took care of organized information [1]. At the season of advancement of PCs the measure of information put away in the PCs are less because of its base stockpiling limit. After the innovation of systems administration, the information put away in PCs are expanded in light of the fact that the improved advancements in the equipment parts. Next, the entry of a web makes a blast to store tremendous accumulations of information and it tends to be utilized for different purposes [2]. This circumstance raised worries about the presentation of new research related ideas like information mining, organizing, picture preparing, network figuring, distributed computing and so on are utilized for examining the diverse kinds of information which are utilized in different areas. Numerous new procedures, calculations, ideas and techniques have been proposed by the scientists for breaking down the static informational indexes. In this computerized time, after the improvement of versatile and remote advances gives another stage in which individuals may share their data through online networking locales for example face book, twitter and google+ [3]. In these spots, the information might be arrived persistently and it can't be put away in PC memory on the grounds that the measure of the information is gigantic and it is considered as "Large Data". This circumstance likewise made an issue about how to perform information examination for this dynamic datasets since the current calculations and their answers are not appropriate for dealing with the enormous information. This circumstance has raised worries about the necessity of improvement of new procedures, techniques and calculations [1][2].

The term 'Huge Data' came into view for first time in 1998 of every a Silicon Graphics (SGI) by John Mashey The development of huge information needs to expand the capacity limit and handling power. Every now and again a lot of information (2.5 quintillion) are made through long range interpersonal communication [1]. It is normal that the development of enormous information is evaluated to achieve 25 billion by 2015 [3]. From the point of view of the data and correspondence innovation, huge information is a vigorous impulse to the up and coming age of data innovation ventures [4], which are comprehensively based on the third stage, fundamentally alluding to huge information, distributed

computing, web of things, and social business. By and large, Data distribution centers have been utilized to deal with the expansive dataset. For this situation extricating the exact learning from the accessible huge information is a principal issue. The greater part of the exhibited methodologies in information mining are not generally ready to deal with the vast datasets effectively. The key issue in the examination of huge information is the absence of coordination between database frameworks just as with investigation apparatuses, for example, information mining and factual examination. These difficulties for the most part emerge when we wish to perform information disclosure and portrayal for its useful applications. A crucial issue is the means by which to quantitatively portray the fundamental qualities of huge information. There is a requirement for epistemological ramifications in depicting information upheaval [5]. Moreover, the investigation on unpredictability hypothesis of enormous information will help comprehend fundamental qualities and arrangement of complex examples in huge information, disentangle its portrayal, improves learning deliberation, and guide the structure of processing models and calculations on huge information [4]. Much research was completed by different scientists on enormous information and its patterns [6], [7], [8].

Huge information applications have presented the extensive scale circulation applications which work with vast informational collections. Information examination issue assumes a fundamental job in numerous segments [1]. The current programming for huge information applications like Apache Hadoop and Google's guide decrease structure, in which these applications creates a lot of middle of the road information [2]. There are numerous utilizations of huge information, for example, fabricating, bioinformatics, human services, interpersonal organization, business, science and innovation and savvy urban areas [3]. Huge information gives a foundation to Hadoop in bioinformatics which joins sequencing people to come, expansive scale information investigation and other organic areas. Parallel circulated processing system and distributed computing consolidates with bunches and web interfaces. [1][3].

## **BIG DATA TECHNOLOGIES**

### **1. Column-oriented databases**

In column-oriented database stores information in segments as opposed to columns, which is utilized to packs monstrous information and quick inquiries [3].

Composition less databases

Composition less databases are generally called as NoSQL databases. Database gives a component to capacity and recovery of information that is displayed in methods other than the unthinkable relations utilized in social databases. There are two kinds of database, for example, record stores and key esteem stores that stores and recovers gigantic measure of organized, unstructured and semi organized information [3].

### **2. Hadoop**

Hadoop is a mainstream open source instrument for taking care of enormous information and implemented in MapReduce. It is java-based programming structure which bolsters vast informational indexes in appropriating registering. Hadoop group utilizes an ace/slave structure. Disseminated record framework in hadoop moves information in fast rates. If there should arise an occurrence of some hub disappointment an appropriated document framework enables the framework

to proceed with the typical task. Hadoop has two fundamental sub extends in particular Map Reduce and Hadoop Distributed File System (HDFS) [4].

### **3. Guide Reduce**

This is a programming worldview which permits execution versatility against a large number of servers and server bunches for expansive undertaking. Guide decrease execution comprises of two assignments, for example, map task and diminish task. In the guide task the info dataset is changed over into various key/esteem sets or tuples where as in decreased assignments a few types of yield of guide task is joined to shape a diminished arrangement of tuples.

### **4. HDFS**

Hadoop dispersed record framework is a document framework which broadens all hubs in hadoop groups for information stockpiling. It connects all the document framework together on neighborhood hub to make into an extensive record framework. To beat the hub disappointments HDFS upgrades the security by portraying information over numerous sources [4].

### **5. Hive**

Hive is an information warehousing foundation which is based on hadoop. It has distinctive capacity types, for example, plain content, RC record, Hbase, ORC and so forth. Worked in client characterized capacities are utilized to deal with dates, strings and other information mining devices It is SQL-like Bridge that enables BI application to run inquiries against Hadoop bunches [4].

### **6. Capacity Technologies**

To store tremendous volume of information, productive and successful procedures are required. The fundamental focal point of capacity advances are information pressure and capacity virtualization [5]. HBase is a versatile distributive database which utilizes Hadoop dispersed record framework for capacity. It bolsters section arranged database and structure information [5].

### **7. Chukwa**

Chukwa examination screens huge dispersed framework and it includes required semantics for log accumulations and it utilizes start to finish conveyance show [5].

## **RESEARCH ISSUES IN BIG DATA**

Enormous information has three principal issue for example capacity issues, the board issues and handling these issues shows an enormous arrangement of specialized research issues though capacity issue manage when a nature of information is detonated, every single time it makes new capacity medium. In addition information is being made for the most part in each spot, for instance, internet based life, 12+ Tbytes of tweets are developing each day and regularly re-tweets are 144 for each tweet. The following issue is the executives issues, which are troublesome issue in huge information space. On the off chance that the information is circulated topographically it tends to be overseen and claimed by various substances. Advanced information accumulation is simpler than manual information gathering where computerized information speaks to the procedure for information

accumulation. Information capability centers around missing information or exceptions rather on approving every thing [5]. Henceforth new methodologies are required for information capability and information approval. In handling issue worries about how to process 1K petabyte of information which requires an absolute start to finish preparing time of approximately 635 years. Along these lines, compelling preparing of exabytes of information will require broad parallel handling and new examination calculations so as to give auspicious data. Numerous issues in huge information can be settled by e-science which requires lattice and distributed computing.

## 1 Big Data Classification

Information characterization is the way toward sorting out information into classes for its best and productive use. An all around arranged information grouping framework makes fundamental information simple to discover and recover. There are three essential and these parts of information grouping in particular techniques, areas and varieties. Strategies depicts regular procedures utilized for arrangement models are probabilistic techniques, choice trees, rule-based techniques, occurrence based techniques, bolster vector machine techniques and neural systems [2]. Spaces look at explicit strategies utilized for information areas, for example, sight and sound, content, time-arrangement, organize, discrete succession and questionable information. It likewise covers vast informational indexes and information streams because of the ongoing significance of the enormous information worldview [4]. Varieties in grouping process talks about outfits, uncommon class learning, separate capacity learning, dynamic learning, visual learning, exchange learning, and semi-regulated learning just as assessment parts of classifiers [5].

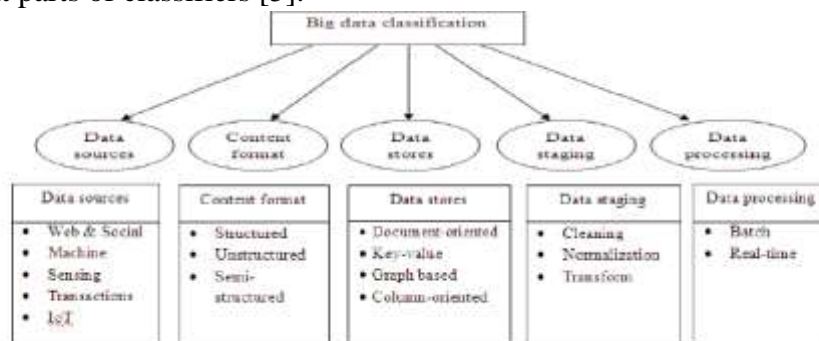


Figure 1. Big Data Classification

## 2. Clusters in Big Data

A gathering of the indistinguishable components firmly together is known as bunching. Information bunching are otherwise called group examination or fragment investigation which arranges an accumulation of n objects into a segment or a chain of command. The primary point of bunching is to arrange information into groups to such an extent that objects are assembled in a similar group when they are "comparative" as per likenesses, characteristics and conduct. The most regularly utilized calculations in bunching are parceling, progressive, lattice based, thickness based, and show based calculations. Dividing calculations is called as the centroid based bunching. Various leveled calculations additionally called as the availability based grouping. Thickness put together bunching is based with respect to the idea of information reachability and information network. Network put together grouping is based with respect to the extent of the matrix rather than the information. Display

put together grouping depends with respect to the likelihood appropriation. Figure 2 speaks to the preparing of information bunching [4],[5],[6].

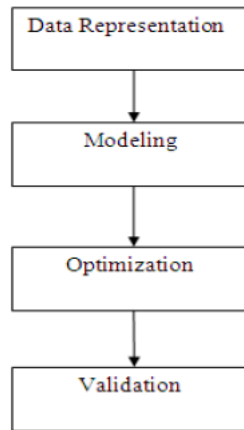


Figure.2. Processes of Data Clustering

algorithm on big data. Big Data are generated at high speed. To guide the selection of a suitable clustering algorithm with respect to the Velocity property shows the criteria and Complexity of algorithm. Many clustering algorithms are available few are listed below. [5][6][7].

- K-means
- Gaussian mixture models
- Kernel K-means
- Spectral Clustering
- Nearest neighbor
- Latent Dirichlet Allocation.

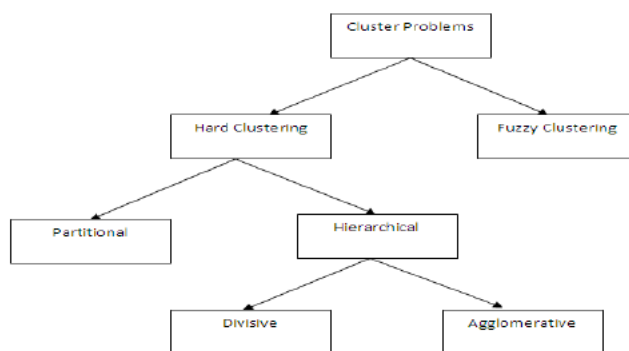


Figure 3. Clustering Algorithm

**3. Association Rules**

Affiliation rules are (assuming/at that point) articulations that assistance to reveal connections between apparently irrelevant information in a value-based database, social database or other data archive. An affiliation rule has two sections, a predecessor (in the event that) and a resulting (at that point) [8]. A forerunner is a thing found in the information. A resulting is a thing that is found in blend with the

precursor. Affiliation rules are made by dissecting information for incessant on the off chance that/at that point examples and utilizing the criteria backing and certainty to recognize the most vital connections. Backing means that how every now and again the things show up in the database. Certainty demonstrates the occasions the in the event that/at that point proclamations have been observed to be valid. In information mining, affiliation rules are helpful for breaking down and foreseeing client conduct [9]. Affiliation rule mining finds the incessant examples, affiliations, connections, or causal structures among sets of things or items in value-based databases, social databases and other data stores. Affiliation rules are utilized for market bin information examination, cross-advertising, index plan, misfortune pioneer investigation, and so forth. A portion of the properties of affiliation rules are the way things or articles are identified with one another and how they will in general gathering together, easy to get (fathomability), give valuable data (utilizability), productive revelation calculations (effectiveness). Distinctive sorts of affiliation rules depend on kinds of qualities took care of for example Boolean affiliation rules and Quantitative affiliation rules. Dimensions of reflection are separated into either single-level affiliation rules or staggered affiliation rules. Measurements of information required into singledimensional affiliation rules and multidimensional affiliation rules [10].

#### **4. BIG DATA VISUALIZATION**

Enormous Data perception is a handling by which numerical information are changed over into important 3-D pictures. It is an introduction of pictorial or graphical arrangement and which relies on visual portrayal, for example, designs, tables, maps and diagrams which sees all the more rapidly and effectively. There are numerous apparatuses in enormous information perception to be specific polymaps, nodebox, flot, preparing, tangle, SAS visual investigation, linkscape, leaflet, crossfilter, openlayer [9]. Perception systems are characterized into three diverse ways (i.e.) in light of the undertaking, in view of the structure of the informational index or dependent on the measurement. Representation can be named whether the given information is spatial or non spatial or whether the showed information to be in 2D or 3D. Visualization parts can be either static or dynamic [10] Visualization is utilized for spatial information and non-spatial information. For speaking to 2D or 3D information additionally different representation instruments are connected. The preparing of information in representation framework can be bunch or intuitive. The bunch preparing is utilized for examination of set of pictures. In information representation connection the client can interface in assortment of ways which incorporates perusing, inspecting, questioning and associative.. Various techniques are accessible in information perception and it depends on kind of information, there are three sorts of information: Univariate, Bivariate and Multivariate. Univariate measures the single quantitative variable, it portrays dispersion and it is spoken to by two techniques they are histogram and pie graph. Bivariate comprises the example pair of two quantitative factors, they are connected with one another. They are spoken to utilizing dissipate plots and line diagram strategies .Multivariate information speaks to multidimensional information and it is spoken to by symbol based strategy, pixel based technique and dynamic parallel organize system.[8][9][10]



## TOOLS FOR BIG DATA PROCESSING

Substantial quantities of instruments are accessible to process enormous information. In this area, we examine some present strategies for dissecting huge information with accentuation on three essential developing devices specifically MapReduce, Apache Spark, and Storm. The vast majority of the accessible instruments focus on bunch handling, stream preparing, and intelligent examination. Most group preparing instruments depend on the Apache Hadoop foundation, for example, Mahout and Dryad. Stream information applications are generally utilized for continuous logical. A few instances of substantial scale gushing stage are Strom and Splunk. The intuitive investigation process enable clients to straightforwardly cooperate progressively for their very own examination and is portrayed in Figure 4.

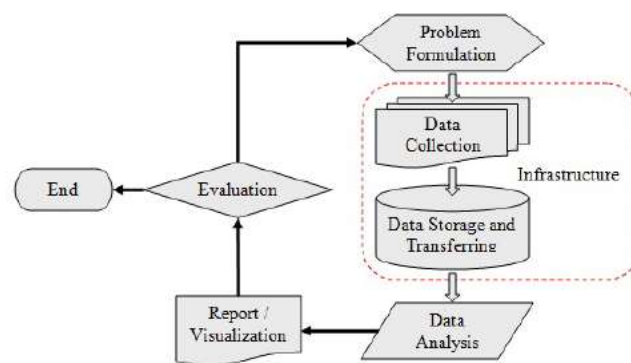


Figure 4: Workflow of Big Data Project

### 1. Apache Hadoop and MapReduce

The most settled programming stage for enormous information examination is Apache Hadoop and Mapreduce. It comprises of Hadoop bit, mapreduce, hadoop dispersed record framework (HDFS) and apache hive and so forth. Guide diminish is a programming model for handling vast datasets depends on separation and overcome strategy. The partition and vanquish strategy is executed in two stages, for example, Map step and Reduce Step. Hadoop chips away at two sorts of hubs, for example, ace hub and specialist hub. The ace hub isolates the contribution to littler sub issues and after that disseminates them to laborer hubs in guide step. From that point the ace hub consolidates the yields for all the subproblems in decrease step. In addition, Hadoop and MapReduce fills in as an amazing programming system for tackling enormous information issues. It is additionally useful in blame tolerant capacity and high throughput information preparing. Predominantly utilized are recorded beneath.

### 2. Apache Mahout

Apache mahout expects to give versatile and business AI procedures for huge scale and clever information examination applications. Center calculations of mahout including bunching, grouping, design mining, relapse, dimensionalty decrease, developmental calculations, and cluster put together community sifting keep running with respect to top of Hadoop stage through guide diminish system. The objective of mahout is to manufacture a dynamic, responsive, various network to encourage discourses on the task and potential use cases. The fundamental target of Apache mahout is to give an

apparatus to alleviating enormous difficulties. The diverse organizations the individuals who have actualized versatile AI calculations are Google, IBM, Amazon, Yahoo, Twitter, and facebook [36].

### 3. Apache Spark

Apache flash is an open source enormous information handling structure worked for speed preparing, and advanced investigation. It is anything but difficult to utilize and was initially created in 2009 in UC Berkeleys AMPLab. It was publicly released in 2010 as an Apache venture. Flash lets you rapidly compose applications in java, scala, or python. Notwithstanding map lessen tasks, it underpins SQL inquiries, gushing information, AI, and chart information preparing. Sparkle keeps running over existing hadoop disseminated document framework (HDFS) foundation to give upgraded and extra usefulness. Sparkle comprises of parts in particular driver program, group administrator and specialist hubs. The driver program fills in as the beginning stage of execution of an application on the sparkle hubs. The group chief apportions the assets and the laborer hubs to do the information preparing as assignments. Every application will have a lot of procedures considered agents that are in charge of executing the undertakings. The real favorable position is that it offers help for sending flash applications in a current hadoop groups. Figure 5 portrays the engineering graph of Apache Spark. The different highlights of Apache Spark are recorded beneath:

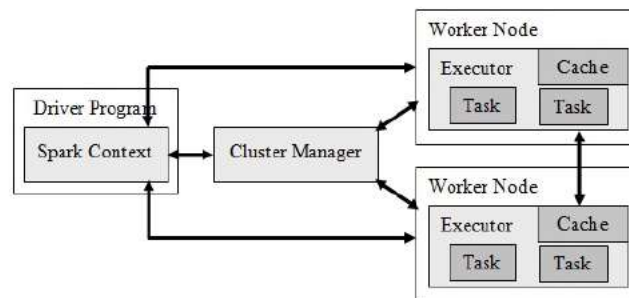


Figure 5: Architecture of Apache Spark

### 4. Dryad

It is another prevalent programming model for executing parallel and appropriated programs for taking care of huge setting bases on dataflow chart. It comprises of a group of figuring hubs, and a client utilize the assets of a PC bunch to run their program in a dispersed manner. To be sure, a dryad client utilize a large number of machines, every one of them with numerous processors or centers. The significant favorable position is that clients don't have to know anything about simultaneous programming. A dryad application runs a computational coordinated chart that is made out of computational vertices and correspondence channels. Subsequently, dryad gives a substantial number of usefulness including creating of occupation chart, planning of the machines for the accessible procedures, change disappointment taking care of in the group, accumulation of execution measurements, imagining the activity, invoking user characterized arrangements and powerfully refreshing the activity diagram in light of these approach choices without knowing the semantics of the vertices [37].

## CONCLUSION

Lately information are produced at an emotional pace. Dissecting these information is trying for a general man. To this end in this paper, we overview the different research issues, difficulties, and devices used to dissect these huge information. From this study, it is comprehended that each enormous information stage has its individual core interest. Some of them are intended for clump preparing though some are great at continuous systematic. Each enormous information stage likewise has explicit usefulness. Distinctive strategies utilized for the investigation incorporate measurable examination, AI, information mining, smart investigation, distributed computing, quantum figuring, and information stream preparing. We believe that in future scientists will give more consideration to these methods to take care of issues of huge information successfully and proficiently.

## REFERENCES

- [1] M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, *International Journal of Application or Innovation in Engineering & Management*, 2(8) (2015), pp.228-232.
- [2] A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, *International Journal of Information Management*, 35(2) (2015), pp.137-144.
- [3] C. Lynch, Big data: How do your data grow?, *Nature*, 455 (2008), pp.28-29.
- [4] X. Jin, B. W.Wah, X. Cheng and Y. Wang, Significance and challenges of big data research, *Big Data Research*, 2(2) (2015), pp.59-64.
- [5] R. Kitchin, Big Data, new epistemologies and paradigm shifts, *Big Data Society*, 1(1) (2014), pp.1-12.
- [6] C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, *Information Sciences*, 275 (2014), pp.314-347.
- [7] K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, *Journal of Parallel and Distributed Computing*, 74(7) (2014), pp.2561-2573.
- [8] S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, On the use of mapreduce for imbalanced big data using random forest, *Information Sciences*, 285 (2014), pp.112-137.
- [9] MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, Health big data analytics: current perspectives, challenges and potential solutions, *International Journal of Big Data Intelligence*, 1 (2014), pp.114-126.
- [10] R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, A look at challenges and opportunities of big data analytics in healthcare, *IEEE International Conference on Big Data*, 2013, pp.17-22.
- [11] Z. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining, *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1997.
- [12] T. K. Das and P. M. Kumar, Big data analytics: A framework for unstructured data analysis, *International Journal of Engineering and Technology*, 5(1) (2013), pp.153-156.
- [13] T. K. Das, D. P. Acharjya and M. R. Patra, Opinion mining about a product by analyzing public tweets in twitter, *International Conference on Computer Communication and Informatics*, 2014.
- [14] L. A. Zadeh, Fuzzy sets, *Information and Control*, 8 (1965), pp.338- 353.
- [15] Z. Pawlak, Rough sets, *International Journal of Computer Information Science*, 11 (1982), pp.341-356.
- [16] D. Molodtsov, Soft set theory first results, *Computers and Mathematics with Applications*, 37(4/5) (1999), pp.19-31.

- [17] J. F.Peters, Near sets. General theory about nearness of objects, *Applied Mathematical Sciences*, 1(53) (2007), pp.2609-2629.
- [18] R. Wille, Formal concept analysis as mathematical theory of concept and concept hierarchies, *Lecture Notes in Artificial Intelligence*, 3626 (2005), pp.1-33.
- [19] I. T.Jolliffe, *Principal Component Analysis*, Springer, New York, 2002.
- [20] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and K. Taha, Efficient machine learning for big data: A review, *Big Data Research*, 2(3) (2015), pp.87-93.
- [21] Changwon. Y, Luis. Ramirez and Juan. Liuzzi, Big data analysis using modern statistical and machine learning methods in medicine, *International Neurourology Journal*, 18 (2014), pp.50-57.
- [22] P. Singh and B. Suri, Quality assessment of data using statistical and machine learning methods. L. C.Jain, H. S.Behera, J. K.Mandal and D. P.Mohapatra (eds.), *Computational Intelligence in Data Mining*, 2 (2014), pp. 89-97.
- [23] A. Jacobs, The pathologies of big data, *Communications of the ACM*, 52(8) (2009), pp.36-44.
- [24] H. Zhu, Z. Xu and Y. Huang, Research on the security technology of big data information, *International Conference on Information Technology and Management Innovation*, 2015, pp.1041-1044.
- [25] Z. Hongjun, H. Wenning, H. Dengchao and M. Yuxing, Survey of research on information security in big data, *Congresso da sociedade Brasileira de Computacao*, 2014, pp.1-6.
- [26] I. Merelli, H. Perez-sanchez, S. Gesing and D. D.Agostino, Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives, *BioMed Research International*, 2014, (2014), pp.1-13.
- [27] N. Mishra, C. Lin and H. Chang, A cognitive adopted framework for iot big data management and knowledge discovery prospective, *International Journal of Distributed Sensor Networks*, 2015, (2015), pp. 1-13
- [28] X. Y.Chen and Z. G.Jin, Research on key technology and applications for internet of things, *Physics Procedia*, 33, (2012), pp. 561-566.
- [29] M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto and R. Buyya, Big data computing and clouds: Trends and future directions, *Journal of Parallel and Distributed Computing*, 79 (2015), pp.3-15.
- [30] I. A. T. Hashem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani and S. Ullah Khan, The rise of big data on cloud computing: Review and open research issues, *Information Systems*, 47 (2014), pp. 98-115.
- [31] L. Wang and J. Shen, Bioinspired cost-effective access to big data, *International Symposium for Next Generation Infrastructure*, 2013, pp.1- 7.
- [32] C. Shi, Y. Shi, Q. Qin and R. Bai Swarm intelligence in big data analytics, H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise, B. Li and X. Yao (eds.), *Intelligent Data Engineering and Automated Learning*, 2013, pp.417-426.
- [33] M. A. Nielsen and I. L.Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, New York, USA 2000.
- [34] M. Herland, T. M. Khoshgoftaar and R. Wald, A review of data mining using big data in health informatics, *Journal of Big Data*, 1(2) (2014), pp. 1-35.
- [35] T. Huang, L. Lan, X. Fang, P. An, J. Min and F. Wang Promises and challenges of big data computing in health sciences, *Big Data Research*, 2(1) (2015), pp. 2-11.

- [36] G. Ingersoll, Introducing apache mahout: Scalable, commercial friendly machine learning for building intelligent applications, White Paper, IBM Developer Works, (2009), pp. 1-18.
- [37] H. Li, G. Fox and J. Qiu, Performance model for parallel matrix multiplication with dryad: Dataflow graph runtime, Second International Conference on Cloud and Green Computing, 2012, pp.675-683.