IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-11 E-ISSN: 2347-2693

Inregration and Interelation of Bigdata With Cloud Computing: A Review

T. Saha[1*], K. Das [2]

[1*]Information Technology, JIS College of Engineering, Kalyani, India
[2] Information Technology, JIS College of Engineering, Kalyani, India

** Corresponding Author: tanusree.saha@jiscollege.ac.in, Tel.: +91 9547274437*

Available online at: www.ijcseonline.org

Received: 17/Oct/2017, Revised: 29/Oct/2017, Accepted: 20/Nov/2017, Published: 30/Nov/2017

***Abstract*:** In this current era, Information technology opens the door through which human enters into the smart , developing society with modern services in all aspects of living as well as several areas of business, medical and scientific studies, engineering and resulting a massive and exponential growth of data. Handling this voluminous data with traditional information technology frame work becomes a challenging and time demanding task in terms of data collection, storage, retrieval, analysis and application. At the same time cloud computing becomes a powerful model for data processing, storing massive amount of data and perform complex computation. We need to influence cloud computing techniques and solutions to deal with big data problems. In this review paper we focus on integration of big data in cloud environment, a comprehensive description of big data and its features, cloud computing and its characteristics and the relationship between both technologies and challenges it's facing.

Key words: Big Data, Big Data Management Tools, Cloud Computing, Cloud Services, Cloud Issues

I. INTRODUCTION

The continuous increase in the volume of data due to explosion of information technology by different organization such as the rise of social media, application of multimedia, internet of Things(IoT) has produce an overwhelming flow of data. Data may be either in structured form or in unstructured form. Data creation is occurring at record rate. Various types of business data are growing by exponential orders of magnitude. These issues have become great obstacles to the realization of a digital society, network society, and intelligent society. The data management and analytics carried out in conventional database systems cannot address the Big Data challenges: data size is too large, values are modified rapidly, and/or they do no longer satisfy the constraints of Database Management Systems (DBMS). Also Big Data environment requires multiple servers to handle large volume of data with high velocity. Cloud Computing has been designed to reduce computational costs and increase the elasticity and reliability of the systems. It is also intended to allow the user to obtain various services without taking into consideration the underlying architecture, hence offering a transparent scalability. The basis of Cloud Computing is the Service-Oriented Architecture which is designed to allow developers to overcome many distributed organization computing challenges including application integration, transaction management, and security policies. Since cloud environments make use of multiple servers and allocate resources on demand, it would be highly beneficial for organizations to make use of cloud services for big data analysis.

Big Data: Definition and Characteristics

At present the industry does not have a unified definition of big data; Big data has been defined in differing ways as follows by various parties:"Big Data refers to datasets whose size is beyond the capability of typical database software tools to capture, store, manage, and analyze"

"Big Data usually includes datasets with sizes beyond the capability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time."

Big data has four main characteristics: Volume, Velocity, Variety, and Value (referred to as "4V," referencing the huge amount of data volume, fast processing speed, various data types, and low-value density). Following are brief descriptions for each of these characteristics.

Volume: It refers to the large amount of data involved with big data. The scale of datasets keeps increasing from gigabytes (GB) to TB, then to the petabyte (PB) level; some even are measured with exabytes (EB) and zettabytes (ZB). For instance, the video surveillance cameras of a medium-sized city in China can produce tens of TB data every day.

Variety: Indicates that the types of big data are complex. In the past, the data types that were generated or processed were simpler, and most of the data was structured. But now, with the emerging of new channels and technologies, such as social networking, the Internet of Things, mobile computing, and online advertising, much semi-structured or unstructured

data is produced, in the form of text, XML, emails, blogs, and instant messages, as just a few examples, resulting in a surge of new data types.

Velocity: The velocity of data generation, processing, and analysis continues to accelerate. There are three reasons: the real-time nature of data creation, the demands from combining streaming data with business processes, and decision making processes. The velocity of data processing needs to be high, and processing capacity shifts from batch processing to stream processing. There is a "one-second rule" in the industry referring to a standard for the processing of big data, which shows the capability of big data processing and the essential difference between it and traditional data mining.

Value: The value is the most important aspect of big data. It refers to the process of discovering huge hidden values from large datasets with various types and rapid generation.

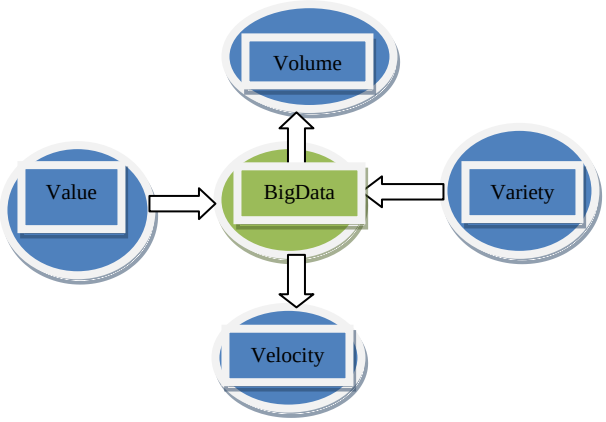

Fig1 : Big Data Characteristics

What is Cloud Computing?

Cloud computing is a type of computing that relies on sharing computing resources rather than having local servers or personal devices to handle applications. Cloud computing is comparable to grid computing, a type of computing where unused processing cycles of all computers in a network are harnesses to solve problems too intensive for any stand-alone machine.

In cloud computing, the word cloud (also phrased as "the cloud") is used as a metaphor for "the Internet," so the phrase cloud computing means "a type of Internet-based computing," where different services — such as servers, storage and applications —are delivered to an organization's computers and devices through the Internet.

The world of Cloud Computing is totally virtual to its users that require minimum effort from user to manage with features like: on-demand, scalability, reliability, maintenance, cost-effective, and flexibility. The services are delivered to users of it through the use of Internet and sharing of resources can be done using network of remote servers to store, manage and process data with distributed data processing system. Its service-oriented architecture supports "everything as a service", offers their "services" according to different models with infrastructure-, platform- and software-as-a-service.

How Cloud Computing Works?

The goal of cloud computing is to apply traditional supercomputing or high-performance computing power, normally used by military and research facilities, to perform tens of trillions of computations per second, in consumer-oriented applications such as financial portfolios, to deliver personalized information, to provide data storage or to power large, immersive online computer games. To do this, cloud computing uses networks of large groups of servers typically running low-cost consumer PC technology with specialized connections to spread data-processing chores across them. This shared IT infrastructure contains large pools of systems that are linked together. Often, virtualization techniques are used to maximize the power of cloud computing.

II. CLOUD COMPUTING ENVIRONMENT FOR BIG DATA MANAGEMENT

Cloud computing technology provides parallel processing for handling big data with advance analytical application. Cloud Computing is an environment based on using and providing services. There are different categories in which the service-oriented systems can be clustered. One of the most used criteria to group these systems is the abstraction level that is offered to the system user. In this way, three different levels are often distinguished: **Infrastructure as a Service (IaaS**), **Platform as a Service (PaaS**), and **Software as a Service (SaaS**) as

IaaS, such as Flexi scale and Amazon's EC2, refers to hardware equipment operating on a cloud provided by service providers and used by end users upon demand

PaaS, such as Google's Apps Engine, Salesforce.com, Force platform, and Microsoft Azure, refers to different resources operating on a cloud to provide platform computing for end users.

SaaS, such as Google Docs, Gmail, Salesforce.com, and Online Payroll, refers to applications operating on a remote cloud infrastructure offered by the cloud provider as services that can be accessed through the Internet.

III. RELATIONSHIP BETWEEN CLOUD COMPUTING AND BIG DATA

Cloud computing and big data are conjoined. Big data provides users the ability to use commodity computing to process distributed queries across multiple datasets and return resultant sets in a timely manner. Cloud computing provides the underlying engine through the use of Hadoop, a class of distributed data-processing platforms. Large data

sources from the cloud and Web are stored in a distributed fault-tolerant database and processed through a programming model for large datasets with a parallel distributed algorithm in a cluster. The main purpose of data visualization is to view analytical results presented visually through different graphs for decision making.

Big data utilizes distributed storage technology based on cloud computing rather than local storage attached to a computer or electronic device. Big data evaluation is driven by fast-growing cloud-based applications developed using virtualized technologies. Therefore, cloud computing not only provides facilities for the computation and processing of big data but also serves as a service model.

Cloud computing and big data are complementary, forming a dialectical relationship. Cloud computing and the Internet of Things' widespread application is people's ultimate vision, and the rapid increase in big data is a thorny problem that is encountered during development. Cloud Computing is a trend in technology development, while big data is an inevitable phenomenon of the rapid development of a modern information society.

To solve big data problems, we need modern means and Cloud computing technologies. Cloud Computing offers scalability with respect to the use of resources, low administration effort, flexibility in the pricing model and mobility for the software user. Under these assumptions, it is obvious that the Cloud Computing paradigm benefits large projects, related with Big Data.

IV. BIG DATA MANAGEMENT TOOLS IN CLOUD

Big Data produces big challenge to manage massive amount of structured and unstructured data to handle. Cloud computing offers scalable solution to manage large amount of data in cloud environment to take advantages of both technologies .To effectively incorporate and manage Big Data in cloud environment it is important to understand the tools and services offered by them.

Hadoop: Hadoop is an open-source Apache Software Foundation project written in Java that enables the distributed processing of large datasets across clusters of commodity. Hadoop has two primary components, namely, HDFS and MapReduce programming framework. The most significant feature of Hadoop is that HDFS and MapReduce are closely related to each other, each are co-deployed such that a single cluster is produced. Therefore, the storage system is not physically separated from the processing system.

HDFS is a distributed file system designed to run on top of the local file systems of the cluster nodes and store extremely large files suitable for streaming data access. HDFS is highly fault tolerant and can scale up from a single server to thousands of machines, each offering local computation and storage. HDFS consists of two types of nodes, namely, a name node called "master" and several data nodes called "slaves." HDFS can also include secondary name nodes. The name node manages the hierarchy of file systems and director namespace (i.e., metadata). File systems are presented in a form of name node that registers attributes, such as access time, modification, permission, and disk space quotas. The file content is split into large blocks, and each block of the file is independently replicated across data nodes for redundancy and to periodically send a report of all existing blocks to the name node.

MapReduce: The MapReduce is a simplified programming model for processing large numbers of datasets pioneered by Google for data-intensive applications. The Map Reduce model was developed based on GFS and is adopted through open-source Hadoop implementation, which was popularized by Yahoo. Apart from the Map Reduce framework, several other current open-source Apache projects are related to the Hadoop ecosystem, including Hive, Hbase etc. MapReduce accelerates the processing of large amounts of data in a cloud, thus, MapReduce, is the preferred computation model of cloud providers. MapReduce is a popular cloud computing framework that robotically performs scalable distributed applications and provides an interface that allows for parallelization and distributed computing in a cluster of servers.

V. ISSUES AND CHALLENGES

Although cloud computing has been broadly accepted by many organizations, research on big data in the cloud remains in its early stages. Several existing issues have not been fully addressed. Moreover, new challenges continue to emerge from applications by organizations: Following are the few challenges:

Scalability: Scalability is the ability of the storage to handle increasing amounts of data in an appropriate manner. Scalable distributed data storage systems have been a critical part of cloud computing infrastructures. The lack of cloud computing features to support RDBMSs associated with enterprise solutions has made RDBMSs less attractive for the deployment of large-scale applications in the cloud.

Availability: Availability refers to the resources of the system accessible on demand by an authorized individual. In a cloud environment, one of the main issues concerning cloud service providers is the availability of the data stored in the cloud. For example, one of the pressing demands on cloud service providers is to effectively serve the needs of the mobile user who requires single or multiple data within a short amount of time. Therefore, services must remain operational even in the case of a security breach. In addition, with the increasing number of cloud users, cloud service providers must address the issue of making the requested data available to users to deliver high-quality services

Data integrity: A key aspect of big data security is integrity. Integrity means that data can be modified only by authorized parties or the data owner to prevent misuse. The proliferation of cloud-based applications provides users the opportunity to store and manage their data in cloud data centers. Such applications must ensure data integrity. However, one of the main challenges that must be addressed is to ensure the correctness of user data in the cloud. Given that users may not be physically able to access the data, the cloud should provide a mechanism for the user to check whether the data is maintained.

Data quality: In the past, data processing was typically performed on clean datasets from well-known and limited sources. Therefore, the results were accurate. However, with the emergence of big data, data originate from many different sources; not all of these sources are well-known or verifiable. Poor data quality has become a serious problem for many cloud service providers because data are often collected from different sources. known and limited sources. Therefore, the results were accurate. However, with the emergence of big data, data originate from many different sources; not all of these sources are well-known or verifiable. Poor data quality has become a serious problem for many cloud service providers because data are often collected from different sources.

Privacy: Privacy concerns continue to hamper users who out-source their private data into the cloud storage. This concern has become serious with the development of big data mining and analytics, which require personal information to produce relevant results, such as personalized and location-based services. Information on individuals is exposed to scrutiny, a condition that gives rise to concerns on profiling, stealing, and loss of control.

Heterogeneity: Variety, one of the major aspects of big data characterization, is the result of the growth of virtually unlimited different sources of data. This growth leads to the heterogeneous nature of big data. Data from multiple sources are generally of different types and representation forms and significantly interconnected; they have incompatible formats and are inconsistently represented.

Legal/regulatory issues: Specific laws and regulations must be established to preserve the personal and sensitive information of users. Different countries have different laws and regulations to achieve data privacy and protection. In several countries, monitoring of company staff communications is not allowed. However, electronic monitoring is permitted under special circumstances. Therefore, the question is whether such laws and regulations offer adequate protection for individuals' data while enjoying the many benefits of big data in the society at large.

VI. OPEN RESEARCH ISSUES

Numerous studies have addressed a number of significant problems and issues pertaining to the storage and processing of big data in clouds. The amount of data continues to increase at an exponential rate, but the improvement in the processing mechanisms is relatively slow. Only a few tools are available to address the issues of big data processing in cloud environments. State-of-the-art techniques and technologies in many important big data applications cannot solve the actual problems of storing and querying big data. For example, Hadoop and MapReduce lack query processing strategies and have low-level infrastructures with respect to data processing and management. Despite the overabundance of work performed to address the problem of storing and processing big data in cloud computing environments, certain important aspects of storing and processing big data in cloud computing are yet to be solved.

VII. CONCLUSION

In this study, we presented a review on theory of big data in cloud computing. The size of data at present is huge and continues to increase every day. The variety of data being generated is also expanding. The velocity of data generation and growth is increasing because of the increase of mobile devices and other device sensors connected to the Inter-net. The use of cloud services to store, process, and analyze data has been available for some time; it has changed the context of information technology and has turned the promises of the on-demand service model into reality. However existing tools and technologies are not fully adequate to provide security and privacy of sensitive information stored in cloud by users. It's a big challenge that should be taken care with the use of strong and efficient encryption techniques for data disguise in cloud environment.

REFERENCES

[1] Zikopoulos PC, Eaton C, DeRoos D, Deutsch T, Lapis G. Understanding big data.New York, NY: McGraw-Hill; 2012.
[2] Bell G, Hey T, Szalay "A. *Beyond the data deluge"*. Science 2009;323(5919):12978.
[3] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big data: the next frontier for innovation, competition, and productivity. MacKinsey Global Institute;2011.
[4] Big Data Research and Development Initiative, http://www.whitehouse.gov/sites/ default/files/microsites/ostp/big_data_press_release_final_2.pdf..
[5] Gupta R, Gupta H, Mohania M. "Cloud computing and big data analytics: what is new from databases perspective? Big data analytics. "Berlin, Heidelberg: Springer; 2012. p. 4261.
[6] Ibrahim Abaker Targio Hashem , Ibrar Yaqoob , Nor Badrul Anuar , Salimah Mokhtar , Abdullah Gani , Samee Ullah Khan "The rise of "big data" on cloud computing: Review and open research issues" ,
[7] "*Big Data Processing in Cloud Computing Environments*" Changqing Ji, Yu Li,Wenming Qiu, Uchechukwu Awada, Keqiu Li
[8] Charlotte Castelino1, Dhaval Gandhi, Harish G. Narula, Nirav H. Chokshi "Integration of Big Data and Cloud Computing"

[9] Alberto Fernández, Sara del Río,Victoria López,Abdullah Bawakid,María J. del Jesus José M. Benítezand Francisco Herrera *"Big Data with Cloud Computing: an insight on the computing environment, MapReduce,and programming frameworks*".

[10] *Big Data Technologies and Cloud Computing available at* "http://scitechconnect.elsevier.com/big-data-technologies-and-cloud-computing-

Authors Profile

Tanusree Saha, pursed Bachelor in Information Technology from JIS College of Engineering in 2009 and Master in Software Engineering fron JIS College of Engineering in 2011. She is currently working as Assistant Professor in Department of Information Technology since from 2011. She is a life member of the FOSET since 2012.Her main research work focuses on Cryptography Algorithms, Network Security, Cloud Computing, Big Data. She has 6 years of teaching experience and 3 years of Research Experience.

Kakali Das, pursuing Bachelor in Information Technology from JIS College of Engineering. She is a 4th year student now and her project work is going on under the guidence of Ms. Tanusree Saha, Assistant Professor, Information Technology, JIS College of Engineering, Kalyani.

