International Journal on Emerging Technologies **10**(4): 125-130(2019)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

The Rise of "Big Data" in the Field of Cloud Analytics

Saiteja Myla[1], Surya Teja Marella[2], K. Karthikeya[3], Preetham B.[4] and S.K. Hasane Ahammad[5]
[1,3]Department of Electronics and Computer Engineering,
Koneru Lakshmaiah Educational Foundation,Vijayawada (Andhra Pradesh), India.
[2,4]Department of Computer Science Engineering,
Koneru Lakshmaiah Educational Foundation, Vijayawada (Andhra Pradesh), India.
[5]Department of Electronics and Computer Engineering,
Koneru Lakshmaiah Educational Foundation, Vijayawada (Andhra Pradesh), India.

(Corresponding author: S.K. Hasane Ahammad)
(Received 19 August 2019, Revised 16 October 2019, Accepted 02 November 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The huge amount of data burst which occurred with the arrival of economical access to the internet led to the rise of market of Cloud Computing which stores this data. And obtaining results from these data led to the growth of "Big Data" industry which analyses this humongous amount of data and retrieve conclusion using various algorithms, Hadoop as a Big Data platform certainly uses Map Reduce Framework to give an analysis report of Big Data. The term "Big Data" can be defined as Modern Technique to store, capture, and manage data's which are in the scale of peta bytes or larger sized dataset with high-velocity and various structures. To address this massive growth of data or Big Data requires a huge computing space to ensure fruitful results through processing of data, and Cloud Computing is that technology which can perform huge-scale and computation which are very complex in nature. Cloud Analytics does enables organizations to perform better business intelligence, data warehousing and Online Analytical Processing (OLAP). This paper includes characteristics, classification of Big Data applications with implementation through Cloud Computing. Moreover, we will have a look into how to apply Big Data analytics to create accurate measures. In this paper we have considered papers from 2010-2019.

Keywords: Cloud Computing, Cloud Analytics, Big Data, Hadoop, Data Analytics.

I. INTRODUCTION

The term "Big Data" used for explaining some detailed information of massive volume. Those data can be in both structured and unstructured form. This data can't be handled using traditional database methodologies and software technologies.

The growth in volume of data is due to large amount of data being captured from organizations, social media, IOT enabled devices etc. The following graphical analysis will give a clearer view of the BIG DATA market scenario.

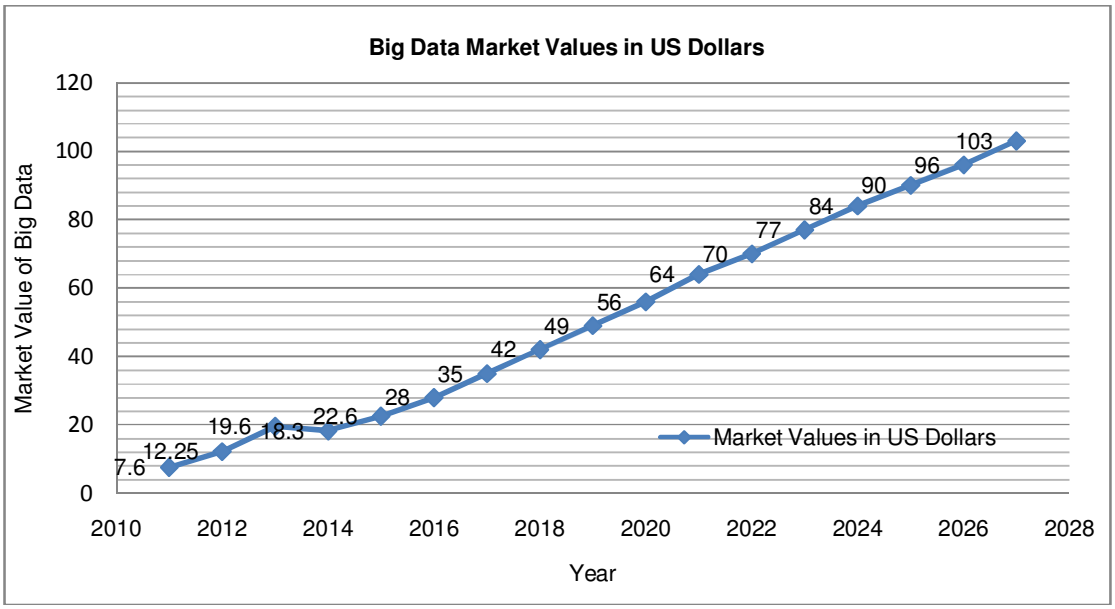

Fig. 1.

Here the graphical growth shows how the industry paced over the few years and how it will go for the next 8-10 years.

The following definition of Big Data is proposed based on the 10 V's namely, Volume, Variety, Velocity, Variety, Variability, Veracity, Validity, Vulnerability, Volatility, Visualization and Value. BIG DATA is a set of techniques and technologies that requires new brand-new forms of integrations to uncover large hidden values from large datasets which are complex, diverse, and of huge scale in nature.

In general, various organizations adopt 3 types of cloud deployment models namely Private Cloud, Public Cloud, Hybrid Cloud [2]. CLOUD COMPUTING is a low cost, economical model that's being used for BIG DATA analytics. The convergence of BIG DATA and Cloud technology make BIG DATA analytics more successful outcomes. Various organizations moved towards dedicated servers for getting better analytical result. H can certainly deliver more accurate and critical results beyond some of the web scale companies such as Yahoo, Spotify [3].

The CLOUD COMPUTING delivers various service models to its user as service-based services in pay you go form. It delivers infrastructure, platform, and software as services. These services are namely, Infrastructure as a Service (IaaS) like Cisco Metapod, Platform as a Service (Paas) like Heroku and Software as a Service (Saas) like Open Shift. [4]. Traditional computing methods are getting out of the plate as the rate of growth of data is too high and with the passing of each day the data is becoming complex in nature, which triggered the BIG DATA application. Large amount of intermediate data is being produced by these softwares like apache H and Google's MAP REDUCE framework. [5]. With the rapid increase of Internet of Thing (IoT) and future internet the concept of smart city evolved in a rapid pace. There is a generation of huge amount of data this data needs to be properly managed and analyzed using integrated Information Communication Technology (ICT) approach. With the arrival of ICT, it brings some significant changes smart cities governance. Processing and integration of cross-disciplinary data is the key for knowledge and intelligence for sustainability, resilience of the governance of the city [6].

Marella *et al.*, (2018) [15] proposes new ways to utilize virtualization technology to improve economic efficiency of data centers.

Gunasekhar *et al.,* (2015) [16] makes a detailed study of insider attacks and their scope in cloud computing.

Bezawada *et al.*, (2018) [17] proposed a new technical analysis of load balancing algorithms in cloud computing.

Marella *et al.,*(2019) [18] proposed a new technology to determine fradulence in credit card transactions.

Myla *et al.,* (2019) [19] proposed a new technology to make decision in carrer counselling.

Bhattacharya *et al.,* (2019) [20] proposed a new technology named cloud federation for cloud task allocation.

Prakash *et al.,* (2019) [21] proposed a new type of braking mechanism.

Gargav *et al.,* (2016) [22] proposed fixed Point Results in Complete Random Convex Metric Spaces.

II. DETAILED STUDY

Big Data Analysis: The definition of the Big Data comes from that 10 V's that will be elaborately discussed in this section.

(a) **Volume:** The term "volume" here is referred to the humungous amount of data that is collected through various sources. The benefit of collecting a massive amount of data is we can find out various hidden patterns or information from these through data analysis. The said approach is called "mobile data challenge" by Nokia. This collection of data leads to various results as predictability of human behavior and human mobility by visualization of data.

(b) **Variety:** Variety refers to the types of different data's which are being collected through various sources such as internet, social media, and sensors and other various devices. All these data can be classified into some of the mainstream types such as images, videos, text, and email. All these collected data are unstructured that is they are in random space.

(c) **Velocity:** It refers to the speed at which data is being transferred.

(d) Value: The most important factor of the 4 V's it refers to the discoveries of hidden data or information from a huge data set collection.

(e) **Variability:** For meaningful analytics to occur it is important to remove outliers and inconsistencies in data.

(f) **Veracity:** If all the aspects of big data are consistent and highly performed only this quantity drops. Veracity is also known as confidence in the data. It generally questions credibility of analytics be it dataset or the method which has been chosen.

(g) **Validity:** It refers to how valid the data is in terms of range of values and the method of collection of data etc.

(h) **Vulnerability:** Any software system is prone to security breaches and big data is no exception. Hence it becomes important to what aspect the data is protected.

(i) **Volatility:** The relevance of data used with regards to time. In simple sense the data used can it be valid in present day scenario.

(j) **Visualization:** The ease of understanding of data it is important that the reports generated from the system are easy to understand and are representation of analytics performed.

(k) **Value:** Lastly it is important that what value does the analytics have what are the inferences we can derive from analysis do they hold any value?

BIG DATA can be classified into various blocks or parts depending on nature or type. The classification is based on the following 5 aspects which are Data Sources, Content Format, Data Stores, Data Staging and Data Process.

A. Cloud Computing

It is one of the fastest growing sectors in the IT industry in today's world. As every organization is seeking for connected storage place and enough good which can be analyzed to obtain results. Cloud service providers have started to provide integrated framework to support parallel data processing [8].

(i) Deployment Model: All the applications which are being used needs to be deployed in the cloud with the variable requirements, now all the deployment model have their own characteristics which are discussed.

(a) **Private Cloud:** One organization operates and maintains the entire cloud infrastructure. People outside the organization don't have any access to the resources in the cloud.

(b) **Public Cloud:** This technology is available for users on pay by subscription basis, it is cost effective technology.

(c) **Hybrid Cloud:** Hybrid cloud can be a merged form of public cloud and private cloud platform where the cloud provider can support some data of the organization and

provide some service in cloud with exchange of some subscription fees. Here consist different types of clouds of different types, and it allows data/ or application to be transfer from one to other cloud [9].

(ii) Service Model: CLOUD COMPUTING provides various models to its consumer in exchange of subscription fees. This provides infrastructure, platform, and software as services. The following models are the models which are being use.

(a) **Software as a Service (Saas):** The Saas model provides user not to install any software, physically in their computer system all the required software is there on the cloud. A Saas provides access to both resource and application to its user.

(b) **Platform as a Service (Paas):** The Paas Platform provides the consumers as a service-based platform. It provides access to the platform that the consumer needs to develop to run their own applications.

(c) **Infrastructure as a Service (Iaas):** The Iaas Platform provides the consumer a computational infrastructure platform to its user. The consumer has total access over storage in cloud and resources which are available in the cloud both hardware and software on a proprietary basis. [10]

B. Big Data Analytics Adoption in Cloud

CLOUD COMPUTING and BIG DATA are co-joined. BIG DATA allows people to compute through distributed queries present in multiple dataset and returns results accurately. While CLOUD COMPUTING provides H, which is a class of distributed data-processing platforms. The data sources are being stored in a distributed fault-tolerance database and proceed through a programming model for large dataset with a parallel distribution algorithm.

Table 1: Comparison of various BIG DATA cloud provider.

	Google Cloud	Microsoft Cloud	Amazon cloud
BIG DATA Storage	Google Cloud Service	Azure	S3
BIG DATA Analytics	Big Query	H on Azure	Elastic MAP REDUCE(H)
MAP REDUCE	App Engine	H on Azure	Elastic MAP REDUCE
Budget	Price per GB 0.010 USD	Price per GB 0.09 USD	Price per Hour 0.010 USD

BIG DATA uses the distributed storage technology service of CLOUD COMPUTING, then the storages installed as hard drives in the computers. In the following table some BIG DATA cloud providers are being listed. [11].

With the increase of the volume of the given data set it also gets more complex in nature; the complexity factor is being solved CLOUD COMPUTING infrastructure which generally addresses this kind of problem. MAP REDUCE has become one of the most important for BIG DATA processing in a cloud environment platform. It provides parallel processing in cluster.

(i) Cloud Analytics and Cloud Analytics Architecture: According to Gartner, an IT research organization "Analytics has been emerged as catch-all term for variety of different Business intelligence (BI) - and application related initiatives". The word 'Analytics' means predicting the future outcome or score by means of analysis of previous or old data. And "Cloud Analytics" is referred as software platform which will be delivered on Software as a Service basis (SaaS).

Cloud Analytics Architecture: The utilization of cloud analytics is to get cloud service for one or all component of an analytical solution this enables the organization to use the social media data, internet data, and all the third-party data to get insights of customer preference and demands. The traditional data base model is no longer in use here in cloud analytics. New database model is needed and used namely NoSQL which can efficiently store and analyze data from cloud.

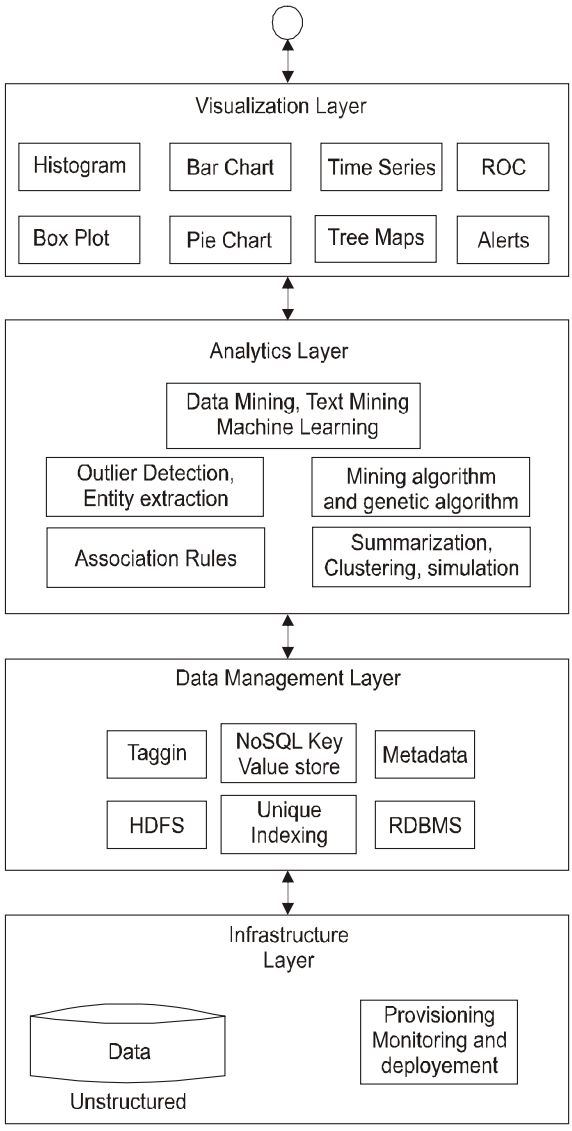

Fig. 2. Cloud Analytics Architecture.

The figure below depicts cloud analytics architecture. There are mainly 4 layers of cloud analytics architecture namely, The Infrastructure layer, The Data storage and Management Layer, The analytics layer, The visualization Layer.

– Infrastructure Layer: The infrastructure layer is the first layer of the cloud analytics. This is also known as the foundation layer. In this layer data is being stored and managed. Data present here are in both structured and unstructured form. Here the data is being collected from various sources like social data, sensor data, streaming data etc.

– Data Storage and Management Layer: The next layer of the cloud analytics architecture is called the data storage and management layer. It is a repository layer here huge amount of raw ordered, raw unordered, and raw semi ordered data is being stored in flat file format. This layer is being managed using H oriented service. Here each data is being attached with a unique ID and this tag is called as Meta data tag. This layer is also known as "data lakes".

– Analytics Layer: The third and most important layer of the cloud analytics architecture is Analytics Layer. This layer mainly holds the business intelligence applications. There are two tools which power this layer viz. pre-analytic tool and analytics tool. The pre-analytic tool chooses and organizes data collected from database, whereas the analytical tool is used to retrieve various patterns and retrieve useful insights from the database. Various Data mining algorithms mainly Supervised Learning where there is a continuous result available are used, algorithms used namely Regression, and Classification modeling is used. Various analytics software is also used viz. Mat Lab, MAP REDUCE, SAS, and R etc.

– Visualization Layer: The last layer of the cloud analytics architecture is the visualization layer. This layer provides user interface and it provides all the visualization and analytics results. Visualization Layer helps the Subject Matter Experts (SMS) to get through all the results without any help of IT personnel [12].

(ii) Big Data Adoption: Various business organizations are moving toward cloud based system as they provide better insights and help them to boost their sales performance in the competitive and ruthless market where the "Survival of the Fittest Statement" stands true each time.

(a) Business Drivers: Nowadays companies are focusing on business analytics trends to get their sales upright. Enterprise uses their private cloud infrastructure to improve risk and it can maintain control over it by analyzing load, cost, and security. The advantage of BIG DATA is being cost saving, revenue growth in an organization. Analytics service used by the companies easily boosts their scalability. Microsoft's solution provides user to analyze the H data from within the excel, with new functioning abilities.

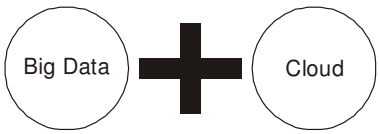

Fig. 3. Interleaving big data and cloud.

Cloud model being interleaved gives us the following benefits as discussed below,

Speed: Private clouds allow its users to do effective BIG DATA analytics and use internal resources from public cloud facilities. This let the users to speed up their infrastructure.

Extract Values: Various organization gives emphasis on the Analytics as a Service (AaaS) model which is being supported by all the three models.

Cost Reduction: We can store huge amount of data in cloud storage economically.

Improved Decision Making: Memory analytics when gets combined with most modern technologies it gives more accurate analyzed results which improves decision making.

Data Valuation: The unstructured or structured data which is collected in raw form is being analyzed and turned into useful data which will give insights of various hidden patterns.

Cloud Security: Cloud service providers are now giving securities at different layers which enable a better security option for the organizations.

Innovation: The convergence of BIG DATA and cloud gives innovative results [13].

(i) Hadoop: There are several software products that would facilitate the BIG DATA analytics and one such software is known as H, in general word it is the solution for most of the BIG DATA Problems. This is the most available Java based framework which supports the processing of huge amount of data in a distributed format. With the help of H, we can analyze an over cluster of servers and applications. The Hadoop model can be upscaled from a one server to many servers, where every single server act as a local storage. When we work with such large scale of data often we face the problem of data redundancy, in H data redundancy is minimized by H D.F.S (HDFS) at application layer itself [14].

(ii) H D.F.S: The backbone of H, it is a D.F.S which is made to be procedure hardware. It's a low cost designed hardware and it's also fault-tolerant. It's effective system when it comes to huge data set its file size is generally in gigabytes and terabytes. The HDFS system is more designed to be work on batch processing model and not in interactive use by user's model. Low Latency user model is more effective in this case. HDFS works on Master-Slave method. It has only one master node data which takes care of the file system namespace and gives regulatory access to files, the master node data is called Name Node. There are numbers of Data Nodes, it usually in the count of one node data per cluster. In internal system the given file is split up into blocks and given the Name Node. The Name Node data works on opening closing, reading file system.

(iii) Calculation of HDFS nodes storage: When we design the new H Cluster, we need some storage place. To estimate the storage amount we need to calculate based on some parameters. As miscalculation of storage place can lead to future problems where shortage of storage leads to non-storage.

The storage is represented by 'H'

Formula:

$$H = \frac{C1 \times R1 \times S1}{(Z)} \times 120\%$$

where

C1= Compression Ratio.

Nature of compression used determines this factor.

C1= 1 is taken when there is no compression

R1 = Replication Factor.

In a production factor it's taken as 3.

S1 = Initial data size that is needed to be sent to cluster.

This can be a mix of both old data and progressive data.

Z = 1–I, where i = Incremental Data Factor.

It's $1/3^{rd}$ $/1/4^{th}$; its H's intermediate working storage space.

120 % = It's 1.2X more than the total size.

For Example, if the cluster size is of 2400 TB, but it is suggested to use upto only 2000 TB.

(iv) Calculation of Number of Data Nodes: The HDFS model follows the Mater-Slave Model. And in that model Data Nodes are the slave node. The Data Nodes stores data in the nodes. The Data Nodes provides replication and balancing. It also keeps an account of whether the slave nodes are alive/dead. The HDFS file is broken in small parts and these parts gets into Data Node.

The data node data is represented by 'D'

Formulae :

$$D = \frac{H1}{d} = \frac{C1 \times R1 \times S1}{d \times (Z)} \times 120\%$$

where,
d =The amount of disk Space available per node.
RAM, IOPS Bandwidth, CPU Configuration of nodes is taken into consideration.
H= HDFS Nodes Storage
Parameters that should be taken into consideration:
– There should be no failure in node.
– Other performance features are not relevant (processor, memory)
Adding new hardware is instantaneous.

Map Reduce Framework: The Map Reduce framework gets executed through parallel data processing. This framework makes small parts of the whole data i.e. dividing input data into small blocks and processes them all at a time into different machines. The Map Reduce works in two separate parts namely, Mapping and Reducing. It scans the whole input data and then it produces key value pairs, it is like one of the data structures in python named as dictionary where inputs are there in the form of key and value. After producing pairs, the Map Reduce library takes up those pairs and groups all the intermediate values attached with the same key and pass it to reduction function. Then the reduction function takes the respective key value and clusters them into smaller set of values.

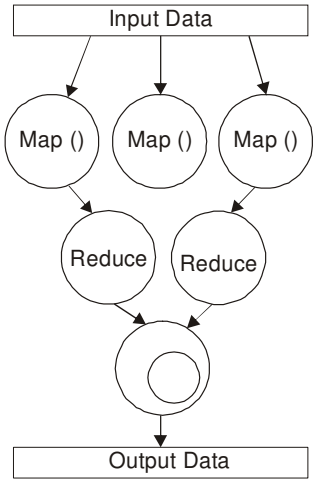

Fig. 4. Map Reduce Function.

III. DISCUSSION

In this paper we have discussed several concepts relating to cloud analytics ranging from its very unique architecture and various techniques which can be used in cloud analytics like HDFS architecture and map based functions with their implementation advantages and disadvantages we have also discussed various cloud services and implementation of HDFS architecture and map reduce framework on cloud analytics architecture.

IV. CONCLUSION

Various organizations are now investing in analysis and storing their data in cloud. Analysis making more informed choice decisions which is reflected as an advantage in their revenue system. Cloud and BIG DATA combination can be used for network optimization and we can identify any system breach which leads to better security over data. The introduction of various new concepts of IoT may lead to various growths in industry and can provide more jobs to data scientist and analysts. Some of the applications of the Cloud Analytics are :

– **Social Media:** The most important aspect of cloud analytics, from the data and analysis report of social media we can easily get an insight of a person's likes and dislikes.

– **Tracking Products:** Various Companies like Amazon uses data analytics to track their product and find out singular preference list for its entire user base.

– **Tracking Preference:** All the information's are stored in cloud and using analytics tool we can determine user preferences.

V. FUTURE SCOPE

As discussed in the paper analytics has huge impact in several industries and being able to implement it in the cloud can also increase its potential to huge amount by reducing time of execution and adding several interesting feature to the current technology. Hence as the research in the area of analytics grows it also improves technologies in cloud computing improving its performance and adding several features.

REFERENCES

[1]. Sujitha, K., & Praveen, K. (2015). Analysing cloud simulation results using big data analytics model. In *2015 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-6). IEEE.

[2]. Arora, R., Parashar, A., & Transforming, C. C. I. (2013). Secure user data in cloud computing using encryption algorithms. *International journal of engineering research and applications*, *3*(4), 1922-1926.

[3]. Knorr, E., & Gruman, G. (2008). What cloud computing really means. *InfoWorld*, *7*, 20-20. (Page No.)

[4]. Deshmukh, S., & Sumeet, S. (2015, December). Big Data Analytics Using Public Cloud Infrastructure: Use Cases and Cost Economics. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 782-784). IEEE.

[5]. Bagheri, H., & Shaltooki, A. A. (2015). Big Data: challenges, opportunities and Cloud based solutions. *International Journal of Electrical and Computer Engineering*, *5*(2), 340-343.

[6]. Khan, Z., Anjum, A., & Kiani, S. L. (2013). Cloud based big data analytics for smart future cities. In *2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing* (pp. 381-386). IEEE.

[7]. Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information systems*, *47*, 98-115.

[8]. Mayilvaganan, M., & Sabitha, M. (2013). A cloud-based architecture for Big-Data analytics in smart grid: A proposal. In *2013 IEEE International Conference on Computational Intelligence and Computing Research* (pp. 1-4). IEEE.

[9]. Ramamoorthy, S., & Rajalakshmi, S. (2013). Optimized data analysis in cloud using BigData analytics techniques. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.

[10]. Hayes, B. (2008). Cloud computing. *Communications of the ACM*, *51*(7), 9-11.

[11]. Sayeed, Z., Liao, Q., Faucher, D., Grinshpun, E., & Sharma, S. (2015). Cloud analytics for wireless metric prediction-framework and performance. In *2015 IEEE 8th International Conference on Cloud Computing* (pp. 995-998). IEEE.

[12]. Vuppala, S. K., Dinesh, M. S., Viswanathan, S., Ramachandran, G., Bussa, N., & Geetha, M. (2017). Cloud based big data platform for image analytics. In *2017 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)* (pp. 11-18). IEEE.
[13]. Elgendy, N., & Elragal, A. (2014). Big data analytics: a literature review paper. In *Industrial Conference on Data Mining* (pp. 214-227). Springer, Cham.
[14]. Dietrich, D., Heller, B., & Yang, B. (2012). EMC Education Services. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. pp. 1–410.
[15]. Marella, S. T., Karthikeya, K., Kushwanth, V. S., & Bezawada, A. (2018). Enhancement of Performance and Economy of Data Centers by Virtualization. *International Journal of Simulation--Systems, Science & Technology*, *19*(6).
[16]. Gunasekhar, T., Rao, K. T., & Basu, M. T. (2015). Understanding insider attack problem and scope in cloud. In *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]* (pp. 1-6). IEEE.
[17]. Bezawada, A., Marella, S. T., & Gunasekhar, T. (2018). A Systematic Analysis of Load Balancing in Cloud Computing. *International Journal of Simulation--Systems, Science & Technology*, *19*(6).
[18]. Marella, S.T., Karthikeya, K., Myla, S., Sai, M.M. & Allam, V. (2019). Detecting Fraudulent Credit Card Transactions Using Outlier Detection. *International Journal of Scientific & Technology Research, 8*(10), 630-637.
[19]. Myla, S., Marella, S. T., Goud, A. S., Ahammad, S. H., Kumar, G. N. S., & Inthiyaz, S. (2019). Design Decision Taking System for Student Career Selection for *Accurate Academic System*, *8*(9), 2199-2206.
[20]. Bhattacharya, A., Choudhury, S., & Cortesi, A. (2019). Replaceability and negotiation in a cloud service ecosystem. *Journal of Cloud Computing*, *8*(1), 1-14.
[21]. Prakash, K., Khasdeo, A., & Barve, A. (2019). Investigation of Regenerative Braking System of Several PMSM Motor with Vector Controller Applicability in HEV. *International Journal of Electrical, Electronics and Computer Engineering*, *6*(2), 4-9.
[22]. Gargav, N., Shrivastava, R., & Jamal, R. (2016). Fixed Point Results in Complete Random Convex Metric Spaces. *International Journal of Theoretical and Applied Sciences*, *8*(1), 1-6.

How to cite this article: Myla, S. Marella, S.T., Karthikeya, K., Preetham, B. and Ahammad, S. K. H. (2019). The Rise of "Big Data" in the Field of Cloud Analytics. *International Journal on Emerging Technologies*, *10*(4): 125–130.