Cloud with Big Data Analytics: A Review

Ashish Soni1 S.N. Panda1  Naveen Sharma1  Ashima Mehta1 Alisha Mehta1

**Abstract: Cloud computing is heavily involved in the IT industry in the present scenario and is going to be a major force in taking the IT industry to greater heights in the forthcoming future. Cloud Computing provides a number of IT services to its clients over the internet with on-demand access. There are huge amounts of data generated in cloud computing termed as “Big Data”; furthermore this data has the potential to provide knowledgeable information when analyzed. This paper discusses cloud computing, its service models, types of cloud, the relation between big data and cloud, analysis of big data in cloud using Hadoop Map Reduce framework. The main aim of the paper is to discuss and gain knowledge about cloud computing , big data , big data analysis using Hadoop, major challenges and security related issues in cloud computing and list some approaches that can help in eradicating these stumbling blocks for a promising future of cloud computing in the IT industry.**

*Index Terms*— Big Data, Cloud Computing, Hadoop, MapReduce

1. **INTRODUCTION**

Cloud Computing is in the process of revolutionizing the modern computing world. It came into existence in early 1960’s when John McCarthy an American computer scientist predicated the eventual emergence of a computing platform which would allow a great degree of flexibility in providing different IT resources in order to fulfill industry demands [1].It provides flexibility to its users as they only pay for the services they use and also provide benefits like improved availability and cost reduction.

*According to National Institute of Standards and Technology “cloud computing is a platform for providing a number of computing resources on demand that are easily accessible to the its users (e.g. networks, servers , storage application and services)and that are easily managed and maintained by the service provider[2].*

Ashish Soni

[ashish.soni2091@gmail.com](mailto:ashish.soni2091@gmail.com)

S.N. Panda

[snpanda@chitkara.edu.in](mailto:snpanda@chitkara.edu.in)

Naveen Sharma

[naveen.sharma@chitkara.edu.in](mailto:naveen.sharma@chitkara.edu.in)

Ashima Mehta  
[ashimamehta03@gmail.com](mailto:ashimamehta03@gmail.com)

Alisha Mehta

[alishachd.mehta20@gmail.com](mailto:alishachd.mehta20@gmail.com)

**1**Chitkara University,Chandigarh-Patiala National Highway (NH- 64), Village Jhansla, Patiala, Punjab 140401, India

Cloud computing is a successful example of service oriented computing. Three of its paradigms are *IaaS* (Infrastructure as a Service), *PaaS* (Platform as a Service),

*and SaaS* (Software as a Service). It helps in providing everything from hardware resources, softwares required and datasets etc to data analysts. Its ability to provide and manage resources with ease makes it ideal to store and analyze big data. Big data has the potential to be analyzed for valuable information, to optimize performance in the transforming modern society. It refers to structured, unstructured and semi structured data. Big data is characterized broadly by five V’s attributes: Volume, Velocity, Variety, Veracity and Values [3]. Big data is defined as -the data that exists in huge volumes, consists of different types of data structures for storage and forms new bits of data daily very quickly which when processed helps

in decision making, finding new insights and optimizing processes [4].

Cloud computing platform helps in solving both the storage and the computational demands of various applications that involve big data analytics. The analysis of big data is done using specific technologies which have been designed to handle such huge amounts of data. Big data analytics makes use of data analysis algorithms that require hardware with high performance capabilities. Cloud computing platform is capable of providing both computational and data processing applications [6]. In today’s data analytics world, MapReduce framework and its open source implementation Hadoop are being widely adopted [7].

Furthermore, there are a number of challenges faced by cloud computing that we are going to discuss and focus on in this paper. These challenges are classified as: 1. Security problems faced by cloud service providers and 2. Security problems faced by customers that are further categorized as security issues in terms of service delivery models [8]. Also various weak spots and dangers that exist in cloud computing, lead to these security problems have been identified. Due to these shortcomings, an attack might occur therefore resources and the data stored can be at risk [8].This paper covers the basics related to cloud computing, its service deployment models, types of clouds, benefits of cloud are also discussed along with big data and big data analytics in cloud computing using Hadoop and Map Reduce. We will also discuss some potential solutions for providing security. The main focus will be on the Security challenges faced in cloud computing. This paper is arranged as follows: Section 2 consists of Literature review. Section 3 defines Big Data using five V’s of big data. Section 4 discusses the relation between big data and the cloud. Sections 5, 6 and 7 focuses on cloud computing architecture, cloud service deployment models and types of cloud. There is a table discussing security challenges and issues associated with big data and cloud. Section 8 and 10 discusses potential solutions for security and big data analysis in cloud using Hadoop respectively. Section 10 consists of conclusion of the paper.

1. **LITERATURE REVIEW:**

**Iyanda, OlukunleA** et.al. discusses Big data and Cloud computing. Two major concerns – Security and Loss of Control are put forward and both of these issues affect cloud computing. that tampers with the Benefits provided cloud computing like Flexibility and storage, Time saving, Reduced Cost [1].

**Rani, B. Kezia** et.al. discusses cloud computing in detail with a view to provide knowledge about the architecture , characteristics and deployment models of cloud computing and illustrates the concept of inter-cloud on which the future technologies are dependent on, so discusses some issues which need to resolved for a better future[2].

**Kumar, Vinay** et.al. defines big data and discusses it from the point of view of large amount data available in India and the role it can play into transforming the country into digital India. Applications of big data: The way big data is being utilized in the various fields is discussed. Challenges in the deployment of big data and solutions for handling the challenges like Streaming algorithm, MapReduce are proposed [3].

**Ji, Changqing** et.al. discusses ways in which we can carry out big data processing in a cloud environment as there has been rapid growth of data over various platforms and this data has to be analyzed. This leads to challenges in management of data which can be solved by using cloud data management. The architecture and platform of cloud has been discussed followed by discussion on MapReduce paradigm and the ways to optimize it [4].

**Zhang, Qi** et.al. discusses cloud computing. Cloud computing is compared with related technologies like Grid computing, Utility computing, Virtualization and Autonomic computing. Some Technologies used in cloud computing and cloud computing products that today are being utilized in the industry are discussed. Research challenges that affect cloud computing and the need to address them for better prospects in cloud computing is put forward [5].

**Manekar, Amit Kumar** et.al. discusses about two technologies – one is Cloud Computing and the other is Big Data. How these technologies are being utilized in the IT industry to handle data and then deliver services of big data based on Cloud Computing model. Using Cloud based big data Analytics to club big data and cloud computing in one environment. Technological aspects that include Hadoop and MapReduce are put forward as a solution to the issues arising in these technologies. [6].

**Agrawal, Divyakant** et.al. focuses on analysis of scalable data management solutions using cloud computing. Both Update heavy applications and decision support systems are discussed. Increasing popularity of Key Value stores due to single tenant systems and the rise of Hadoop (Open source implementation of MapReduce) has been illustrated. Large multitenant databases for data management consisting of a number of small applications, each of which has smaller footprints are also discussed. So, the design principles and problems in cloud data management that need to be addressed are put forward [7].

**Khan, Nabeel** et.al. discusses the definition and objective of cloud computing. Framework for adoption of cloud computing is also mentioned. The main focus was on identifying security threats current and future. The outcome was the discovery of total 18, current and future, security problems that could affect cloud computing, to eradicate them so that organizations are encouraged to adopt cloud computing [8].

**Purcell, Bernice M.** discusses big data and the latest advances in technologies that are currently being used in the industry. The hardware and processing resources these technologies make use of are mentioned, its installation cost to various small and medium sized enterprises is also discussed. Reasons for companies to implement these technologies for their benefits are put forward. Major concerns regarding cloud computing are also addressed [9].

**Fan, Jianqing** et.al. discusses Big data along with its background. The objective of analyzing big data is put forward. Paradigm shift is discussed as we need new methods to handle big data challenges. Rise of big data and its description in various fields is discussed. Features of Big that pose challenges to analysis of data are discussed. Impact of big data on cloud computing infrastructure and tools like Hadoop, Map reduce are also discussed. Issues related to big data and the need to eradicate those using statistical methods is also addressed [10].

**Samuel, S Justin Rvp** et.al. defines Big Data. Big data classes i.e. categorization of big data and some research areas are discussed in detail. Technical challenges are also put forward. A table about some recent evolvements in big data research is depicted and issues related to big data that need to be addressed are brought to light [11].

**H. Bhosale** et.al. discusses big data and problems with big data processing that need to be addressed. Hadoop is put forward as a solution for big data processing and architectures of HDFS and MapReduce are also explained [12].

**I. Sriram** et.al. discusses Cloud computing and its Methodology in detail. Lessons from related technologies have been mentioned. Cloud computing standards and interfaces have been described. Cloud Interoperability and Novel Protocols are discussed. Description on building clouds is given and the new technologies and use cases that become possible through cloud computing are explained [13].

**Lola Yorita Astri** et.al. aims to find how certain factors affect an organization that uses cloud computing. Cloud computing is defined along with the three services that are provided in cloud computing, namely- SaaS, PaaS, IaaS. The three models of cloud computing are also discussed. Factors that impact cloud adoption in various organizations are discussed [14].

**Manogaran** et.al. discusses about challenges and potential solutions for protecting big data in cloud computing. MetaCloudDataStorage architecture is proposed for security- which categorizes data into three levels such as Sensitive, Critical and Normal for protecting big data in cloud computing environment to ensure efficient processing of big data in cloud computing environment and gains more business insights. MapReduce Algorithm is also explained. An example of processing big data in Amazon web service with Apache Hive is also discussed [15].

**Venkatesh** et.al. discusses big data in cloud computing. Three V’s of big data are discussed along with Hadoop, MapReduce and big data applications and also advantages of big data. Big data management and need of security in big data is addressed. Some possible solutions for security are proposed [16].

**Assunção, Marcos D** et.al. discusses various ways for performing analytics on Clouds for Big Data applications. The four major areas are discussed. How data is managed and stored on cloud in different ways is explained. Storage and analysis of data using Hadoop is illustrated. The research challenges in this field, business and non-technical challenges are discussed and the need to address those using proper tools is highlighted [17].

**W. Fan** et.al. discusses big data in a detailed manner. Implementations of big data in various fields are illustrated. The way in which Global Pulse is using big data for development to improve life in countries which are still not fully developed is discussed. Some Contributions are mentioned that show traditional research in big data mining. Controversies involving big data are mentioned in brief. Tools like Hadoop, MOA, and Vowpal wabbit are put forward. The challenges that could arise in Big Data in the future are discussed briefly [18].

**A.Abouzeid** et.al. discusses Hadoop and Map reduced systems and how they are being utilized in today’s industry for data analysis. The desired properties of a system designed for performing data analysis are mentioned. Background and shortfalls of available approaches is also mentioned. Hadoop’s implementation background and its components are discussed. Hadoop is compared with two benchmarked systems and also concludes that HadoopDB is a hybrid of parallel DBMS and Hadoop approaches to analysis of data.[19].

**Gandomi, Amir** et.al. discusses big data in a broad sense along with its characteristics and unique features. Although the main focus is on analytical techniques implied in big data analysis. Analysis involving data that is unstructured and predictive analytics regarding structured data are discussed and the need to develop tools and techniques to analyze big data is addressed [20].

1. **FIVE V’s OF BIG DATA**

Big data is defined/characterized using five V’s. These characteristics describe the data that is being generated everyday in today’s world [11, 18, 20].

**3.1 Volume:**

The amount of data is growing exponentially day by day. The data accumulated around the world daily is in the size of zeta bytes. The quantity of data that is collected these days is incomprehensible.

**3.2 Variety:**

There are various forms of data that originate from different sources. Data can be in structured, unstructured and semi-structured format.

**3.3 Velocity:**

The speed at which data is generated and comes in for processing refers to velocity of the data. The data has to be stored, processed and then later retrieved. This has to be done in a minimum amount of time.

**3.4 Variability:**

It refers to the difference in the structures of data and how it looks to a user that wants to analyze the data. It also depends upon variation in data flow rates.

**3.5 Value:**

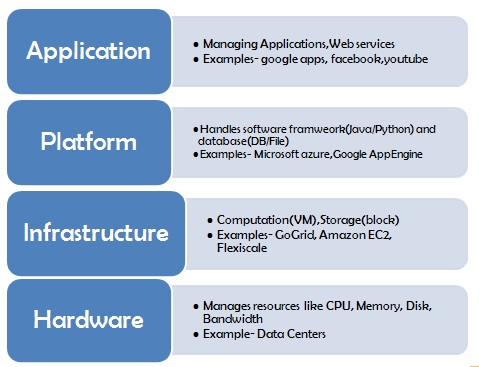
It refers to the value that can be extracted from the data i.e. the knowledge that can be gained after analyzing the data that can help in better decision making.

1. **BIG DATA IN RELATION TO CLOUD**

The term big data implies that the datasets are so large that the current systems in the industry do not have the capability to store and analyze these datasets. There is huge volumes of data that is emerging in the current market, this data comes in many forms namely structured , unstructured and semi structured data, also data comes from many different sources that include email, social networks, historical information, e-Governance services, medical hospitals, banks, business transactions media houses, defense and corporate sectors and from sensors like CCTV. Storing and analyzing big data presents a big challenge to the IT industry. The most common model used for storing big data is known as clustered Network-Attached Storage. A number of computers are connected to a single computer in a network-attached storage (NAS) pod using the (NAS) device [9]. Several NAS pods are connected together through the computer used as the NAS device [9]. It is an expensive model to implement in practice. Cloud services provide a better prospect at substantially lower costs. The analysis of big data is performed by a software framework called Map Reduce. In Map Reduce, two functions take key part in the process of data analysis; they are known as map and reduce functions. Map function takes input in the form of key/value pairs and gives output in the form of intermediate set of key/value pairs and then the Reduce function merges that output associated with the task/query to be implemented and gives the reduced result as the final output. The mapping is performed by every NAS device separately; it requires processing to be done in a parallelized manner. Processing in parallel manner in Map Reduce is expensive. The Cloud Computing environment provides a better platform for the processing needs of the data at much lower cost.

1. **Cloud Computing Architecture:**

Architecture of Cloud consists of 4 layers: [5], as shown in Fig.1



**Fig. 1 Layers of Cloud Architecture**

**5.1 The hardware layer:** also known as the datacenter layer, it manages the physical resources of the cloud that include servers, routers, switches, power and cooling systems. In practice, this layer is generally implemented in different data storage sources. Issues that can be encountered in this layer include, tolerating faults, Congestion control, and utility resource management.

**5.2 The infrastructure layer:**recognized as the virtualization layer, it takes a number of resources for storing and fulfilling computational needs, it is done by splitting the resources available by making use of the technologies available in the market such as VMware, Xen and KVM. This layer is a key part of cloud computing as it provides capabilities like dynamic resource management, that is possible through virtualization.

**5.3 The platform layer:** situated above the infrastructure layer, it contains operating systems and application frameworks. The main objective of the platform layer is to decrease the workload of integrating applications directly into VM containers.

**5.4 The application layer:** The topmost layer in the hierarchy, it contains the real time cloud applications which offer scalability, on demand access and pay as per daily usage in the cloud computing environment that help to achieve better performance at lower costs.

Each layer in cloud computing architecture is loosely coupled with the layers above and below, allowing each layer to evolve separately.

1. **Cloud Service Deployment models:**

Software as a Service, Platform as a Service, Infrastructure as a Service are the common service deployment models in cloud computing. SPI is an acronym used to describe these service models. There is one more service model known as hardware as a service, but it is not used as extensively as the other three models.

**6.1 Software as a Service (SaaS):**

In SaaS model, multiple users can access the applications over the internet. However, a user cannot customize a service being provided in this model. Salesforce.com, Rackspace and SAP Business by Design are some examples [5].It provides benefits like easy administration of software, automatic updating of system, compatibility across the business organization, easy access etc [9].

Characteristics of SaaS are:

1. Service is available on the web.
2. Maintenance is taken care of by companies who provide facilities to the user.
3. Follows “one to many model”, so it is very cost effective.
4. Offers on-demand availability and scalability according to requirements.

**6.2 Platform as a Service (PaaS):**

In PaaS model, the users have platforms for developing and customizing applications according to their requirement over the internet. It differs from SaaS as it allows users to customize applications and services based on their needs. Improved Security, user-friendly, using technologies which have been tried and tested before etc are some advantages of PaaS [9]. Google AppEngine and Windows Azure Platform are examples of PaaS [2].

Characteristics of PaaS are:

1. Offers scalability and security
2. Offers development tools to develop and customize applications
3. Provides integration capabilities

**6.3 Infrastructure as a Service (IaaS):**

In IaaS model, users have the complete cloud computing infrastructure that includes servers, storage, network and operating system in accordance with the demands and needs of the user. IaaS provides benefits like flexibility, security, choice of service and business agility. Amazon ECC, GoGrid and Linode are examples of IaaS [2].

Characteristics of IaaS:

1. Availability of various resources on-demand.
2. Offers replication feature that lowers the risk of losing vital data.
3. Offers Scalability and reliability.

**6.4 Hardware as a Service (HaaS):**

Haas model is still in early stages and not being used that extensively. Haas is based upon timesharing model which was used in 1960s and 1970s on minicomputers and mainframes [1].The user can directly license the hardware. Timesharing developed into the practice of managed services [1]. Google with its Chrome books for business, Char Tec and Equus are examples of HaaS [1].

1. **TYPES OF CLOUD:**

Various cloud models exist in Cloud computing. The four models that are currently at work in the industry are discussed as follows [1]:

**7.1 Public cloud:** A public cloud offers pay-as-you-use services to the people. It is also known as external cloud. In public cloud setup the users do not own the resources and services; they are just using them via web applications/web services [2].

**7.2 Private cloud:** A private cloud is used exclusively by a single organization. The cloud may be managed by the organization itself or a third party cloud service provider. It is also known as an internal cloud. The private cloud can exist onshore or offshore [2].

**7.3 Hybrid cloud**: A Hybrid cloud [23] is setup using both public and a private cloud. They operate individually but

are connected through a proprietary technology [17]. In this, an organization uses the public cloud to work on extra

tasks and manage non-confidential information. The private cloud is used to maintain and manage critical and important

information within the organization [2].

**7.4 Community Cloud:** A community cloud is shared among a number of organizations those have shared interests. It is owned and managed in the community or by a third party service provider [2].

**7.5 Virtual Private Cloud (VPC):** has gained recognition recently, it offers a new way to address the drawbacks of public and private clouds. A VPC runs on top of public clouds as a platform [5].

**Table 1 Security Challenges and Issues in Big data and Cloud [8]:**

|  |  |
| --- | --- |
| Security Threat | Description |
| **1. Account Hacking** | A client’s account is taken over by an intruder to conduct illegal or unauthorized activities |
| **2. Data Scavenging** | The attacker tries to find confidential or sensitive data by searching through a system’s data storage. |
| **3. Data Leakage** | The unauthorized transfer of confidential information to the outside world. |
| **4. Denial of service** | An authorized user is denied access to the computing resources by a malicious user who has taken over all the resources for own agenda. |
| **5. Customer Data manipulation** | The data sent to server’s application by the application module is tampered with to conduct unauthorized activities the applications. For example- SQL injection, command injection etc. |
| **6. VM Escape** | It affects the infrastructure (IaaS) as the hypervisor is used unfairly to take control of the complete infrastructure. |
| **7. VM Hopping** | It allows one VM the right to use other VMs. This happens because there is no proper management of utilization of resources of VMs. |
| **8. Non-trustable VM Creation** | VM images are created only by an authenticated user that uses a certain virus to harm the storage places. The root cause of this is the lack of control in placing of VM images in public storage source. |
| **9. Lack of security in VM migration** | Data is accessed illegally by an attacker, the VM travels to a susceptible source by creation and migration of many VMs. It is caused by exposure of data of the VM. |
| **10. Sniffing/Spoofing virtual networks** | Virtual networks are monitored by an intruder by using VM and some of the packets are also tampered with by using ARP spoofing (location). It is caused because of sharing of virtual bridges among many networks. |
| **11. Eaves Dropping** | A system is hacked and services like video calls, messaging, audio calls are tampered with by an intruder.. |
| **12. Hypervisor viruses** | The hypervisor layer is affected by a malicious piece of code. |
| **13. Authorized interception spot** | It helps in analyzing the data that comprises of different types of content. The affect of losing this data to an unauthorized user will be destructive. |
| **14. Virtual machine security** | Virtualization is an important part of the cloud setup. There are a lot of threats that can affect the security of the system if proper protocols are not utilized while using the system. |
| **15. Trusted transaction** | Security is a challenge in transactions like e-business, data etc. Trust is necessary in these transactions. |
| **16. Smartphone data slinging** | Smartphones are being used to access private data that can have a harmful affect on the whole system. |
| **17. Lack of Security in APIs** | Cloud users make use of various services by utilizing a number of APIs. If there’s a problem in these APIs, then the whole system is prone to issues like privacy, security etc. |
| **18. Risk of Sharing Resources** | For better efficiency and performance, resources are shared among many users but there is a risk to the security of the system if one system gains access to confidential information of the system. |

1. **Potential Solutions for Security:**

**8.1MetaDataCloudStorage Security Architecture:**

MetaDataCloudStorage Security Architecture is proposed to protect big data stored on cloud from hackers. Here, data is stored on the basis of its scope and importance. Data is categorized into three dimensions namely- Critical, Sensitive and Normal. Each category of data is stored in distinct datacenters with more than one copy of similar data being stored on different places provided. MetaDataCloudStorage Interface can access the appropriate data center as per the request of the user. AWS (Amazon Web Service) Cloud Trail is utilized to analyze the log files. AWS Key Management Service (KMS) incorporated with AWS Cloud Trail that helps in delivering the log files to an Amazon S3 bucket. It is easy to combine Cloud Trail with other applications using API.

Here, datacenters will be divided into**’t’** parts, where each part can be represented by part **r (r(s, t)).** The **‘u’** different storage providers are responsible for storing and acknowledging them as provider **q(q(s, u)).**

**‘t’** will be always more than **’u’**, the **’u’** storage providers are a part of different enterprises.

Data parts are stored on certain cloud storage providers will be allocated to some physical storage media that belongs to the storage provider. The big data stored in the data center will form a distinctive storage path given as:

MappingStorage\_Path=**{Data((A1(D1,D2,……DX))(A2(D1,D2,……Dy)(An(D1,D2,……Dz))**

Where **’A’** specifies the storage provider and **‘D’** specifies the physical storage media. In this MetaDataCloudStorage framework, the path where big data is stored is encrypted and a cryptographic value is obtained, which is known as cryptographic virtual mapping of big data i.e. MetaDataCloudStorage framework safeguards mapping of different data elements to each storage provider with the help of MetaDataCloudStorage Interface [15].

To ensure high availability and robustness, a number of copies of same data will be stored at distinct cloud storage providers. The Administrator that controls the system will keep the storage index information for each data part; so that if a problem occurs, another copy of data part can be found easily according to storage index information.

Fig.2 depicts the way the user will access the application and data in distributed cloud. Fig.3 shows the architecture for Security in MetaCloudDataStorage in Cloud. The security algorithm which is shown below that prevents the illegal access when trying to login to application that is being utilized in cloud. Although this algorithm updates the following tables such as 1) Threat updated table, 2) MetaData Storage Cloud table, 3) AmazonCloud data storage table, 4) GoogleCloud data storage table, 5) XCloud data storage table, 6) XnCloud data storage table.

Threat updated table will store the entries related to illegal attempts and Metadata storage cloud stores information regarding data storage. Distinct tables are used for keeping data.

**ALGORITHM TO PROCESS THE LOG FILES:**

//Processing the log file

Selection of userid, password!, lastlogin, status, location\*, @browser, ipconfig, sysdate from updatedauditlog;

//Attempt to login by an intruder

{

Put into Threatupdated values (userid, password!, location\*, @browser, sysdate);

Return -1;

}

else {

List 1 = Selection of Storage provider,

ScopeofData from MetaDataCloudStorage where

user\_login$\_AppId#@;

If (1.DataStoragepoviderName = = “AmazonCloud” && 1.ScopeofData = = “critical”)

{

Goto->Amazon Data Cloud Storage

Select\*fromAmazonDataCloudStoragewhere user\_login$\_AppId#@;

}

elseif (1.DataStoragepoviderName = = “GoogleCloud” && 1.ScopeofData = = “sensitive”)

{

Goto->Google Data Cloud Storage

Select\*from GoogleDataCloudStorage where

user\_login$\_AppId#@;

}

elseif(1.DataStoragepoviderName==“XCloud”&& 1.ScopeofData = = “normal”)

{

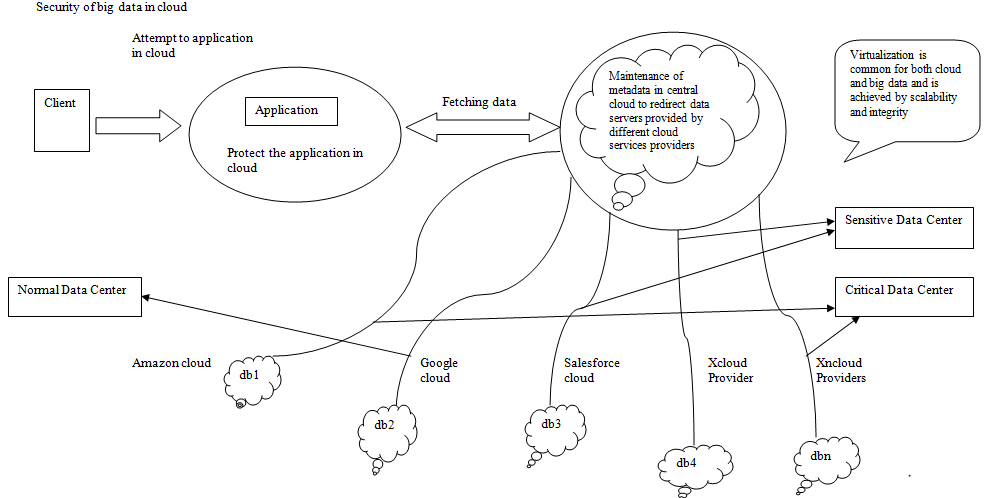
Goto->X Data Cloud Storage;

Select\*fromXDataCloudStoragewhere user\_login$\_AppId#@;

}

}

}//endelse



**Fig.2 Client accessing the application [15]**

**8.2 Encryption:**

Data stored in any system can be easily accessed by an unknown and unauthorized user. So, to safeguard the data everyone should encrypt the data. Various encryption techniques are utilized for different interfaces and the keys generated should be kept a secret. This will keep the data of the user secure [16].

**8.3 Nodes Authentication:**

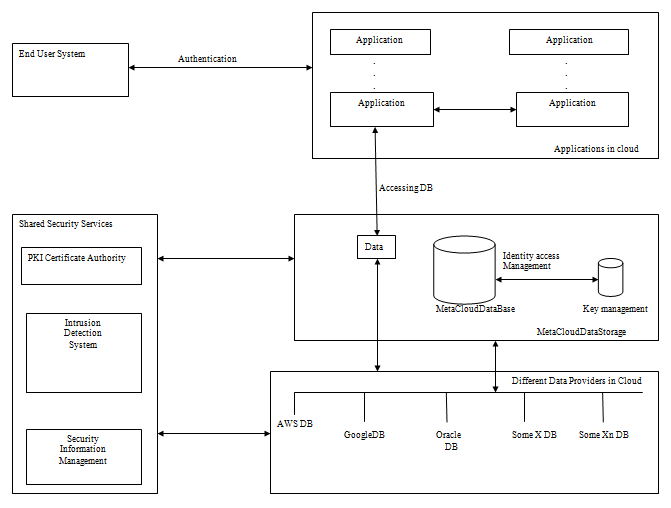
The node in the system should be authenticated whenever it accesses the system. If there is any doubt over node’s credibility, it must not be authenticated [16].

**8.4 Honeypot Nodes:**

These nodes seem like normal nodes but are actually traps that are placed to defend against unwanted attacks. They automatically trap the intruder and not allow for any damage of data to happen [16].

**8.5 Access Control:**

Implementation of different access control and privacy measures in a cloud computing environment will provide a better sense of security [16].



**Fig.3 Architecture for Security in MetaCloudDataStorage in Cloud [15]**

1. **BIG DATA ANALYTICS IN CLOUD USING HADOOP**

There are potentially huge amounts of data available in cloud that can be utilized for making better and informed decisions by extracting knowledgeable information using analytics. The analytics on data available on the cloud is performed using Hadoop, which is discussed in detail below:

**9.1 Hadoop:**

Hadoop framework has open source capabilities; developed by Google it provides storage space and processing ability for big data available in a shared computing environment

across clusters of computers. Hadoop consists of two major parts that are HDFS (Hadoop Distributed File System) and MapReduce. HDFS is used for storage and retrieval of data and MapReduce helps to handle and execute jobs for users. HDFS is used for the purpose of storing data in abundance while MapReduce is used for analyzing that data to extract valuable information. Both these components are discussed in detail below [12].

**9.2 Hadoop Distributed File System (HDFS) Architecture:**

Hadoop uses HDFS which is based on Google File System (GFS). HDFS is developed to work on huge clusters, designed to provide reliability and tolerate faults. HDFS is capable of storing huge amounts of information, scale up according to the need and avoid failures in storage infrastructure without losing data [17].HDFS follows a master/slave architecture where master contains one NameNode that manages the metadata and a number of slave DataNodes that contain the definite data. HDFS maintains data on the cluster by dividing large files into blocks that are stored redundantly in different set of DataNodes. 64MB is the default block size, which helps in reducing the amount of metadata storage required per file. The metadata is maintained by the NameNode, the metadata structures (filenames, directories) are allowed to be modified by many users. The NameNode decides the mappings of blocks to the DataNodes. The DataNodes implement the read and write processes and moreover manages the procedure of creating, deleting and replicating blocks (each DataNode consists of three copies of a file by default) [10].

**9.3 MapReduce Architecture:**

MapReduce is known as the processing pillar of the Hadoop framework. MapReduce is a programming model for processing large datasets in parallel [10]. Large datasets which are stored across a number of nodes in a cluster are processed via three basic operations. Firstly, a set of map tasks are implemented in each node in parallel in the cluster without communicating with other nodes. Then, that data is distributed again across the all nodes of the cluster. At last, a set of reduce tasks are performed in parallel by each node on the data it received. MapReduce is a fault tolerant model and can operate in heterogeneous environments with ease.

The two functions in MapReduce are as follows [12, 19]:

**Map -** this function takes input in the form of key/value pairs and generates output as in the form of intermediate set of key/value pairs.

**Reduce-** this function combines the output of the map function that is associated with the same intermediate key.

There are three stages in MapReduce which are described below [10].

**First stage: mapping.**

The first stage in MapReduce is known as mapping. In this, a dataset is provided to the ‘map’ function to transform it into key, value pairs.

**Intermediate stage: shuffle and sort.**

The intermediate output from the mapping stage is provided to different ‘reducers’ which process those inputs and store them in a sorted manner i.e. all the pairs with the same key are stored together on a node. This step is known as shuffling and sorting.

**Final stage: reducing.**

In this reducing stage, the result of an instance that the user is searching for is returned as the output.

1. **CONCLUSION:**

Cloud Computing and Big data have made a major impact on the IT industry and will continue to do so in the forthcoming future. They have changed the face of the present industry by bringing their attributes into play as they contribute to the work being done in the industry by providing simpler and efficient solutions to every problem. In this paper, we have defined big data, discussed the relation between cloud and big data, cloud computing and its architecture, deployment models, and types of cloud. The Hadoop MapReduce framework that is used for analysis big data in cloud has been discussed. The analytics of data stored in cloud provides knowledgeable patterns of data that help in making better decisions, planning and optimizing customer experience. Some challenges related to these technologies like security, have been mentioned and some approaches that can be used to overcome these challenges have been discussed. MetaDataCloudStorage Architecture has great potential to provide the perfect solution to every problem that may occur in these areas. However, more efforts need to be put in to look into that possible solution while also looking for other solutions to help in creating a perfect system for the industry.

REFERENCES:

(1) Iyanda, O., 2014. Big data and current cloud computing issues and challenges. *IJARCSSE*, *4*(6), pp.1192-1197.

(2) Rani, B.K., Rani, B.P. and Babu, A.V., 2015. Cloud Computing and Inter-Clouds–Types, Topologies and Research Issues. *Procedia Computer Science*, *50*, pp.24-29.

(3) Kumar, V., & Chaturvedi, A. (2016). Applications of Big Data in the Digital India: Opportunities and Challenges. *IRA-International Journal of Technology & Engineering (ISSN 2455-4480)*, *3*(3). http://doi.org/10.21013/jte.v3.n3.p7

(4) Ji, C., Li, Y., Qiu, W., Awada, U., & Li, K. (2012). Big Data Processing in Cloud Computing Environments. *2012 12th International Symposium on Pervasive Systems, Algorithms and Networks*. http://doi.org/10.1109/i-span.2012.9

(5) Zhang, Q., Cheng, L. and Boutaba, R., 2010. Cloud computing: state-of-the-art and research challenges. *Journal of internet services and applications*, *1*(1), pp.7-18.

(6) Manekar, A.K. and Pradeepini, G., 2015, December. Cloud Based Big Data Analytics a Review. In *Computational Intelligence and Communication Networks (CICN), 2015 International Conference on* (pp. 785-788). IEEE.

(7) Agrawal, D., Das, S. and El Abbadi, A., 2011, March. Big data and cloud computing: current state and future opportunities. In *Proceedings of the 14th International Conference on Extending Database Technology* (pp. 530-533). ACM.

(8) Khan, N. and Al-Yasiri, A., 2016. Identifying cloud security threats to strengthen cloud computing adoption framework. *Procedia Computer Science*, *94*, pp.485-490.

(9) Purcell, B.M., 2014. Big data using cloud computing. *Journal of Technology Research*, *5*, p.1.

(10) Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*(2), 293–314. http://doi.org/10.1093/nsr/nwt032

(11) Samuel, S.J., RVP, K., Sashidhar, K. and Bharathi, C.R., 2006. A SURVEY ON BIG DATA AND ITS RESEARCH CHALLENGES. *ARPN Journal of Engineering and Applied Sciences*.

(12) Bhosale, H.S. and Gadekar, D.P., 2014. A Review Paper on Big Data and Hadoop. *International Journal of Scientific and Research Publications*, *4*(10), p.1.

(13) Sriram, I. and Khajeh-Hosseini, A., 2010. Research agenda in cloud technologies. *arXiv preprint arXiv:1001.3259*.

(14) Astri, L.Y., 2015. A Study Literature of Critical Success Factors of Cloud Computing in Organizations. *Procedia Computer Science*, *59*, pp.188-194.

(15) Manogaran, G., Thota, C. and Kumar, M.V., 2016. MetaCloudDataStorage architecture for Big Data security in cloud computing. *Procedia Computer Science*, *87*, pp.128-133.

(16) Venkatesh, H., Perur, S.D. and Jalihal, N., 2015. A Study on Use of Big Data in Cloud Computing Environment. *International Journal of Computer Science and Information Technologies*, *6*(3), pp.2076-2078.

(17) Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A. and Buyya, R., 2015. Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, *79*, pp.3-15.

(18) Fan, W. and Bifet, A., 2013. Mining big data: current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter*, *14*(2), pp.1-5.

(19) Abouzeid, A., Bajda-Pawlikowski, K., Abadi, D., Silberschatz, A. and Rasin, A., 2009. HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads. *Proceedings of the VLDB Endowment*, *2*(1), pp.922-933.

(20) Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." *International Journal of Information Management* 35.2 (2015): 137-144.