RESEARCH REVIEW JOURNALS
Volume-04
Issue-07
July-2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com[Peer Reviewed Journal]

Big Data Security Issues and Challenges – A Survey

*[1]V. Ganesan & [2]Dr. N. Umadevi

[1]*Doctoral Scholar, Department of Computer Science, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts & Science, Coimbatore – 641 005 (India)*
[2]*Head & Associate Professor, Department of Computer Science, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts & Science, Coimbatore – 641 005 (India)*

ARTICLE DETAILS

Article History
Published Online: 15 July 2019

Keywords
Data Security, Big data, Cloud computing, Dataset and Big data Analytics.

***Corresponding Author**
Email: ganesanvgg[at]gmail.com

ABSTRACT

Big data and cloud computing are two important issues within the recent years, enables computing resources to be provided as data Technology services with high efficiency and effectiveness. Now a day's big data is one among the most problems that researchers try to solve it and focusing their researches over it to get ride the problem of however big data could be handling within the recent systems and managed with the cloud of computing, and also the one among the foremost important issue is a way to gain an ideal security for big data in cloud computing. Cloud computing could be a most powerful technology that performs massive-scale and complex computing. It eliminates the need to keep up expensive computing hardware, dedicated space requirement and connected software. Massive growth within the scale of knowledge big data generated through cloud computing has been identified. Thought of big data could be a challenging and time-demanding task that needs a large computational space to confirm successful processing and analysis. The term 'Big data' defines innovative techniques and technologies to capture, store, distribute, manage and analyze petabyte or larger-sized datasets with fast and totally different structures. Big data is also structured, unstructured or semi-structured, leading to incapability of typical knowledge management strategies. Data will be generated from various relevant sources and might store within the system at various rates. In order to research these massive amounts of knowledge in a cheap and economical means, correspondence technique is used. Our paper reviews a Survey of big data with cloud computing security and also the mechanisms that used to shield and secure also have privacy for big data with an available clouds.

1. Introduction

Big data is known as a datasets with size beyond the ability of the software tools that used nowadays to manage and process the data among a dedicated time. With variety, Volume, velocity big data such military knowledge or different unauthorized knowledge need to be protected in a scalable and efficient way [1]. Information privacy and security is one amongst most involved problems for Cloud Computing because of its open environment with terribly limited user side control. It's also a vital challenge for big data. when few years later a lot of knowledge globally would be touched with Cloud Computing that provides robust storage, computation and distributed capability in support of big data processing. Different issues are that information privacy and security challenges in each Cloud Computing and large knowledge should be investigated. the privacy and security providing such forum for researchers, and developers to exchange the most recent experience, analysis ideas and development on fundamental issues and applications about security and privacy problems in cloud and big data environments [2]. The cloud helps organizations and allows rapid on demand provisioning of server resources such as CPUs, manage, storage, bandwidth, and share and analyze their big data in a reasonable and easy to use. The cloud infrastructure as a service platform, supported by on demand analytics solution seller that creates the big size of information analytics terribly affordable. As location independent cloud computing Involving shared services providing resources , code and knowledge to systems and the hardware on demand, really the storage networking in cloud may be a terribly robust because use driver for high performance.

Figure 1.cloud computing in big data

Big data is as a collection of large dataset that cannot be processed using traditional computing techniques .Big data is not simply an information rather it's become a whole subject that involve various tools, techniques and framework. the requirement of big data generated from the big companies like

face book, yahoo, Google, YouTube etc for the aim of study of huge quantity of information also Google contains the massive amount of information big data could be a term that refers to dataset whose volume (size), complexity and rate of growth (velocity) create them to difficult to captured, managed, processed or analyzed by standard technology and tools like relative databases.

Information securities are often improved by big data technology that is useful from security tools like network monitoring, security information, and event management [5], [6]. However, on the down-side, there are additional security challenges brought by the large information technology, together with cryptography algorithms, information provenance, secure information storage, access management, real time observance and then on [7]. Identifying and analyzing the safety problems can bring a better usage of massive information. Thus, in this paper, we'll first survey existing analysis on security and privacy. Then, we'll specialize in an essential type of data: trajectory.

2. Big Data

The Big data is a high volume of data with a range of datasets that explodes in exponential pace. The conventional database system cannot handle humongous dataset, and thus, the big data paradigm has emerged. The mammoth sized dataset may be a very perplex to store, process and manage [12]. There are various data types that are stitched along to make the big data [13]. As a consequence, the information warehouses are additional vulnerable to security breach. There are three key knowledge types, namely, structured, semi structured and unstructured data [15]. Interestingly, 90th of big data are unstructured data types. The big data doesn't deal with petabytes anymore. The trending of big data is beyond exabytes, for instance, Google Inc. and NSA have 15 EB and 10 EB knowledge within the warehouse respectively.

Imagine a world whereas not data storage; a neighborhood where every detail a few person or organization, every group action performed, or each side which can be documented is lost directly once use. Organizations would thus lose the ability to extract valuable knowledge and information, perform careful analyses, additionally as offer new opportunities and edges. Something starting from client names and addresses, to product out there, to purchases created, to workers hired, etc. has become essential for normal continuity. Knowledge is that the building block upon that any organization thrives.

Now consider the extent of details and also the surge of information and data provided today through the advancements in technologies and also the online. With the increase in storage capabilities and methods of data assortment, large amounts of data became simply out there. Every second, a lot of data is being created and desires to be hold on and analyzed thus as to extract value. Moreover, information has become cheaper to store, so organizations got to be compelled to get the maximum quantity worth as potential from the large amounts of hold on data. The size, variety, and speedy modification of such data want a replacement kind of huge data analytics, still as altogether different storage and analysis strategies.

3. Cloud Computing

The Cloud Computing can be termed as internet based and are connected through the remote servers. Through this sharing of data process tasks, online access to laptop resources or services and centralized knowledge storage. The best examples are electric station, during which consumer use power without having the information of infrastructure to produce the service. Within the same manner, the cloud vendors use the resources as a service and pay just for resources that they use. Majority cloud computing infrastructures includes services delivered through common centers and rest on servers. Cloud Computing provides a surroundings for resource sharing in terms of control frameworks, middleware's and application development platforms, and business applications. The operation models of cloud computing grasp free infrastructure services with value another platform services, subscription-based infrastructure services with supplemental application services, and free services for sellers but sharing of revenues generated from shoppers [1]. The term Cloud Computing has been out lined in some ways in which by analyst firms, academics, business practitioners and IT firms. Clouds is an oversized pool of simply usable and accessible virtualized resources. These resources could also be dynamically reconfigured to control to a variable load (scale), permitting additionally for an optimum resource utilization [2].

4. Big Data and Cloud Computing

The Big data conception has been strongly leveraged and have become a major force of innovation across academics, governments and corporate. The paradigm is regarded as an effort to know and obtain data from knowledge (Big data Analytics), providing insights and knowledge over huge datasets. Therefore, it's seen by governments as a way to enhance cities (smart cities [11]) and get proper insights over their people. corporate regard this technology as a way to better understand and perceive their clients, to get closer to them and gain competitive advantage over their competitors. At last, massive knowledge is viewed by scientists as a mean to store and method huge amounts of information like those yielded by CERN's large hadron collider (LHC) in Switzerland [4].

Big data not only concerns the ability to storage huge amounts information however also ways that to method and extract knowledge from it. Table one presents some examples of big data sizes in several domains. In observe, a big knowledge database will contain structured and unstructured knowledge which will come at different velocities, be varied and have different volumes. These are known by the three "V's" of big data. To alternative "V's" – veracity and price – are also vital to clarify that amount is nice however valuable and trustful data also are important. The following paragraphs briefly describe the 5 V's model:

Volume issues the huge loads that generally big data needs to modify. Process and storing massive volumes of information is rather difficult, since it concerns

(among others): scalability (vertical, horizontal or both) so as to facilitate the storage and process power growth; availability, that guarantees access to knowledge and ways that to perform operations over them; and bandwidth and performance, that guarantee the access to knowledge at the right-time.

Variety issues the different} types of knowledge from various sources that big data frameworks have to deal with (typically, completely different sources output different sorts of data). Big data is a way to overcome these differences and unify knowledge. Internet of Things (IoT) could be a big data connected topic that studies knowledge from individual objects of everyday life that may be very varied: internet traffic, smartphones, wearable technology, and others. So as to method varied sorts of knowledge, massive data should provide data-type abstraction frameworks.

Velocity concerns the different rates from every knowledge source. For instance, an Enterprise data Warehouse (EDW) is typically updated once a day, while data from wireless sensor systems is constantly being updated. So as to aggregate knowledge from several knowledge sources, big data should be able to deal that knowledge arriving at different velocities.

Value concerns verity value of information (i.e., the potential value of the data relating to the data they contain). Huge amounts of information are useless if they are doing not provide value for who is exploring it.

Veracity refers to the trustfulness of the data (i.e., it addresses the confidentiality, integrity, and accessibility of the data). Knowledge is meaningless if their source is unreliable. Therefore, organizations got to make sure that the data is correct similarly because the analyses performed on the data are correct.

The 5 V's of massive knowledge complement one another so as to produce solutions that are ready to store and method knowledge a lot of with efficiency.

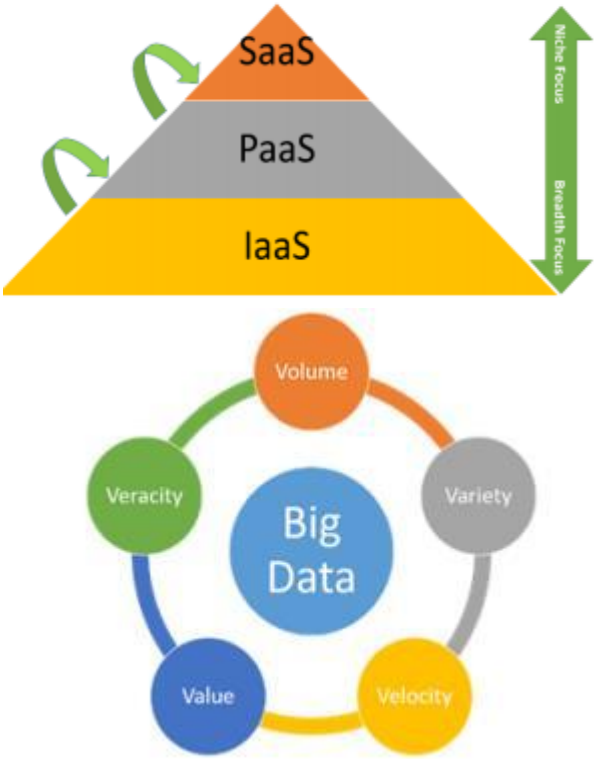
Cloud computing is another fashionable movement that provides theoretically unlimited on-demand services to its users. Cloud's ability to virtualizes resources permits abstracting from hardware, requiring little interaction with cloud providers and swimmingly facultative users to access terabytes of storage, high process power and high availability as a pay-as-you-go model. Moreover cloud computing transfers all costs and responsibilities from the user to the cloud provider, leverage corporations in their early days.

IaaS delivers storage, process power, virtual machines, and so on. From the purpose of read of cost reduction, it is sensible to hire computer power as virtual machines. All that's required could be a few low cost computers to function front-end to access the virtual machines hold on within the cloud. The cloud supplier satisfies the requirements of the client by virtualizing resources according to the service level agreements (SLAs). Some examples of IaaS are Amazon EC21 and Google compute Engine2.

PaaS is built on top of IaaS. The service allows the user to deploy cloud applications created using the programming and runtime environments supported by the provider. Once more, by contracting this service, one is released from server maintenance and software system updates, transferring those concerns to the cloud supplier. Examples of PaaS are Google App Engine3 and Microsoft Azure4.

SaaS is one of the most known cloud models. It consists of applications running directly in the cloud provider. Some of the most used SaaS applications are Google Docs and Drop box. These three basic services are closely related: SaaS is developed over PaaS and ultimately PaaS is built on top of IaaS. Also, from these basic services several others emerged, including Database as a Service (DBaaS) and Big Data as a Service (BDaaS). DBaaS (Database as a Service), as well as BDaaS (Big Data as a Service), usually consist in a SaaS that allows users to hire database services. AaaS (Analytics as a Service) is another service that allows users to hire analytics tools to perform calculations over data.

Since cloud virtualizes resources that are often distributed in clusters or datacenters, it is the most suitable framework for Big Data processing. By virtualizing thousands of machines we can create the high processing power and high storage levels to store and process big amounts of data



Big Data processing

5. Privacy and security issues of cloud computing

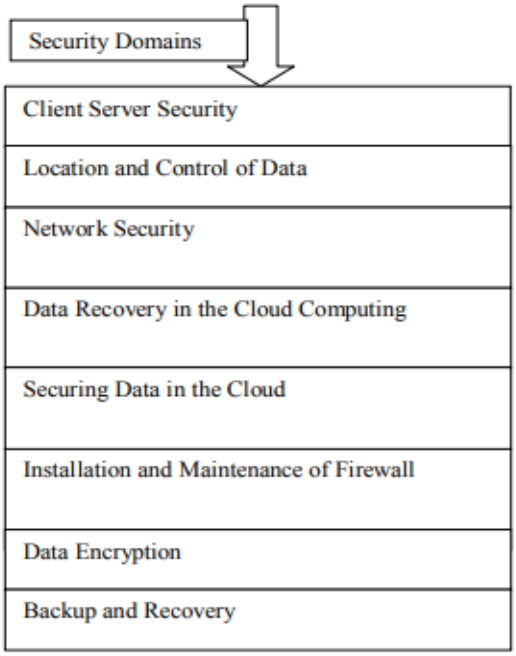
5.1 Privacy issues

1. Compelled Disclosure to the government Cloud can be subject to different levels of protection than on the information it contained
2. Data Security and Disclosure of Breaches: How does cloud provider protect customer's data how can

customer ensure security compliance when storing information on the cloud?

3. Data Accessibility, Transfer and Retention: Can companies and consumers have access to data on cloud? [4]Can the data be destructed by cloud owners or should it be returned to customers?
4. Location of Data: The physical location of the server storing the data may have legal implications

5.2 Security issues



5.3 Client server security

Cloud computing encompasses a client and a server. To maintain secure client, organizations should review existing security practices and use further ones to make sure the safety of its information. Clients should consider secure VPN to attach to the provider. Web browsers are utilized in client side to access cloud computing services. Cloud providers usually offer the customers with APIs that is used by the latter to control, monitor the cloud services. It's important to make sure the safety of those APIs to protect against each accidental and malicious makes an attempt to evade the safety. The varied plug-ins and applications offered within the internet browsers also causes a serious threat to the client systems used to access the provider. Several of the web browsers don't enable automatic updates which can append to the safety concerns. Cloud providers should also incorporate these measures to assure secure transaction among its customers.

5.4 Location and control of data

In traditional data centers business had the privilege to know about the information flow, exact knowledge location, precautions used to protect knowledge from unauthorized access. The physical location' raises the question of legal governance over the information. Another impediment issue is in case of disputes arises between the provider and also the customer. Public cloud has the attraction of cost saving and low maintenance but the enticement comes with a disadvantage. The infra-structure should be shared with unknown people. A cyber invader will act as a subscriber and may spread malicious viruses within the system. It's a responsibility of the provider to check the believability of the customers. The vendor could grant some privileged third parties access to your keep knowledge. The identity of such parties, if any, should be disclosed to the customer. Here, the third party may be a legal authority or even an internal employee. The customer should be informed before the vendor allows third parties to access the stored data. Non cloud services even have security concerns however cloud has additional risk of external party involvement and exposure of critical and confidential knowledge outside organizations management. Modifying security measures or introducing pristine Cloud provider stores the information in provider's side and maintenance is exclusively done by the providers, thus the clients don't have any means to check on the providers security practices, providers employees, their skills specializations etc.

5.5 Network security

Public cloud services are delivered over the internet, exposing the information that was previously secured within the internal firewalls. Applications which individuals used to access at intervals organizations intranet are thus exposed to networking threats and web vulnerabilities which includes distributed denial of service attacks, phishing, malwares and Trojan horses. If an attacker gains access to client credentials, they can eavesdrop on all activities and transactions, manipulate information, falsified information, and redirect clients to illegitimate sites.

5.6 Data recovery in cloud computing

Usually cloud users do not know their data location and the vital query of data recovery in all circumstances may not be possible. The difficulty in retrieving data if there is a change in provider or a need to roll to different platform adds to the apprehension to embrace cloud computing.

5.7 Securing data in the cloud

A Proper implementation of security measures is necessary in cloud computing. The fact that application is launched over the web makes it inclined for security risks. Cloud providers should assume beyond the customary security practices like restricted user access, password protection etc. Physical location of keep knowledge is additionally vital and it's the responsibility of the provider to choose the correct location of storage.

5.8 Installation and maintenance of firewall

Installation of firewall and its maintenance is mandatory to ensure the protection. A firewall ought to be present in all external interfaces. Assessment of firewall policies and rule sets and reconfiguration of router ought to be done in regular intervals. Build and deploy a firewall that denies access from untrusted sources or applications, and adequately logs these events. Build and deploy a firewall that restricts access from systems that have direct external connection and those that contain confidential knowledge or configuration knowledge.

5.9 Data encryption

Data encryption is one common approach the providers to protect their clients' knowledge but the question is whether the data is obtaining stored in encrypted format or not. Several providers follow private/public key coding to confirm knowledge security. To store crucial knowledge organizations will think about private or hybrid cloud wherever the data are in secure corporate firewall.

5.10 Backup and recovery

In cloud computing knowledge is stored in distributed location. Backup computer code ought to include public cloud APIs, enabling simple backup and recovery across major cloud storage vendors, such as Amazon S3, Nirvanix Storage Delivery Network. It's important for the backup application to code confidential knowledge before sending it offsite to the cloud, protective each detain-transit over a WAN to a cloud storage vault and data-at-rest at the cloud storage web site. Customers need to verify that the cloud backup computer code they choose is certified and compliant with the Federal information processing Standards (FIPS) 140 necessities issued by the National Institute of Standards and Technology.

6. Ensuring security against the various types of attacks

Problems associated with the network level security comprise of:

- DNS attacks,
- Sniffer attacks,
- Issue of reused IP address, Denial of Service (DoS) and
- Distributed Denial of Service attacks (DDoS) etc.

6.1 DNS attacks

A Domain Name Server (DNS) server performs the translation of a domain name to a science address. Though victimization DNS security measures like: domain name System Security Extensions (DNSSEC) reduces the consequences of DNS threats however still there are cases once these security measures influence be low once the path between a sender and a receiver gets rerouted through some evil connection. It should happen that even finally the DNS security measures are taken, still the route selected between the sender and receiver cause security problems.

6.2 Sniffer attacks

A sniffer program, through the NIC (Network Interface Card) ensures that the data/traffic linked to different systems on the network additionally gets recorded. It may be achieved by placing the NIC in promiscuous mode and in promiscuous mode it will track all knowledge, flowing on a similar network. A malicious sniffing detection platform based on ARP (address resolution protocol) and RTT (round trip time) may be used to detect a sniffing system running on a network.

6.3 Issue of Reused IP Addresses

Each node of a network is provided an IP address. IP address is largely a finite quantity. A large range of causes related to utilize IP-address issue is observed lately. Once a particular user moves out of a network then the IP-address related to him (earlier) is assigned to a replacement user. This generally risks the protection of the new user as there's a precise time lag between the change of associate IP address in DNS and the clearing of that address in DNS caches. We can say that generally though the recent IP address is being assigned to a new user still the probabilities of accessing the data by another user. It's not negligible because the address still exists within the DNS cache and therefore the information belonging to a particular user may become accessible to another user violating the privacy of the initial user.

6.4 DBGP Prefix Hijacking

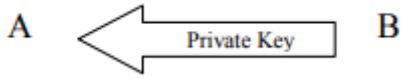
Prefix hijacking could be a kind of network attack during which a wrong announcement related to the IP addresses related to an Autonomous system is formed malicious parties get access to the untraceable IP addresses. On the net, IP area is associated in blocks and remains under the management of AS's. An autonomous system will broadcast data of an IP contained in its regime to all its neighbors'. These ASPs communicate using the Border gateway Protocol (BGP) model. Sometimes, due to some error, a faulty AS could broadcast incorrectly concerning the IPs related to it [7]. In such case, the particular traffic gets routed to some IP other than the intended one. Hence, information is leaked or reaches to another destination that it actually should not.

7. Security against the various types of attacks

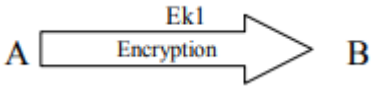
7.1 Symmetric Key Cryptography

It is equally important to secure information in transit and security of transmitted data will be achieved through varied encryption and decryption schemes. In such a situation, though the information gets into the hands of a hacker, he won't be ready to make any unauthorized use till he is aware of a way to decrypt it. Some of the coding-decryption techniques include private and public key encryption. During a symmetric key (private key) coding such as: DES, Triple DES, RC2, RC4 etc, identical secret is used for coding and decryption. Before the information is transferred, the key is shared between each the receiver and also the sender. Sender then sends the information after having encrypted it using the key and also the receiver decrypts it using the same key.

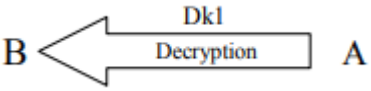
Step.1. Receiver sends its private Key to sender



Step.2. Sender encrypts the Data using sender's Private Key and sends it to Receiver



Step.3. Receiver using his Private Key and Decrypts the same data



7.1 Asymmetric Key Cryptography

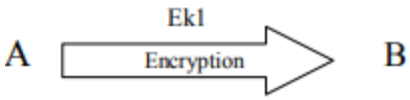
In case of Asymmetric key algorithm (RSA, DSA etc...) there are two types of keys known as Public Key and Private

Key. Public key is common for both sender and receiver and the Private Key is used for decrypts the data from the sender

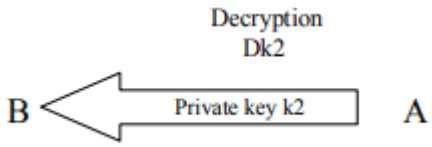
Step.1. Receiver sends its Public key to sender

In Public key encryption bit processing time is more than private key encryption. But the security is more concern rather than the speed, public-key encryption provides more secure data transmission in comparison to private-key encryption. Security issues in a virtualized environment wherein a malicious virtual machine tries to take control of the hypervisor and access the data belonging to other [8].

Step.2. Sender encrypts the Data using sender's Public Key and sends it to Receiver



Step.3. Receiver using his Private Key and Decrypts the same data



8. Conclusion

Cloud computing is artifact of extremely advanced analysis done for virtualization, distributed computing with usages of software and its connected services and also networking. It completely opens a new advanced and secured world of occasions for businesses, however mixed with the offers and high level of security challenges that has to be positively considered once society using the advanced cloud computing ideas. We are presenting the assorted hidden security challenges to be exactly and closely monitor. During this paper we are discussed the intrinsic use of virtual systems as a tool for implementing an improved and advanced cloud environment.

References

[1] X. Cheng, L. Xu, et al., A Novel Big Data Based Telecom Operation Architecture[C], in Proc. 2015 International Conference on Signal and Information Processing(ICSINC), Beijing, China, Oct. 2015

[2] A. A. Cardenas, P. K. Manadhata, et al., Big Data Analytics for Security[J], IEEE Security and Privacy, vol. 11, no. 6, pp. 74-76, Nov.-Dec. 2013.

[3] Cloud Security Alliance Big Data Working Group, Expanded Top Ten Big Data Security and Privacy Challenges[R], Apr. 2013.

[4] B. Matturdi, X. Zhou, et al., Big Data Security and Privacy: A Review [J], China Communications Magazine, 2014

[5] L. Xu, C. Jiang, et al., Information Security in Big Data: Privacy and Data Mining [J]. IEEE Access, vol.2, pp. 1149-1176, Oct. 2014.

[6] E. Sahafizadeh, and M.A.Nematbakhsh, A Survey on Security Issues in Big Data and NoSQL[J]. Advances in Computer Science: an International Journal (ACSIJ), vol.4, no.16, pp.68-72, Jul. 2015.

[7] E. Bertino, Big Data-Security and Privacy[C]. in Proc. 2015 IEEE International Congress on Big Data, Jun. 2015.

[8] K. Saranya, K. Premalatha, et al., A Survey on Privacy Preserving Data Mining[C], in Proc. 2015 IEEE Sponsored 2nd International Conference on Electronics and Communication System (ICECS'15), 2015.

[9] G. P. Silvana and A. Costanzo, "Cloud control and management planes for service provisioning," in Proceedings of the 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, ser. DASC '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 540–546.512

[10] Shyam Patidar; Dheeraj Rane; Pritesh Jain "A Survey Paper on Cloud Computing" in proceeding of Second International Conference on Advanced Computing & Communication Technologies, 2012.

[11] Yashpalsinh Jadeja; Kirit Modi, "Cloud Computing - Concepts, Architecture and Challenges" in Proceeding of International Conference on Computing, Electronics and Electrical Technologies [ICCEET], 2012.

[12] Kuyoro S. O.; Ibikunle F; & Awodele O., "Cloud Computing Security Issues and challenges" in Proceeding of International Journal of Computer Networks (IJCN), Volume (3), Issue (5), 2011.

[13]. Anup H. Gade, A Survey paper on Cloud Computing and its effective utilization with Virtualization, International Journal of Scientific & Engineering Research, Volume 4, Issue 12, December-2013

[14] K Hashizume et al., An analysis of security issues for cloud computing, Journal of Internet Services and Applications, a Springer open journal, pp 1-13, 2013.

[15] Venkata Narasimha Inukollu, Sailaja Arsi ,and Srinivasa Rao Ravuri, SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUD COMPUTING, international Journal of Network Security & Its Applicatis (IJNSA), Vol.6, No.3, May 2014 .

