ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 6, Issue 1, pp: (50-59), Month: January - March 2018, Available at: www.researchpublish.com

Review of Big Data Analysis Techniques

Balraj[1], Rosy[2]

[1]M.Tech Scholar, Lovely Professional University, Punjab, India
[2]Assistant Professor, Lovely Professional University, Punjab, India

Abstract: **The Big data is the type of data which is non-relational and very large in size. The big data is very difficult to analyze due to its non-relation properties. The big data can be handled can handles in the HDFC file system which can divide the data into certain number of chucks and each chuck is assigned to different processes. In the past years, various techniques have been proposed for the data analysis which is reviewed in this paper in terms of various parameters.**

Keywords: **HDFC, Big Data, Utilities, Map Reduce.**

I. INTRODUCTION

The term “big data” is relatively new in IT and business. The Big data is a term used where the large volume of data is difficult to process, store and analyze by using traditional existing database technologies. As the nature of big data is indistinct so, there is need to involves considerable processes to identify and translate the data into new insights. There are number of definitions of big data some researchers also define big data as a large volume of scientific data for visualization. Other researchers define big data as “the amount of data just beyond technology's capability to store, manage, and process efficiently.”

Big Data is one of the major challenges of statistical science and a lot of recent references start to think about the numerous consequences of this new context from the algorithmic viewpoint and for the theoretical implications of this new framework. Big Data always involve massive data. They also often include data streams and data heterogeneity.

Fig: 1.1 Big Data

Here are some examples of Big Data applications:

- **Smart Grid case:** it is crucial to monitor and manage the smart grids operations and the national electronic power consumption in real time. To achieve the objective above mentioned there is need to make multiple connections among smart meters, sensors, control centres’ and other infrastructures. To detect the abnormal behaviours of the connected devices and to identify at-risk transformers Big Data analytics need to use. With the help of Big data Grid Utilities can choose the best treatment or action. The real-time analysis of the generated Big Data allow to model incident scenarios.

E-health: To personalize health users are already using services connected health platforms (e.g., CISCO solution). Big Data is generated from different heterogeneous sources (e.g., laboratory and clinical data, patients symptoms uploaded from distant sensors, hospitals operations, and pharmaceutical data). There are number of beneficial applications for using advanced analysis of medical data sets. It enables to personalize

- Health services (e.g., doctors can monitor online patients symptoms in order to adjust prescription); to adapt public health plans according to population symptoms, disease evolution and other parameters. To decrease the cost expenditure and optimize the operations of hospital the big data has been used.

- **Internet of Things (IoT):** IoT represents one of the main markets of big data applications. Because of the high variety of objects, the applications of IoT are continuously evolving. Nowadays, there are various Big Data applications supporting for logistic enterprises. With the help of sensors, wireless adapters and GPS it becomes possible to detect the position of vehicles. Thus, such data driven applications enable companies not only to supervise and manage employees but also to optimize delivery routes.

- **Public utilities:** In complex water supply network the sensors have been placed in pipelines to monitor the flow of water. It is reported in the Press that Bangalore Water Supply and Sewage Board is implementing a real-time monitoring system to detect leakages, illegal connections and remotely control valves to ensure equitable supply of water to different areas of the city. It helps to reduce the need for valve operators and to timely identifying and fixing water pipes that are leaking.

- **Transportation and logistics:** The RFID (Radiofrequency Identification) and GPS have been used by many public road transport companies to track buses and explore interesting data to improve their services. We are able to optimize bus routes and the frequency of trips by collecting data about the number of passengers using the buses in different routes. Various real-time systems has been implemented not only to provide

- Passengers with recommendations but also to offer valuable information on when to expect the next bus which will take him to the desired destination. By predicting the demand about public or private networks the by using mining of Big Data helps in improving the travelling business. Making predictions from such data is a complicated issue because it depends on several factors such as weekends, festivals, night train, starting or intermediate station. By using the machine learning algorithms, it is possible to mine and apply advanced analytics on past and new big data collection. In fact advanced analytics can ensure high accuracy of results regarding many issues.

- **Political services and government monitoring:** Many governments such as India and United States are mining data to monitor political trends and analyze population sentiments. There are many applications that combine many data sources: social network communications, personal interviews, and voter compositions. Such systems enable also to detect local issues in addition to national issues. Furthermore, governments may use Big Data systems to optimize the use of valuable resources and utilities. For instance, sensors can be placed in the pipelines of water supply chains to monitor water flow in large networks. So it is possible for many countries to rely on real-time monitoring system to detect leakages, illegal connections and remotely control valves to ensure equitable supply of water to different areas of the city.

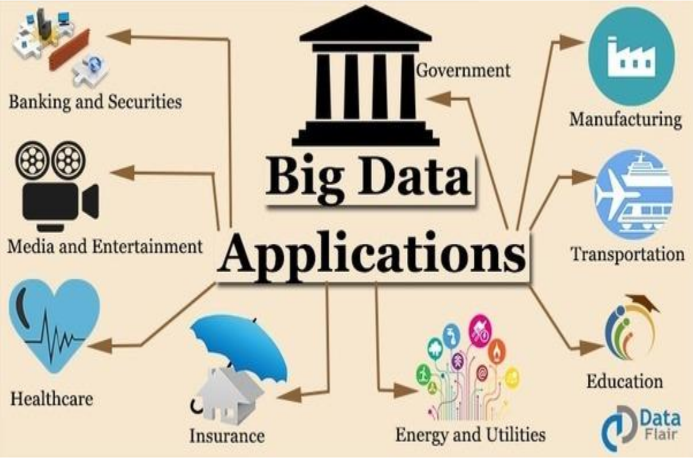

Fig: 1.2 Applications of Big Data

II. ISSUES OF BIG DATA

There are numerous issues arising within this technology some of which are listed below:

- **Storage and Transport Issues:** Each time when new storage medium has been invented the quantity of data has exploded. Now there is no storage area is left due to the increase in social media. Moreover, data is being created by everyone and everything (e.g., devices, etc) by professionals such as scientist, journalists, writers, etc. Current disk technology limits are about 4 terabytes per disk. So, 1 Exabyte would require 25,000 disks. Even if an Exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. Current communication networks has been overwhelm due to access to that data. Assuming that a 1 gigabyte per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes. Thus, transferring an Exabyte would take about 2800 hours, if we assume that a sustained transfer could be maintained. It would take longer to transmit the data from a collection or storage point to a processing point than it would to actually process it.

- **Management Issues:** Management is one of the most difficult problems to address with big data. This problem first surfaced a decade ago in the UK eScience initiatives where data was distributed geographically and “owned” and “managed” by multiple entities. Resolving issues of access, metadata, utilization, updating, governance, and reference (in publications) have proven to be major stumbling blocks. Collection of digital data is easier than the collection of data by manual methods. Due to increase in digital data representation will reduce the need of data collection methodologies. Data is often very fine-grained such as click stream or metering data. Given the volume, it is impractical to validate every data item: new approaches to data qualification and validation are needed. The sources of this data are varied both temporally and spatially, by format, and by method of collection. Individuals contribute digital data in mediums comfortable to them: documents, drawings, pictures, sound and video recordings, models, software behaviours, user interface designs, etc – with or without adequate metadata describing what, when, where, who, why and how it was collected and its provenance. Yet, all this data is readily available for inspection and analysis.

- **Privacy**: In big data privacy and security is an important issue. The Big data security model gets disabled in case of event of complex applications. However, in its absence, data can always be compromised easily. Privacy is the privilege to have some control over how the personal information is collected and used. In case of information privacy it define the capacity of an individual or group to stop information about them from becoming known to people other than the people whom that information is need to send. The identification of personal information during transmission over the Internet is one of the serious issues of user privacy issue.

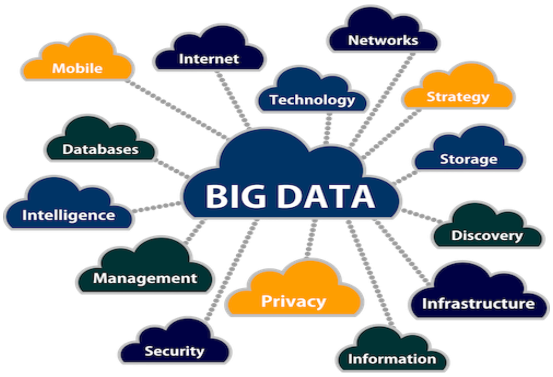

Fig: 1.3 Issues of Big Data

III. TECHNIQUES OF BIG DATA ANALYSIS

Predictive analytics software and hardware solutions which firm determine, assess, optimize, deploy and predictive models to analysis big data source to improve mitigation risk and performance. NoSQl, databases graph database key-value document and. Search and knowledge discovery tools which extract the data from structure and unstructured data from multiple data sources. Stream analytics software can filter, aggregate enrich and analyze high throughput of data from multiple disparate live data sources and any data formats. Data integration tools for data orchestration across solutions such as Amazon Elastic MapReduce (EMR), Apache Hive, Apache Pig, Apache Spark, MapReduce, Couch base, Hadoop, MongoDb.

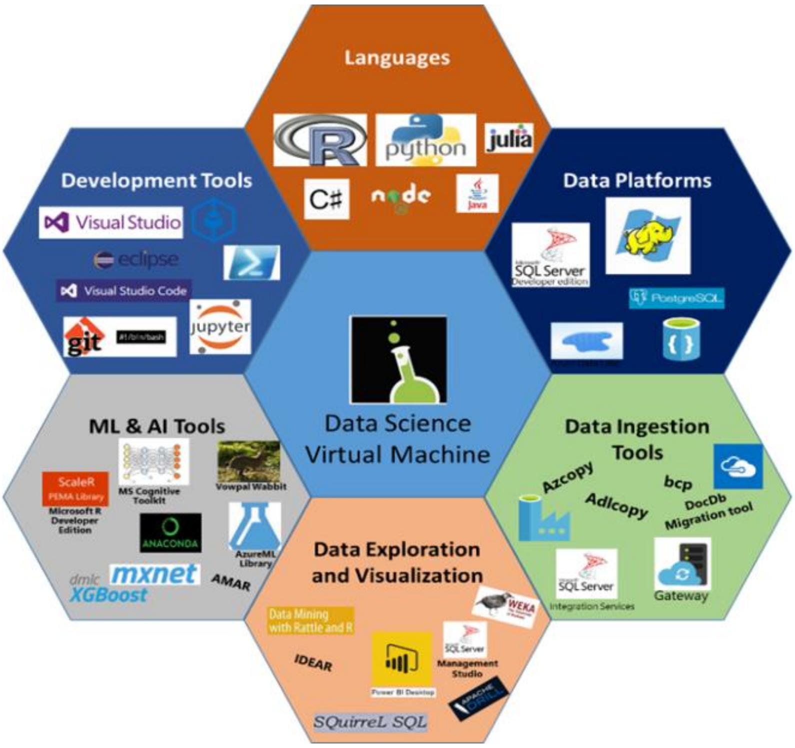

Fig: 1.3 Techniques of Big Data Analysis

IV. LITERATURE REVIEW

ZakiaAsad, et.al, (2015), have concluded that pressure is increased on the data centres' network due to movement of massive volume of data in cloud. So, in this paper the authors have used the mixing technique, spate coding along with software defined network control to propose a new scheme to dynamically reduce the volume of communication. For this purpose they have introduced a novel spate coding algorithm which helps in achieving the real world use cases for networks of data centers. The results are compared in terms of communication volume which is up to 62% better than existing schemes, in terms of good put it is improved by 76%, disk utilization is by 38% and in terms of number of bits that can be transmitted per joule of energy is up to 200%. The results show that the proposed scheme is advantageous from the existing techniques in terms of different parameters [1].

ZakiaAsad, et.al, (2015), have proposed a network coding technique CodHoop employ by system for the same purpose mentioned in previous paper. In this paper, authors have used a network middle box service and specifically index coding for controlling the dynamically reduction in communication volume. A result shows that the proposed scheme is in average 31% better than use case translates depending vanilla Hadoop implementation. This shows that there is 31% less utilization of equipment energy in Hadoop scheme and in proposed scheme 31% jobs can run simultaneously or can say job completion time is reduced by 31%. In this case the given memory has larger bandwidth by which authors are able to process closer to link rate. Even in the worst case this coder has 809 Mbps of throughput on a 1 Gbps link [2].

Xuelian Lin, et.al, (2012), have recommended that for job analysis and optimization of MapReduce the accurate performance model is required. The numbers of steps are needed to be performed in case MapReduce that make it a challenging task. In MapReduce the number of steps is directly proportional to complexity, with increase in number of steps complexity increased at steady rate. To calculate the cost of each item they have decomposed the major cost items and make a new cost model based on vector, equation. The result of model is verified on a several clusters of Hadoop and it shows the effectiveness of proposed model. By improving the proposed scheme the results can be improve in terms of serialization. In case of resource contentions in the cluster, proposed model will not be able to accurately predict the execution time of task [3].

Chang Liu, et.al, (2013), have recommended that for data intensive computation in application of big data, a low cost and high efficiency can be achieved by an environment of cloud computing. In this paper, the authors have proposed a Cloud Background Hierarchical Key Exchange (CBHKE) novel hierarchical scheme. This key exchange scheme will help in achieving the secure and efficient scheduling for cloud computing environment. The experimental and theoretical results of proposed scheme Cloud Background Hierarchical Key Exchange (CCBKE) and Internet Key Exchange (IKE) is superior in terms of efficiency. The proposed scheme CBHKE key exchange scheme help in improving the efficiency but at the same time they become slow in case of large datasets [4].

VidushiVashishth, et.al, (2017), have recommended that with the development in cloud and Internet of Thing (IoT) integration will continuously generate a stream of sensor data. Because of the above mentioned reason number of researchers has started working on the integration of Cloud with Big Data. In this paper, the authors have proposed a predictive scheme for task scheduling on the cloud in case of high velocity (Big Data). The results of proposed algorithm are compared with existing traditional algorithms and it shows that the proposed scheme is 10 times faster than traditional algorithm. In case of small datasets acceptable accuracy can be achieved by appropriate choice of classifiers. By using the classifiers for allocation tasks, researchers are able to achieve the load balancing allocation [5].

E. Goldin, et.al, (2017), have aimed to make a novel infrastructure of cloud computing for the Big Data analytics. Current innovations in the field of Process Analyzer Techniques (PAT), big data and wireless technologies have created a new environment. In this environment with the help of different techniques almost all stages of the industrial process can be recorded and utilized for safety, real time optimization. The sensors of Big Data continuously record a data which require a huge investment in hardware and software. In this paper, authors have presented pilot cloud based architecture for data driven modeling applications. In process control field pilot based architecture will help in getting the optimal control configuration. As it was presented, these developments have been carried in close relationship with the process industry. They also overlay a way for a generalized application of the cloud based approaches, towards the future of Industry 4.0 [6].

Chu-Hsing Lin, et.al (2017) proposed in this paper a study in which the data mining was performed on the land price data of past ten years of Taichung City. The clustering algorithms were utilized here in order to perform data extraction techniques. The K-means and Fuzzy C-means clustering algorithms were executed in Hadoop cloud and a stand-alone PC in order to analyze their respective performances. As per the achieved results it was seen that in a cloud that included 9 compute nodes, an acceleration of around 3.5 times was achieved. Thus, it was concluded that in order to solve insufficient memory related problems within the big data applications, the Hadoop cloud with R provided better results. With the execution of proposed method the computation time was minimized to great extent. Also, with the utilization of adequate number of computing nodes, the issue related to insufficient memory was resolved [7].

Ahmed S. Kaseb, et.al (2017) presented in this paper that there are innumerable applications that use network cameras in order to visualize real time data within different environments. However, there is a need of adequate number of resources in order to analyze such large amount of data which is being generated regularly. A cloud resource manager is proposed in this paper that helps in solving all such problems. A heuristic algorithm is utilized in order to formulate the resource allocation issue. The allocated resources are monitored with the help of proposed manager. Any requirement of a new resource within the application is fulfilled and in case there are any unused resources present, they are removed from the application which results in minimizing the overall cost of the system. Experiments were conducted in this paper to analyze the performance of proposed system. As per the results achieved it was seen that the proposed method minimized the overall cost of the system by up to 60% [8].

Ming-ShenJian, et.al (2017) presented in this paper that there are numerous sites available on Internet these days which help in selecting an appropriate location for spending your vacations as per your comforts. However, with so many choices available, it becomes difficult to select an appropriate destination. A solution to this problem was proposed in this paper which combined cloud computing with big data. The emotions of sentences present on web helped in generating the data which could be given as input to Hadoop in order to perform distributed computing. In order to sort all the data, the K-means algorithm was used here which also helped in updating the database on daily basis. The best choice for the user was selected with the help of intelligent learning mechanism. With the help of this generated system, the travelers were easily able to select perfect destination as per their requirements and choices [9].

RezvanPakdel, et.al (2016) proposed in this paper a mechanism which can handle all types of unstructured data present within the medical applications. Here, both image and textual data needs to be handled in a proper manner. The proposed method needs to be designed in such a manner that it is very general and highly efficient so that the all different types of data can be analyzed easily. As the solution provided here is cloud-based, there is a dynamic improvement in the efficiency which relies on the real-time performance of the computing nodes. Various experiments are conducted to analyze the performance of proposed system. As per the results achieved, it is seen that a scalable solution is provided through this framework. The analysis performance can be enhanced to greater extents in case when there are larger datasets are available. With the help of proposed framework, all types of unstructured data can be processed easily [10].

JiangfanPeng, et.al (2016) presented in this paper [11], that there is a need to attain the space information from the tunnel simulation scenario, in which the data is to be analyzed and compared to provide reliability and quality. Two technology solutions are designed in this paper as per the engineering application of tunnel measurement which utilizes the 3-d laser scanner individually with separate distance measurement principle. In order to detect the target recognition and measure the accuracy, numerous styles identify the cloud of target in both types of 3-d laser scanner. The factors that are related to the design of the technical solution are tested before the implementation process. The accuracy and log size of the two station 3-D laser scanner is tested along with the impact of incident angle on accuracy as well as technology of the systems. As per the results achieved after conducting experiments, it is seen that the error is very less and the precision and reliability have enhanced with the application of proposed technique.

Peter Brezany, et.al (2017) presented in this paper [12], that the evolution of cloud computing towards the Dew computing is presented in this paper which will help in providing various advancements in the scientific computational productivity with the usage of automation. There are various advancements being made recently for maximizing the productivity of the applications related to big data scenarios. Various measures have been proposed to generate automated data science platforms. However, most of the platforms generated fall into the category of business and engineering application areas. The automatic data analysis which is generated by Cloud-Dew computing is presented in this paper. The two application domains namely breath das analysis and brain damage restoration are focused upon in this paper. A novel Dew-enabled balance disorder rehabilitation approach was utilized for presenting various guidelines in order to provide improvements in these techniques. On the basis of various experiments conducted and comparisons made it was seen that various enhancements in terms of accuracy were achieved with the application of this proposed approach.

Mohammad Hossein Ghahramani, et.al (2017) presented in this paper [13], that there is a huge growth in the percentage of processed heterogeneous data with the increase in amount of data being generated each day. The popularity of mobile phones for example is growing on huge rate due to the presence of sensors and the cost effectiveness they include. The collection of contextual data which can further be utilized in engineering and business domains is done in a very easy manner due to such technologies. Amongst the various challenges the researchers are facing related to this technology, the amount of data being generated and the need to analyze this information closer to real-time are gaining attention these days. From the academia, industry and government applications these days, the demand of big data has been arising lately. New technologies are to be presented such that the huge amount of data can not only be processed and analyzed but also ingested quickly at an easy location. A dynamic data analysis framework is proposed in this paper which explores and analyzes the mobile phone data being generated. An interactive exploratory spatial data analysis algorithm is presented in this paper once the data of cell phone communication records is processed. A neighbourhood function is defined here using the nearest neighbour function. The frequency of calls at each cell tower is also analyzed. On the basis of various calculations made, the huge amount of data is processed and stored in efficient manner within less time duration.

Giuseppe Agapito, et.al (2017) [14], Both, Parallel Bioinformatics Algorithms and Cloud-based Healthcare and Biomedicine Services and Systems are reviewed in this paper which is a part of the parallel computing and cloud computing in life sciences. These methods have been utilized within the parallel preprocessing and statistical and data mining analysis of omics data as well as large scale applications respectively. There are various issues that arise when such platforms are utilized in order to store and analyze the health data which are also presented in this paper. The major focus here is made on preserving the security and privacy of the records of patients. Further, the study is proposed related to the parallel and distributed modeling and simulation within the fields of medicine and biology. High performance methods were reported to model, simulate and design these cases present in biological and clinical applications. In order to verify the speed-up of the proposed mechanism, the algorithms and applications were tested and validated with the datasets of real clinics. The performances of these techniques were measured within the parallel computation environments which showed that the proposed technique provided better results.

Albino Altomare, et.al (2017) presented in this paper [15], that the minimization of power consumption of cloud data centers is a major concern within the consolidation of virtual machines. Thus, various studies have been presented within this area. Along with the satisfaction of Service Level Agreement made by the users, it is the objective of consolidation to allocate the virtual machines on minimum number of physical server possible. On the basis of forecast of the virtual machine resource that is required, the effectiveness of the consolidation strategy can change. In order to develop intelligent consolidation policies, the data-driven predictive models are exploited. The various consolidation techniques of virtual machines in cloud systems that are driven by the predictive data mining models are compared in this paper. In order to allocate the requirements present on the present servers, the migrations of future computational requirements of virtual machines are made. There is huge improvement seen within the results in terms of energy saving and most efficient consolidation techniques as per the submission result achieved.

V. TABLE OF COMPARISON

Table 1: Literature Review Comparison

Author	Year	Description	Outcome
1.ZakiaAsad, Mohammad Asad RehmanChaudhry, David Malone	2015	The Authors have concluded that pressure is increased on the data centers network due to movement of massive volume of data in cloud. So, in this paper the authors have used the mixing technique, spate coding along with software defined network control to propose a new scheme to dynamically reduce the volume of communication.	The outcomes compared in terms of communication volume which is up to 62% better than existing schemes, it is improved by 76%, and disk utilization is by 38% and in terms of number of bits that can be transmitted per joule of energy is up to 200%. The results shows that proposed scheme is advantageous from the existing techniques in terms of different parameters
2. ZakiaAsad, M. AsadRehmanChaudhry, D. Malone	2015	Authors have proposed a network coding technique CodHoop employ by system for the same purpose mentioned in previous paper. In this paper, authors have used a network middle box service and specifically index coding for controlling the dynamically reduction in communication volume. Further they have presented the motivating use case for this class of applications and used Hadoop as a representative	A result shows that the proposed scheme is in average 31% better than use case translates depending vanilla Hadoop implementation. This shows that there is 31% less utilization of equipment energy in Hadoop scheme and in proposed scheme 31% jobs can run simultaneously or can say job completion time is reduced by 31%. Even in the worst case this coder has 809 Mbps of throughput on a 1 Gbps
3. Xuelian Lin, ZideMeng, ChuanXu, Meng Wang	2015	They have recommended that for job analysis and optimization of MapReduce the accurate performance model is required. The numbers of steps are needed to be performing in case MapReduce that make it a challenging task. In this paper to measure the MapReduce task complexity, authors have used a new concept which helps in analyzing the detail composition	In this proposed model, authors have not considered the combine operation and serialization cost. By improving the proposed scheme the results can be improve in terms of serialization. In case of resource contentions in the cluster, proposed model will not be able to accurately predict the execution time of task
4. Chang Liu, Xuyun Zhang, Chengfei Liu, Yun Yang, Rajiv Ranjan, DimitriosGeorgakopoulos, Jinjun Chen	2013	They have recommended that for data intensive computation in application of big data, a low cost and high efficiency can be achieved by an environment of cloud computing. In this paper, the authors have proposed a Cloud Background Hierarchical Key Exchange (CBHKE) novel hierarchical scheme. This key exchange scheme will help in achieving the secure and efficient scheduling for cloud computing environment. They have	The experimental and theoretical results of proposed scheme Cloud Background Hierarchical Key Exchange (CCBKE) and Internet Key Exchange (IKE) is superior in terms of efficiency. The proposed scheme CBHKE key exchange scheme help in improving the efficiency but at the same time they become slow in case of large datasets

		designed a layer by layer iterative key exchange strategy to achieve a more efficient Authentication Key Exchange (AKE)	
5. VidushiVashishth, AnshumanChhabra, ApoorviSood	2017	They have recommended that with the development in cloud and Internet of Thing (IoT) integration will continuously generate a stream of sensor data. Because of the above mentioned reason number of researchers has started working on the integration of Cloud with Big Data. In this paper, the authors have proposed a predictive scheme for task scheduling on the cloud in case of high velocity (Big Data).	The results of proposed algorithm are compared with existing traditional algorithms and it shows that the proposed scheme is 10 times faster than traditional algorithm. Due to volume of Big Data, fast processing is required which is achieved by using a proposed scheme.
6. E. Goldin, D. Feldman, G. Georgoulas, M. Castano, G. Nikolakopoulos	2017	They have aimed to make a novel infrastructure of cloud computing for the Big Data analytics. The sensors of Big Data continuously record a data which require a huge investment in hardware and software. In this paper, authors have presented pilot cloud based architecture for data driven modeling applications	In process control field pilot based architecture will help in getting the optimal control configuration. As it was presented, these developments have been carried in close relationship with the process industry. They also overlay a way for a generalized application of the cloud based approaches, towards the future of Industry 4.0
8. Ahmed S. Kaseb, Anup Mohan, YoungsolKoh, Yung-Hsiang Lu	2017	This paper that there are innumerable applications that use network cameras in order to visualize real time data within different environments. However, there is a need of adequate number of resources in order to analyze such large amount of data which is being generated regularly. A heuristic algorithm is utilized in order to formulate the resource allocation issue	The allocated resources are monitored with the help of proposed manager. Any requirement of a new resource within the application is fulfilled and in case there are any unused resources present, they are removed from the application which results in minimizing the overall cost of the system. Experiments were conducted in this paper to analyze the performance of proposed system. As per the results achieved it was seen that the proposed method minimized the overall cost of the system by up to 60%
9. Ming-ShenJian, Yi-Chi Fang, Yu-Kai Wang, Chih Cheng	2017	This paper that there are numerous sites available on Internet these days which help in selecting an appropriate location for spending your vacations as per your comforts. However, with so many choices available, it becomes difficult to select an appropriate destination. A solution to this problem was proposed in this paper which combined cloud computing with big data. The information present on the web was collected and sorted in a proper manner.	The results achieved were ranked on the basis of the analysis made such that they could be read and understood very easily. The emotions of sentences present on web helped in generating the data which could be given as input to Hadoop in order to perform distributed computing. In order to sort all the data, the K-means algorithm was used here which also helped in updating the database on daily basis. The best choice for the user was selected with the help of intelligent learning mechanism.
10. RezvanPakdel, John Herbert	2016	This paper a mechanism which can handle all types of unstructured data present within the medical applications. Here, both image and textual data needs to be handled in a proper manner. The proposed	The analysis performance can be enhanced to greater extents in case when there are larger datasets are available. With the help of proposed framework, all types of unstructured

		method needs to be designed in such a manner that it is very general and highly efficient so that the all different types of data can be analyzed easily.	data can be processed easily
11. JiangfanPeng, XingwangShen, Ming Guo	2016	In this paper as per the engineering application of tunnel measurement which utilizes the 3-d laser scanner individually with separate distance measurement principle. In order to detect the target recognition and measure the accuracy, numerous styles identify the cloud of target in both types of 3-d laser scanner.	As per the results achieved after conducting experiments, it is seen that the error is very less and the precision and reliability have enhanced with the application of proposed technique
12. Peter Brezany, Thomas Ludeschery and Thomas Feilhauer	2017	In this paper the evolution of cloud computing towards the Dew computing is presented in which will help in providing various advancements in the scientific computational productivity with the usage of automation. There are various advancements being made recently for maximizing the productivity of the applications related to big data scenarios.	On the basis of various experiments conducted and comparisons made it was seen that various enhancements in terms of accuracy were achieved with the application of this proposed approach.
13. Mohammad HosseinGhahramani, MengChu Zhou, and Chi Tin Hon	2017	New technologies are to be presented such that the huge amount of data can not only be processed and analyzed but also ingested quickly at an easy location. A dynamic data analysis framework is proposed in this paper which explores and analyzes the mobile phone data being generated	A neighborhood function is defined here using the nearest neighbor function. The frequency of calls at each cell tower is also analyzed. On the basis of various calculations made, the huge amount of data is processed and stored in efficient manner within less time duration.
14. Giuseppe Agapito, Barbara Calabrese, Pietro H. Guzzi, GionataFragomeni	2017	Both, Parallel Bioinformatics Algorithms and Cloud-based Healthcare and Biomedicine Services and Systems are reviewed in this paper which is a part of the parallel computing and cloud computing in life sciences. These methods have been utilized within the parallel preprocessing and statistical and data mining analysis of omics data as well as large scale applications respectively	The algorithms and applications were tested and validated with the datasets of real clinics. The performances of these techniques were measured within the parallel computation environments which showed that the proposed technique provided better results.
15. Albino Altomare, Eugenio Cesario	2017	In this paper, that the minimization of power consumption of cloud data centers is a major concern within the consolidation of virtual machines. Thus, various studies have been presented within this area. Along with the satisfaction of Service Level Agreement made by the users, it is the objective of consolidation to allocate the virtual machines on minimum number of physical server possible	The various consolidation techniques of virtual machines in cloud systems that are driven by the predictive data mining models are compared in this paper. In order to allocate the requirements present on the present servers, the migrations of future computational requirements of virtual machines are made. There is huge improvement seen within the results in terms of energy saving and most efficient consolidation techniques as per the simulation results achieved

VI. CONCLUSION

In this paper, it is concluded that big data is the type of data which is very large in size and database is very un-relational. The big data can be handled with the HDFC file system. The Big data can be processed with the Map Reduce technique. In this paper, various data analysis techniques of big data are analyzed in terms of various parameters.

REFERENCES

[1] ZakiaAsad, Mohammad AsadRehmanChaudhry, David Malone, “Greener Data Exchange in the Cloud: A Coding Based Optimization for Big Data Processing”, IEEE Journal on Selected Areas in Communications, vol. 5, pp.1-18, 2015.

[2] ZakiaAsad, M. AsadRehmanChaudhry, D. Malone, “Codhoop: A system for optimizing big data processing”, in IEEE InternationalSystems Conference (SysCon), 2015, pp. 295–300.

[3] Xuelian Lin, ZideMeng, ChuanXu, Meng Wang, “A practical performance model for hadoop mapreduce”, in IEEE CLUSTER Workshops, vol.4 pp. 231– 239, 2012.

[4] Chang Liu, Xuyun Zhang, Chengfei Liu, Yun Yang, Rajiv Ranjan, DimitriosGeorgakopoulos, Jinjun Chen, “An Iterative Hierarchical Key Exchange Scheme for Secure Scheduling of Big Data Applications in Cloud Computing”, 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, vol. 4, pp. 9-15, 2013.

[5] VidushiVashishth, AnshumanChhabra, ApoorviSood, “A predictive approach to task scheduling for Big Data in Cloud environments using classification algorithms”, IEEE 2017 7th International Conference on Cloud Computing, Data Science & Engineering – Confluence, vol.7, pp. 1888-192, 2017.

[6] E. Goldin, D. Feldman, G. Georgoulas, M. Castano, G. Nikolakopoulos, “Cloud Computing for Big Data Analytics in the Process Control Industry”, 2017 25th Mediterranean Conference on Control and Automation (MED), vol. 5, pp.1373- 1378, 2017

[7] Chu-Hsing Lin, Jung-Chun Liu, Tsung-Chi Peng, "Performance Evaluation of Cluster Algorithms for Big Data Analysis on Cloud", 2017, IEEE

[8] Ahmed S. Kaseb, Anup Mohan, YoungsolKoh, Yung-Hsiang Lu, "Cloud Resource Management for Analyzing Big Real-Time Visual Data from Network Cameras", 2017, IEEE

[9] Ming-ShenJian, Yi-Chi Fang, Yu-Kai Wang, Chih Cheng, “Big Data Analysis in Hotel Customer Response and Evaluation based on Cloud”, 2017, ICACT

[10] RezvanPakdel, John Herbert, “Scalable Cloud-based Analysis Framework for Medical Big-data”, 2016 IEEE 40th Annual Computer Software and Applications Conference

[11] JiangfanPeng, XingwangShen, Ming Guo, “Research on Processing and Analysing of Point Cloud Data of a variety of Lidar”, 2016 Fourth International Workshop on Earth Observation and Remote Sensing Applications

[12] Peter Brezany, Thomas Ludeschery and Thomas Feilhauer, “Cloud-Dew Computing Support for Automatic Data Analysis in Life Sciences”, MIPRO 2017

[13] MohammadhosseinGhahramani, MengChu Zhou, and Chi Tin Hon, “Analysis of Mobile Phone Data under a Cloud Computing Framework”, 2017, IEEE

[14] Giuseppe Agapito, Barbara Calabrese, Pietro H. Guzzi, GionataFragomeni, “Parallel and Cloud-based Analysis of Omics Data: Modelling and Simulation in Medicine”, 2017 25th Euromicro International Conference on Parallel, Distributed and Network-Based Processing

[15] Albino Altomare, Eugenio Cesario, “A Comparative Analysis of Data-Driven Consolidation Policies for Energy-Efficient Clouds”, 2017 25th Euromicro International Conference on Parallel, Distributed and Network-Based processing.