International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 185-191
Published Online March 2021 in IJEAST (http://www.ijeast.com)

BIG DATA IN CLOUD COMPUTING: A LITERATURE REVIEW

Neelay Jagani
Department of IT
KJSCE Vidyavihar, Mumbai, Maharashtra, India

Parthil Jagani
Department of Computer Science
Florida Institute of Technology, Melbourne, Florida, USA

Suril Shah
Department of Computer Science
K.J. Somaiya College of Science and Commerce, Mumbai, Maharashtra, India

Abstract—**Big Data and Cloud Computing are two of the most important technologies of the day. Since data is being generated exponentially every day, Big Data has gained a lot of significance in any technology. The daily explosion of data means that it's better to have big data included in the applications. Whereas cloud computing is allowing users to use platforms according to their time, convenience and affordability. It is providing users ability to collaborate and work efficiently more than ever. Combining these two technologies can give a hands down advantage to the users in terms of knowledge and efficiency. Big Data, when used in cloud computing has applications in different fields such as Finance, Management, supply chain, planning, data storage, warehouses and many more. In this paper, we have discussed Big Data implementation and application in Cloud Computing. 4 V's in big data can be applied in Cloud computing to get better performance, higher input details, better insights, reliable and secure platforms at comparatively lower costs. Different analytics, technology involved in coupling of big data with cloud computing, the challenges involved in this process, trends applications of the domain and security factors involved are discussed in this paper.**

Keywords— **Cloud Computing, Big Data, Efficiency.**

I. INTRODUCTION

Big data treats ways to analyse, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Big data, as its name suggests, simply means a very huge amount of data. The typical data characteristics can be explained using 4V's. Cloud computing can be simply called as on-demand availability of computer system resources, particularly data storage and computing power. Cloud computing typically allows users to access, use, work and modify their work while collaborating with peers. Cloud computing allows users to work according to their convenience while big data provides insights and information. The analytics performed involves characteristics analysis, storage management and cloud, big data processing and finally deriving insights i.e. piece of knowledge from the huge data available. Nowadays, digital security is one of the most important factors. In case of big data, security is absolutely critical, since the data consists of confidential information, secret keywords, passwords, which, if compromised, can have very dangerous consequences. So security is extremely important while considering big data and cloud computing. The security can be achieved through different ways such that Node Authentication, encryption, access control, honeypot nodes etc. The implementation of this system may face different challenges such as data storage, speed, security, processing, transmission, visualization, architecture, integration, quality etc. Cloud computing with Big Data has applications in many fields such as Management, Finance etc.

II. BIG DATA ANALYTICS

Big Data in cloud refers to enormous size of the dataset perhaps in few dozens of terabytes and petabytes and thus working with them in a traditional local computer based Database Management System becomes enormously difficult. The ability to scale storage, visualize data, manage and capturing becomes very tedious and highly costly and thus use of cloud is the most apt solution. Many of the world's largest organization are storing all of their data on cloud. These enterprises are able to explore large volumes of highly detailed data so as to discover facts they didn't know with the help of inbuilt cloud features or deploying their own functionality on the cloud. Naturally businesses can benefit from large data with almost real time capability, and thus the cloud needs to have different data architecture, analytical methods, and tools.

A. Characteristics of Big Data–

The feature characteristics of big data are divided into 4V's, namely Volume, Variety, Velocity, Veracity. The first V, volume refers to the size of the data and how big the data is.

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 185-191
Published Online March 2021 in IJEAST (http://www.ijeast.com)

This is the primary and most looked attribute of big data. Velocity refers to the rate at which data is been collected or is changing. Some Big Data's like Stock Market prices are monitored and collected at a very high velocity with frequency as small as one second. The third feature Variety refers to from how many different sources the data is coming from, the data can be coming from logs, social media or even click streams. The last feature Veracity describes how good the data is. The quality of data is measured by observing patterns on how much data is inconsistent, missing, incomplete, approximated, decepted, ambiguous, or latent.

B. Storage Management and Cloud –

There are several software packages available on cloud to facilitate cloud computing. Enterprise data warehouse can be used or if there is presence of unstructured data like large texts use of NoSQL can be used. Majorly Hadoop, Spark, Map Reduce, HBase are used. Hadoop is the most available programming framework, written in Java that supports processing of large amount of data. With Hadoop large amount of data can be analysed by use of clusters of servers on these servers we can have thousands of nodes running the application. Hadoop framework helps in risk of system failure even multiple nodes fail. The framework has a flexible and fault tolerant computing solution. The Hadoop Distributed File System (HDFS) defines a very efficient yet high tech file system where in the file system spans all nodes in Hadoop cluster for data storage and connects the file system on local nodes, this improves the reliability significantly. HBase is a NoSQL software with Hadoop framework of HDFS. It is an open source database that was modelled after Google's Big table and like Hadoop is written in Java. It is widely used to store Big Data, more accurately when big data is of variety (unstructured). Spark is also an open source tool with unified analytics engine for large scale data processing. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. As part of storing data on cloud with appropriate tools Azure HDInsight and Amazon EMR are most popular. Both these cloud services for storing data have proven to be enormously effective. They are cloud native, meaning they enable us ML services, create clusters of Hadoop, Spark, Map Reduce, and even HBase. When the velocity of data is high, cloud storage enables to scale workloads up and down accordingly. The services allow for creating data pipelines to optimize work and by creating cluster on demand and paying for what you use; cost can be reduced significantly. The cloud also provides with various other clusters that are open source like Kafka, Apache Interactive Query, Apache Storm suiting to customer's needs. The cloud supports many programming languages as well for operations on data like Java, .NET, Python, Go, Scala and Clojure. The cloud also provides powerful visualization tools inbuilt, removing the stress of purchasing software's. Azure HDInsight comes with Microsoft Power BI at the same time AWS comes with Tableau.

C. Big Data Processing –

There are four fundamental requirements for processing.

1. The primary requirement is the ability to load the data quickly.
2. Fast query processing.
3. Efficient utilization of storage space.
4. Strong adaptivity to highly dynamic workload.

To satisfy all the four requirements efficiently, the cloud service providers help us by providing Map Reduce Software, both Azure HDInsight and Amazon EWS provide Map Reduce framework. The framework helps enormously in processing as it is a parallel programming model. The Map Reduce framework rather than increasing the storage capacity of a server or a computer, or increasing computational power, it adds more servers and computer. Therefore, the fundamental concept is that we do not scale up rather we scale out.

In Map Reduce a task is broken down into stages and are executed parallely thus increasing the efficiency. The working is quite simple as the word suggests; the first word Map is used to "map", the smaller tasks and assign them appropriate key value pair. Like for example if we have unstructured data like text, the key could be any word and the value can be the number of occurrences of that word. Next is the reduce function. The reduce function performs collection and combination of the output generated by "map", by combing all values which share the same key value, to provide the final result of the computational task. This is very advantageous as cloud architecture is very fast and when clubbed with parallel processing, the performance is unmatched to a general local computer. When processing speeds are this high, we can analyse data in real time whilst getting the output in real time as well. Such a system when implemented on cloud is very advantageous, and Big Data with high velocity and high volume, companies, exchanges like NASDAQ, BSE, NSE can all benefit. The storage, analytics and processing all are carried out with more efficiency and lower cost when compared to traditional normal computers.

III. CHALLENGES

In spite of all the advantages of the integration between cloud computing and big data, there are some challenges and risks that ought to be thought while deploying big data on a cloud environment.

A. Data Storage –

With the advancement of technology, we are able to wit- ness an exponential growth in data. However, most of the generated data is ignored or deleted because enough space is not available to store them. So, the primary challenge for Big Data analysis is storage mediums and better transmission rates. The available storage technologies do not possess the required

ability to process Big Data. Storing data on traditional physical storage systems is a complicated task as hard disk drives often fail, and traditional data protection mechanisms are not efficient. In addition to this, the velocity of Big Data must be such that the storage systems must be able to scale up quickly when required, which is actually difficult to achieve with these traditional storage systems. Due to this ever-growing data, data mining tasks has increased considerably which has led to wide diversity of data. There's a need to pay more attention for designing storage systems and to make efficient data analysis tools that will provide guarantees on the output since the data is gathered from different sources. Moreover, machine learning algorithms can be designed for analyzing the data which will help in improving the efficiency and scalability. The unlimited storage along with high fault tolerance offered by Cloud storage services (such as: Amazon S3, Elastic Block Store) provides solutions to address Big Data storage challenges. But, it's very expensive to host and transfer Big Data on the cloud since the size of data is gigantic.

B. Data Transmission –

Another challenge is how to move vast amounts of big data (let's take for example hundreds of terabytes of data) into a public cloud in a short period of time? How will we deal with the storage, reliability, privacy, and security issues? Transferring gigantic volumes of data in different stages of data life cycle poses challenges in each of these stages. Therefore, we need to devise smart pre-processing techniques and data compression algorithms to effectively reduce the data size before transferring the data. For transferring data from local data centers to cloud platforms, we need to develop efficient algorithms which will automatically recommend the appropriate cloud service (location) based on the geotemporal principles (since data can be at different locations) to maximize the data transfer speed while the minimizing cost.

C. Computational Complexities –

For processing large volumes of data, we require dedicated computing resources, which we usually handle by the increasing speed of storage, network and CPU. However, the processing power and the computing resources provided by the traditional computing system is insufficient for processing the data. The virtually unlimited and on-demand processing power offered by cloud computing acts as a partial solution. However, shifting to the cloud results in some issues. First, the network bandwidth of cloud computing is very limited which affects the efficiency of computation over large volumes of data. Second, the data is dispersed at different locations which makes it difficult to gather it for pre-processing. The essential features of cloud computing such as virtualization, pooled resources of data and high computing power makes it a difficult task to track and ensure data locality, and hampers its ability to support data processing which involves intensive communication and exchange of data.

D. Data Security–

Some security vulnerabilities arise due to the integration of Big Data and Cloud Computing. Also, the data security policies and schemes work with the structured data which is stored in conventional DBMS and aren't effective in handling highly heterogeneous and unstructured data. Therefore, we need to make effective policies for data access control and safety management so as to incorporate new data management systems and storage structures. Ensuring data confidentiality, integrity and availability becomes elemental in this cloud era since the data owners have limited control over the data and various resources. Heterogeneity is one of the most known Big Data's cloud security vulnerability. In many cases the deployment of Big Data requires it to deploy on a new cloud platform which will need new security tools to be developed as the existing security tools and practices won't work for such platforms. These security tools should include encryption, authentication, intrusion detection, access control, monitoring and event logging. Along with the security policies, while integrating Big Data to the cloud environment, consolidation plans should be taken into consideration.

E. Data Privacy –

It's been noticed that the cultural challenge of cloud computing and big data lies in the aspects of privacy. According to many researchers, most of big data sources takes the styles of documents, messages, images, audio and video posts and also very sensitive information like an individual's location, behaviour, transactions or companies tracking the employees' movement and productivity which are digitally recorded via social media implies that the most important resources for big data is relatively the social media and hence, accessing users' private information - which accounts for a major risk.

F. Different Conceptual Ideas of these Domains –

The concepts of consolidation and resource pooling comprises the base concepts of Cloud Computing whereas, big data systems (such as Hadoop) are built on the shared nothing principle, where each node is self-sufficient and independent. Integrating big data with cloud computing technologies can led businesses and educational institutes in a better direction for the future. Cloud computing has the capability to store enormous amount of data in various forms processing it at very large speeds which will result in data that can guide the education and business institutes for fast development. Nonetheless, there is a huge concern regarding security and privacy issues while moving to the cloud environment which is the main reason why the educational and business institutes aren't willing to move to cloud.

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 185-191
Published Online March 2021 in IJEAST (http://www.ijeast.com)

IV. TECHNOLOGY

The cloud service types for Big Data analytics as a service includes infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS).

A. Infrastructure as a Service (IaaS) –

To enable enterprises for allocating or buying time on shared server resources (which are often virtualized) for handling the computational and storage needs for Big Data analytics IaaS can be deployed on premise or via a cloud provider. Managing high-performance network, servers and storage resources is done by the cloud provider. Enterprises involved in Big Data do not need to maintain the hardware and software required for such performance. Hadoop, is an open-source solution, which employs distributed data storage and processing.
The technologies that are used for IaaS purposes are: the Hadoop framework, or a NoSQL database, such as MongoDB, Apache Cassandra or Couchbase technologies. IaaS solutions providers are Amazon Web Services, Windows Azure, Citrix CloudPlatform, Microsoft System Centre, OpenStack software, Rackspace etc.

B. Platform as a Service (PaaS) –

For providing higher-level programming models and database systems PaaS is used. It provides tools and libraries for building, testing, deploying, and to run applications on cloud environment.

Amazon ElasticMapReduce can be used which provides a basic Hadoop framework PaaS environment. Windows Azure's data service HDInsight brings Hadoop to the cloud environment coupled with Power Map, Power View and other Microsoft BI tools. Common PaaS offerings from DynamoDB for NoSQL database services and AWS including Redshift for data warehousing. PaaS capabilities are also offered by Google such as: Bigtable, BigQuery.

C. Software as a Service (SaaS) –

For delivering applications over the internet Software as a service (or SaaS) is used. SaaS offers jKool which provides business related cloud-based solutions and a real-time analysis of time-sensitive information. Concur is one of the fast-growing TE SaaS company, which runs only a single instance of its software and contains preferences, history of millions of business travelers on a global scale for airlines, hotels, car rentals, taxi services, etc. Karmasphere also offers a pay-as-you-go application which analyses data stored with Amazon S3 using Amazon Elastic Map Reduce.

V. SECURITY

A. Need for security in Big Data –

Big data is used by too many of business but they may not have environment from perspective of the security. If any safety problem occurs to big data, it may come out with even more serious issue. Generally, companies use this technology to store data of zeta byte range regarding to the company. This potentially results in severe criticality for classification of information.to secures the data we either need to encrypt, log or use honeypot techniques. The challenge of detecting attacks and intruders, must be solved using big data style analysis. Analysis and computation of big data: Fastness is the main thing when we look up for database in the big data. However, the process may be hectic only because of the reason that it cannot traverse all related data in the whole database in a little time. While the big data is getting complex, the indices in the big data are aiming at the simple type of the data. The traditional series algorithm is inefficient for this big data.

B. Challenges of security in cloud computing –

The challenges of safety in cloud computing environments can be categorized into network level, user authentication level, data level, and general issues.

1. Network level: The problems that can be categorized in a network level deal with network protocols and network security, such as distributed data, Internode communication, distributed nodes
2. Network level: The challenges that can be categorized under user authentication level deals with various encrypting/decrypting techniques, authentication methods such as authentication of applications and nodes, and logging, administrative rights for nodes.
3. Data level: The challenges that can be categorized under data level deals with availability such as data protection and distributed data.
4. General types: The challenges that can be categorized under general purpose are traditional security tools, and use of various technologies

C. Ways to tackle security problems –

1. Encryption: Since the data in any computer will be present in a cluster, a person can easily steal the data from the system. This may become a serious problem for any company or organization to safeguard their very important data. To avoid this thing, we may go for encryption of the data. Different encryption mechanisms can be used for different systems and the keys generated should be stored safely behind

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 185-191
Published Online March 2021 in IJEAST (http://www.ijeast.com)

firewalls. By choosing this way the data of the customer is kept secure.

2. Node authentication: The node must be go from authentication whenever it joins the cluster. If the node turns out to be a malicious cluster then such nodes should not be authenticated.
3. Honeypot nodes: The honeypot nodes are disguised to be like a regular node but is a trap. It automatically traps the hackers and will not allow any harm to happen to the system or the data.
4. Access control: The various privacy and access control in the distributed environment will be a good measure of security. To prevent the information from leaking we use Linux operating system. The Linux is a feature that provides the mechanism for supporting access control security policy through the use of Linux Security modules in Linux kernels.

D. Ways to tackle security problems –

Cloud computing helps in storage of data at a remote site so that we can maximize resource utilization. Therefore, it is very important for this data to protect and access should be given only to authorized people. Therefore, this amounts to secure third party publication of data that is required for data outsourcing, as well as for outside publications. In the cloud computing , the machine serves the role of a third party publisher, which stores the sensitive data in the cloud. The data needs to be protected, and the above techniques have to be used to ensure the timely maintenance of authenticity and completeness.

VI. ADVANTAGES

There are many advantages of integrating cloud computing to big data. Big data raises the need for multiple servers because of the large amount of data and size it deals with, and it demands high velocity and variability. These various servers operate in parallel to meet the high demands of big data. Cloud computing already uses multiple servers and allow resource allocations. Because of that, it is a great fit to build the big data on these cloud multi-servers and using the resource allocation availability offered by the cloud environments which in result improve the efficiency for big data analysis. The performance of both will be enhanced by using a cloud infrastructure as a storage system for big data. Since cloud services are built on remote multi-servers, they can manage huge amounts of data at the same time. This feature enables big data to deal with large volumes of data using advanced analytics techniques. Cost reductions will benefit from the convergence of cloud computing and big data. To accommodate the large amount of data, big data requires clusters of servers and volumes. Instead of building new servers and volumes for big data, cloud computing systems will act as the foundation for all of them, allowing for greater flexibility and scalability while still eliminating the significant investments in big data machines and servers. In addition to that, using cloud computing provides faster provisioning to big data as provisioning servers in the cloud is so easy and feasible. As a result, the cloud environment used can be scaled depending on the big data processing requirements of the big data. This fast provisioning is crucial for big data since the value of the data decreases rapidly over time.

Cloud computing, in general, complements big data by offering a simple, on-demand, and shared computing platform with minimal management effort and overhead. It also makes the environment more robust, automated and provides multi-tenancy.

Big data allows the end users to visualize the data and companies can explore new market opportunities. Data analytics is another significant benefit of big data, as it helps individuals to personalise information or access and communicate with real-time websites. Furthermore, the convergence of the two makes big data resources more manageable, monitor able, and reportable. Furthermore, this integration allows for a reduction in complexity and an increase in efficiency. Because of all of these benefits, cloud-based approaches are the best models for deploying big data.

VII. APPLICATIONS

The big data application refers to the large-scale distributed applications which usually work with large data sets. In the span of big data, data exploration and analysis became a difficult problem in many industries. With large and complex data, conventional data processing applications fail to manage computation, prompting the development of big data applications. Google's map reduce framework and apache Hadoop are the software systems for big data applications, in which these applications produce a large amount of intermediate data. Big data systems are primarily used in manufacturing and bio- informatics. Big data offers a transparent infrastructure for the manufacturing sector, allowing it to address uncertainties such as inconsistency, component performance and availability. A conceptual structure of predictive manufacturing starts with data acquisition in these big data applications, where various types of sensory data such as pressure, vibration, acoustics, voltage, current, and controller data can be acquired.

Big data in manufacturing is generated by combining sensory data with historical data. The input is the generated big data from the above combination. The parallel distributed computing system is combined with computer clusters and web interfaces in cloud computing. Software packages for Big Data include a rich range of tools and options that allow an individual to map the entire data field across the company,

enabling the individual to evaluate the risks he or she faces internally. This is regarded as one of the most significant benefits, as big data ensures data security. This helps a person to identify potentially sensitive information that is not properly protected and ensures that it is processed in compliance with regulatory requirements.

If big data are combined with predictive analytics, it produces a challenge for many industries. The combination results in the exploration of these four areas: Calculate the risks on large portfolios, Detect, prevent, and re-audit financial fraud, improve delinquent collections, execute high value marketing campaign.

Companies may use big data to build new products and services, boost existing ones, and even invent completely new business models. Big data analytics can be used to obtain such benefits in a number of fields, including consumer intelligence, supply chain intelligence, performance quality, and risk management, and fraud detection. In the field of risk management, sectors such as investment or retail banking, as well as insurance, will benefit from big data analytics. Big data analytics can aid in the selection of investments by analysing the probability of gains versus the probability of losses, which is a crucial feature of the financial services industry. Internal and external big data may also be evaluated for the full and dynamic appraisal of risk exposures.

Big data analytics can be used to detect and prevent fraud, especially in the government, banking, and insurance industries. While analytics are still widely used in automated fraud detection, organisations and sectors are increasingly looking to big data to improve their systems. They can use big data to match electronic data from various sources, both public and private, and perform faster analytics. Manufacturing, retail, central government, healthcare, telecom, and banking are few of the sectors that can benefit from big data analytics.

VIII. CONCLUSION

Big data and cloud computing play a huge role in the current digital world. The application of Big Data in Cloud Computing seems to have a huge potential in the coming years. While using Software as Service, typically, big data plays a pretty important role in giving insight, in cloud computing applications. Big Data when applied in cloud computing, has many applications in different fields. Some of these applications include improved analysis due to large data size, creation of an efficient infrastructure while reducing the cost in the long run and allowing better integrity and availability and security of the cloud platform, letting the businesses and platforms grow through the means of big data.

IX. REFERENCES

[1] Alsghaier, Hiba & Al-Shawakfa, Emad. (2018). An empirical study of cloud computing and big data analytics. International Journal of Innovative Computing and Applications. 9. 180. 10.1504/IJICA.2018.10014870.

[2] Yadav S., Sohal A. (2017) "Review Paper on Big Data Analytics in Cloud Computing" in International Journal of Computer Trends and Technology (IJCTT) V49(3):156-160, July 2017. ISSN:2231-2803.

[3] Hariharan, U. & Kotteswaran, Rajkumar & Pathak, Nilotpal. (2020). The Convergence of IoT with Big Data and Cloud Computing. 10.1201/9781003054115-1.

[4] Agrawal, Divyakant & Das, Sudipto & Abbadi, Amr. (2011). Big Data and Cloud Computing: Current State and Future Opportunities. ACM International Conference Proceeding Series. 530-533. 10.1145/1951365.1951432.

[5] Ibrahim Abaker Targio Hashem, Nor Badrul Anuar, Salimah Mokhtar (2014) Information Systems 47:98-115 "The rise of the "Big Data" on cloud computing: Review an open research issues". DOI:10.1016/j.is.2014.07.006

[6] Alyass, Akram & Turcotte, Michelle & Meyre, David. (2015). From big data analysis to personalized medicine for all: Challenges and opportunities. BMC medical genomics. 8. 33. 10.1186/s12920-015-0108-y.

[7] Assuncao, Marcos & Calheiros, Rodrigo & Bianchi, Silvia & Netto, Marco & Buyya, Rajkumar. (2014). Big Data computing and clouds: Trends and future directions. Journal of Parallel and Distributed Computing. 75.

[8] Wang, Lizhe & von Laszewski, Gregor & Younge, Andrew & He, Xi & Kunze, Marcel & Tao, Jie & Fu, Cheng. (2010). Cloud Computing: A Perspective Study. New Generation Comput.. 28. 137-146. 10.1007/s00354-008-0081-5.

[9] Sengupta S., Kaulgud V., Sharma V. (2011) "Cloud Computing Security- Trends and Research Directions" in IEEE Computer Society,Pg 524-531. DOI:10.1109/SERVICES.2011.20

[10] Turcotte M., Alyass A. (2015) "Form Big Data Analysis to Personalized Medicine for all: Challenges and opportunities" Article in BMC Medical Genomics. DOI:10.1186/s12920-015-0108-y

[11] Martian A.,Vulpe A.,Suciu G., Cranciunescu R.(2015) Big Data, Internet of Things and Cloud Convergence- An Architecture for Secure E-Health Applications.Article in Journal of Medical Systems. DOI:10.1007/s10916-015-0327-y

[12] Qi, Lianyong & Khosravi, Mohammad & Xu, Xiaolong & Zhang, Yiwen & Menon, Varun. (2021).

Cloud Computing. 10.1007/978-3-030-69992-5.

[13] Marchiori, Massimo. (2017). Learning the way to the cloud: Big Data Park. Concurrency and Computation: Practice and Experience. 31. 10.1002/cpe.4234.

[14] Brevini, Benedetta. (2015). Book Review: To the Cloud: Big Data in a Turbulent World. Media, Culture & Society. 37. 1111-1113. 10.1177/0163443715596318a.