International Journal of Information Technology and Management

Vol. VII, Issue No. X, November-2014, ISSN 2249-4510

A COMPARATIVE STUDY ABOUT VARIOUS SECURITY AND PRIVACY CHALLENGES OF BIG DATA ON CLOUD COMPUTING

AN INTERNATIONALLY INDEXED PEER REVIEWED & REFEREED JOURNAL

www.ignited.in

International Journal of Information Technology and Management
Vol. VII, Issue No. X, November-2014, ISSN 2249-4510

A Comparative Study about Various Security and Privacy Challenges Of Big Data on Cloud Computing

Charles R.[1] Dr. P. Selva Kumar[2]

[1]Research Scholar, Madurai Kamaraj University, Tamilnadu

Abstract – In this paper, we discuss security issues for cloud computing, Big data, Map Reduce and Hadoop environment. The main focus is on security issues in cloud computing that are associated with big data. Big data applications are a great benefit to organizations, business, companies and many large scale and small scale industries. We also discuss various possible solutions for the issues in cloud computing security and Hadoop. Cloud computing security is developing at a rapid pace which includes computer security, network security, information security, and data privacy.

Cloud computing plays a very vital role in protecting data, applications and the related infrastructure with the help of policies, technologies, controls, and big data tools. Moreover, cloud computing, big data and its applications, advantages are likely to represent the most promising new frontiers in science.

Big Data and cloud computing are two important issues in the recent years, enables computing resources to be provided as Information Technology services with high efficiency and effectiveness. Now a day's big data is one of the most problems that researchers try to solve it and focusing their researches over it to get ride the problem of how big data could be handling in the recent systems and managed with the cloud of computing, and the one of the most important issue is how to gain a perfect security for big data in cloud computing, our paper reviews a Survey of big data with clouds computing security and the mechanisms that used to protect and secure also have a privacy for big data with an available clouds.

----------♦----------

INTRODUCTION

In order to analyze complex data and to identify patterns it is very important to securely store, manage and share large amounts of complex data. Cloud comes with an explicit security challenge, i.e. the data owner might not have any control of where the data is placed. The reason behind this control issue is that if one wants to get the benefits of processing at cloud, he/she must also utilize the allocation of resources and also the scheduling given by the controls. Hence it is required to protect the data in the midst of untrustworthy processes. Since cloud involves extensive complexity, we believe that rather than providing a holistic solution to securing the cloud, it would be ideal to make noteworthy enhancements in securing the cloud that was ultimately provide us with a secure cloud.

Google has introduced MapReduce framework for processing large amounts of data on commodity hardware. Apache's Hadoop distributed file system [HDFS] is evolving as a superior software component for processing at cloud combined along with integrated parts such as MapReduce. Hadoop, which is an open-source implementation of Google MapReduce, including a distributed file system, provides to the application programmer the abstraction of the map and the reduce. With Hadoop it is easier for organizations to get a grip on the large volumes of data being generated each day, but at the same time can also create problems related to security, data access, monitoring, high availability and business continuity.

In this study, we come up with some approaches in providing security. We ought a system that can scale to handle a large number of sites and also be able to process large and massive amounts of data. However, state of the art systems utilizing HDFS and MapReduce are not quite enough/sufficient because of the fact that they do not provide required security measures to protect sensitive data. Moreover, Hadoop framework is used to solve problems and manage data conveniently by using different techniques such as combining the k-means with data mining technology .

Processing at cloud is a technology which depends on sharing of processing resources than having local

servers or personal devices to handle the applications. In Processing at cloud, the word “Cloud” means “The Internet”, so Processing at cloud means a type of processing in which services are delivered through the Internet. The goal of Processing at cloud is to make use of increasing processing power to execute millions of instructions per second. Processing at cloud uses networks of a large group of servers with specialized connections to distribute data processing among the servers. Instead of installing a software suite for each computer, this technology requires to install single software in each computer that allows users to log into a Web-based service and which also hosts all the programs required by the user. There's a significant workload shift, in a processing at cloud system.

Big Data is the word used to de scribe massive volumes of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. The term “Big Data ” is believed to be originated from the Web search companies who had to query loosely structured very large distributed data. The three main terms that signify Big Data have the following properties:

a] Volume: Many factors contribute towards increasing Volume - storing transaction data, live streaming data and data collects from sensors etc.,

b] Variety: Today data comes in all types of formats – from traditional databases, text documents, emails, video, audio, transaction s etc.,

c] Velocity: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand. The other two dimensions that need to consider with respect to Big Data are Variability and Complexity .

d] Variability: Along with the Velocity, the data flows can be highly inconsistent with periodic peaks.

Complexity: Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked,matched, cleansed and transformed into required formats before actual processing.

Big data is known as a datasets with size beyond the ability of the software tools that used today to manage and process the data within a dedicated time. With Variety, Volume, Velocity Big Data such military data or other unauthorized data need to be protected in a scalable and efficient way . Information privacy and security is one of most concerned issues for Cloud Computing due to its open environment with very limited user side control. It is also an important challenge for Big Data. After few years later more data globally would be touched with Cloud Computing which provides strong storage, computation and distributed capability in support of Big Data processing. Other considerations are that information privacy and security challenges in both Cloud Computing and Big Data must be investigated. the privacy and security providing such forum for researchers, and developers to exchange the latest experience, research ideas and development on fundamental issues and applications about security and privacy issues in cloud and big data environments.

The cloud helps organizations and enables rapid on demand provisioning of server resources such as CPUs, manage, storage, bandwidth, and share and analyze their Big Data in a reasonable and simple to use. The cloud infrastructure as a service platform, supported by on demand analytics solution seller that makes the large size of data analytics very affordable. As location independent cloud computing Involving shared services providing resources , software and data to systems and The hardware on demand, actually the storage networking in cloud is a very strong because use driver for high performance.

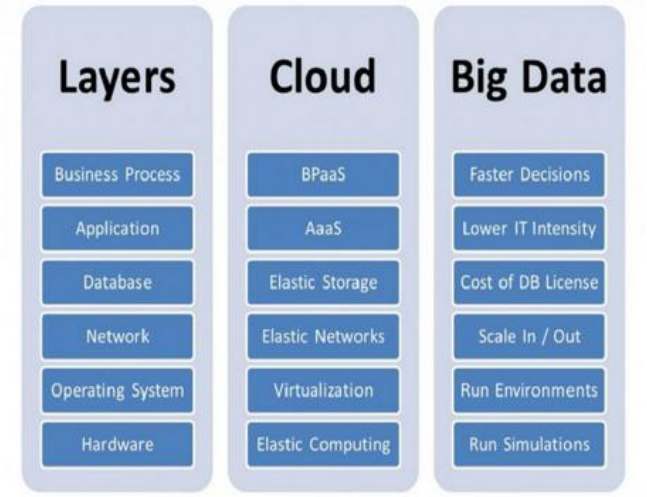

Fig. 1. Big Data and clouds

For example Arista provides Networks with Specifications and product line of switching solutions as shown in figures 2 and 3 below. However, the requirements of cloud storage needs hypothesized to a group of sub nodes operations performed with some of the units and CPUs advanced.

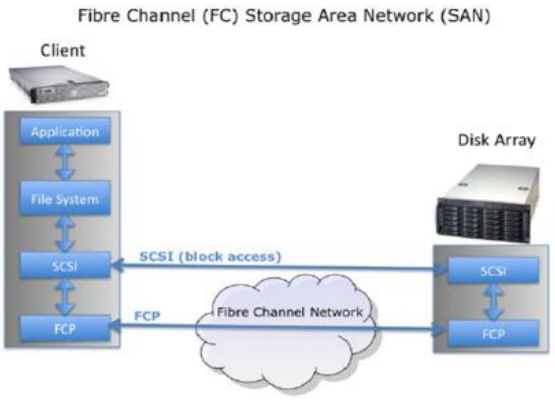

Fig. 2. Fiber Channel Storage area Network

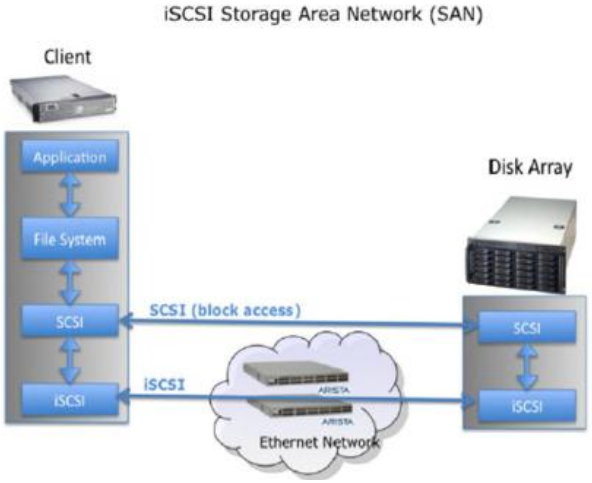

Fig. 3. iSCSI storage area network

CLOUD COMPUTING

Cloud Computing is a technology which depends on sharing of computing resources than havinglocal servers or personal devices to handle the applications. Computing means a type of computing in which services are delivered through the Internet. The services are Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS). The goal of Cloud Computing is to make use of increasing computing power to execute millions of instructions per second. Cloud Computing uses networks of a large group of servers with specialized connections to distribute data processing among the servers. Instead of installing a software suite for each computer, this technology requires to install a single software in each computer that allows users to log into a Web-based service and which also hosts all the programs required by the user. Cloud computing technology is being used to minimize the usage cost of computing resources. The cloud network, consisting of a network of computers also the costof software and hardware on the user decreases.

Cloud computing architecture refers to the components and subcomponents required for cloud computing. These components typically consists of a front end platform (fat client, thin client, mobile device), back end platforms (servers, storage), a cloud based delivery, and a network (Internet, Intranet, Inter-cloud). Combined, these components make up cloud computing architecture.

CLOUD COMPUTING IN BIG DATA

The rise of cloud computing and cloud data stores has been a precursor and facilitator to the emergence of big data. Cloud computing is the commodification of computing time and data storage by means of standardized technologies. It has significant advantages over traditional physical deployments. However, cloud platforms come in several forms and sometimes have to be integrated with traditional architectures. This leads to confuse for decision makers in charge of big data projects, leads to a question of how and which cloud computing is the optimal choice for their computing needs, especially if it is a big data project? These projects regularly exhibit unpredictable, bursting, or immense computing power and storage needs.

At the same time business stakeholders expect swift, inexpensive, and dependable products and project outcomes. This article introduces cloud computing and cloud storage, the core cloud architectures, and discusses what to look for and how to get started with cloud computing.

BIG DATA CLOUD PROVIDERS

Cloud providers come in all shapes and sizes and offer many different products for big data. Some are household names while others are recently emerging. Some of the cloud providers that offer IaaS services that can be used for big data include Amazon.com, AT&T, GoGrid, Joyent, Rackspace, IBM, and Verizon/Terremark. Currently, one of the most high-profile IaaS service providers is Amazon web Services with its Elastic Compute Cloud (Amazon EC2). Amazon didn't start out with a vision to build a big infrastructure services business.

The cloud computing space has been dominated by Amazon Web Services until recently. Increasingly serious alternatives are emerging like Google Cloud Platform and other clouds that mentioned above. Amazon Web Services compatible solutions, i.e. Amazon's own offering or companies with application programming interface compatible offerings, and Open Stack, an open source project with a wide industry backing. Consequently, the choice of a cloud platform standard has implications on which tools are

available and which alternative providers with the same technology are available.

Amazon EC2 offers scalability under the user's control, with the user paying for resources by the hour. The use of the term elastic in the naming of Amazon's EC2 is significant. Here, elasticity refers to the capability that the EC2 users have to increase or decrease the infrastructure resources assigned to meet their needs. Amazon also offers other big data services to customers of its Amazon web Services portfolio. These include the following : Amazon Elastic MapReduce: Targeted for processing huge volumes of data. Elastic MapReduce utilizes a hosted Hadoop framework running on EC2 and Amazon Simple Storage Service (Amazon S3). Users can now run HBase.

Amazon DynamoDB: A fully managed not only SQL (NoSQL) database service. DynamoDB is a fault tolerant, highly available data storage service offering selfprovisioning, transparent scalability, and simple administration. It is implemented on SSDs (solid state disks) for greater reliability and high performance.

Amazon Simple Storage Service (S3): A web-scale service designed to store any amount of data. The strength of its design center is performance and scalability, so it is not as feature laden as other data stores. Data is stored in "buckets" and you can select one or more global regions for physical storage to address latency or regulatory needs.

Amazon High Performance Computing: Tuned for specialized tasks, this service provides low-latency tuned high performance computing clusters. Most often used by scientists and academics, HPC is entering the mainstream because of the offering of Amazon and other HPC providers. Amazon HPC clusters are purpose built for specific workloads and can be reconfigured easily for new tasks.

Amazon RedShift: Available in limited preview, RedShift is a petabyte-scale data warehousing service built on a scalable MPP architecture. Managed by Amazon, it offers a secure, reliable alternative to in-house data warehouses and is compatible with several popular business intelligence tools.

SECURITIES ISSUES AND CHALLENGES

Cloud computing comes with numerous security issues because it encompasses many technologies including networks, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, concurrency control and memory management. Hence, security issues of these systems and technologies are applicable to cloud computing. For example,it is very important for the network which interconnects the systems in a cloud to be secure. Also, virtualization paradigm in cloud computing results in several security concerns. For example, mapping of the virtual machines to the physical machines has to be performed very securely.

Data security not only involves the encryption of the data, but also ensures that appropriate policies are enforced for data sharing. In addition, resource allocation and memory management algorithms also have to be secure. The big data issues are most acutely felt in certain industries, such as telecoms, web marketing and advertising, retail and financial services, and certain government activities. The data explosion is going to make life difficult in many industries, and the companies will gain considerable advantage which is capable to adapt well and gain the ability to analyze such data explosions over those other companies. Finally, data mining techniques can be used in the malware detection in clouds.

The challengesof security in cloud computing environments can be categorized into network level, user authentication level, data level, and genericissues.

Network level: The challenges that can be categorized under a network level deal with network protocols and network security, such as distributed nodes, distributed data, Internode communication.

Authentication level: The challenges that can be categorized under user authentication level deals with encryption/decryption techniques, authentication methods such asadministrative rights for nodes, authentication of applications and nodes, andlogging.

Data level: The challenges that can be categorized under data level deals with data integrity and availability such as data protection and distributed data.

Generic types: The challenges that can be categorized under general level are traditional security tools, anduse of different technologies

Distributed Nodes-

Distributed nodes are an architectural issue. The computation is done in any set of nodes. Basically, data is processed in those nodes which have the necessary resources. Since it can happen anywhere across the clusters, it is very difficult to find the exact location of computation. Because of this it is very difficult to ensure the security of the place where computation is done.

Distributed Data-

In order to alleviate parallel computation,a large data set can be stored in many pieces across many machines. Also, redundant copies of data are made to ensure data reliability. In case a particular chunk is corrupted, the data can be retrieved from its copies. In the cloud environment, it is extremely difficult to find exactly where pieces of a file are stored. Also, these pieces of data are copied to another node/machines based on availability and

maintenance operations. In traditional centralized data security system, critical data is wrapped around various security tools. This cannot be applied to cloud environments since all related data are not presentedin one place and it changes.

Internode Communication-

Much Hadoop distributions use RPC over TCP/IP for user data/operational data transfer between nodes. This happens over a network, distributed around globe consisting of wireless and wired networks. Therefore, anyone can tap and modify the inter node communication for breaking into systems.

Data Protection-

Many cloud environments like Hadoop store the data as it is without encryption to improve efficiency. If a hacker can access a set of machines, there is no way to stop him to steal the critical data present in those machines.

Administrative Rights for Nodes-

A node has administrative rights and can access any data. This uncontrolled access to any data is very dangerous as a malicious node can steal or manipulate critical user data.

3.6 Authentication of Applications and Nodes-

Nodes can join clusters to increase the parallel operations. In case of no authentication, third part nodes can join clusters to steal user data or disrupt the operations of the cluster.

Logging-

In the absence of logging in a cloud environment, no activity is recorded which modify or delete user data. No information is stored like which nodes have joined cluster, which Map Reduce jobs have run, what changes are made because of these jobs. In the absence of these logs, it is very difficult to find if someone has breached the cluster if any, malicious altering of data is done which needs to be reverted. Also, in the absence of logs, internal users can do malicious data manipulations without getting caught.

Traditional Security Tools-

Traditional security tools are designed for traditional systems where scalability is not huge as cloud environment. Because of this, traditional security tools which are developed over years cannot be directly applied to this distributed form of cloud computing and these tools do not scale as well as the cloud scales.

Use of Different Technologies-

Cloud consists of various technologies which has many interacting complex components. Components include database, computing power, network, and many other stuff. Because of the wide use of technologies, a small security weakness in one component can bring down the whole system. Because of this diversity, maintaining security in the cloud is very challenging.

BIG DATA PRIVACY AND SECURITY

Big Data remains one of the most talked about technology trends in 2013. But lost among all the excitement about the potential of Big Data are the very real security and privacy challenges that threaten to slow this momentum. Security and privacy issues are magnified by the three V's of big data: Velocity, Volume, and Variety. These factors include variables such as large-scale cloud infrastructures, diversity of data sources and formats, streaming nature of data acquisition and the increasingly high volume of intercloud migrations. Consequently, traditional security mechanisms, which are tailored to securing small-scale static (as opposed to streaming) data, often fall short.

The CSA's Big Data Working Group followed a three step process to arrive at top security and privacy challenges presented by Big Data; Interviewed CSA members and surveyed security practitioner oriented trade journals to draft an initial list of high priority security and privacy problems studied published solutions. Characterized a problem as a challenge if the proposed solution does not cover the problem scenarios. Following this exercise, the Working Group researchers compiled their list of the Top 10 challenges as shown in figure 4 below. The Expanded Top 10 Big Data challenges have evolved from the initial list of challenges presented at CSA Congress to an expanded version that addresses three new distinct issues :

- Modeling: formalizing a threat model that covers most of the cyber-attack or data-leakage scenarios.
- Analysis: finding tractable solutions based on the threat model.
- Implementation: implanting the solution in existing infrastructures.

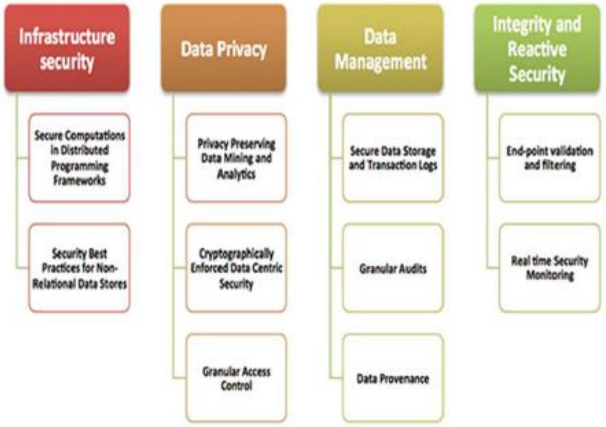

Fig.4. 10 Challenges of CSA's Big Data Working Group

The information security practitioners at the Cloud Security Alliance know that big data and analytics systems are here to stay. They also agree on the big questions that come next: How can we make the systems that store and compute the data secure? And, how can we ensure private data stays private as it moves through different stages of analysis, input and output? The answers to those questions that prompted the group's latest 39-page report detailing 10 major security and privacy challenges facing infrastructure providers and customers. By outlining the issues involved, along with analysis of internal and external threats and summaries of current approaches to mitigating those risks, the alliance's members hope to prod technology vendors, academic researchers and practitioners to collaborate on computing techniques and business practices that reduce the risks associated with analyzing massive datasets using innovative data analytics.

Existing encryption technologies that don't scale well to large datasets. Real-time system monitoring techniques that works well on smaller volumes of data but not very large datasets. The growing number of devices, from smartphones to sensors, producing data for analysis. General confusion "surrounding the diverse legal and policy restrictions that lead to ad hoc approaches for ensuring security and privacy .

APPROACHES TO IMPROVE THE SECURITY OF PROCESSING AT CLOUD ENVIRONMENT:

We present various security measures which would improve the security of processing at cloud environment. Since the cloud environment is a mixture of many different technologies, we propose various solutions which collectively wasmake the environment secure. The proposed solutions encourage the use of multiple technologies/ tools to mitigate the security problem specified in previous segments. Security recommendations are designed such that they do not decrease the efficiency and scaling of cloud systems.

Following security measures should be taken to ensure the security in a cloud environment.

- **File Encryption**

Since the data is present in the machines in a cluster, a hacker can steal all the critical information. Therefore, all the data stored should be encrypted. Different encryption keys should be used on different machines and the key information should be stored centrally behind strong firewalls. This way, even if a hacker is able to get the data, he cannot extract meaningful information from it and misuse it. User data was be stored securely in an encrypted manner.

- **Network Encryption**

All the network communication should be encrypted as per industry standards. The RPC procedure calls which take place should happen over SSL so that even if a hacker can tap into network communication packets, he cannot extract useful information or manipulate packets.

- **Logging**

All the map reduce jobs which modify the data should be logged. Also, the information of users, which are responsible for those jobs should be logged. These logs should be audited regularly to find if any, malicious operations are performed or any malicious user is manipulating the data in the nodes.

- **Software Format and Node Maintenance**

Nodes which run the software should be formatted regularly to eliminate any virus present. All the application software's and Hadoop software should be updated to make the system more secure.

- **Nodes Authentication**

Whenever a node joins a cluster, it should be authenticated. In case of a malicious node, it should not be allowed to join the cluster. Authentication techniques like Kerberos can be used to validate the authorized nodes from malicious ones.

- **Rigorous System Testing of Map Reduce Jobs**

After a developer writes a map reduce job, it should be thoroughly tested in a distributed environment instead of a single machine to ensure the robustness and stability of the job.

- **Honeypot Nodes**

Honey pot nodes should be present in the cluster, which appear like a regular node but is a trap. These honeypots trap the hackers and necessary actions would be taken to eliminate hackers.

➢ **Layered Framework for Assuring Cloud**

A layered framework for assuring processing at cloud as shown in Figure 5, consists of the secure virtual machine layer, secure cloud storage layer, secure cloud data layer, and the secure virtual network monitor layer. Cross cutting services are rendered by the policy layer, the cloud monitoring layer, the reliability layer and the risk analysis layer.

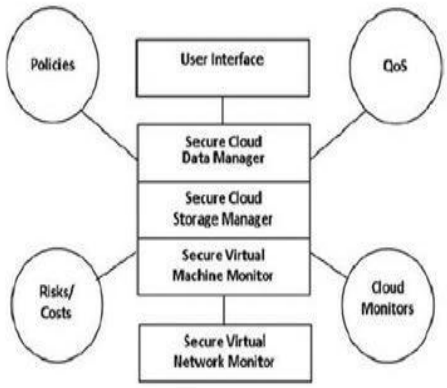

Fig 5: Layered framework for assuring cloud.

CONCLUSION

Cloud environment is widely used in industry and research aspects; therefore security is an important aspect for organizations running on these cloud environments. Using proposed approaches, cloud environments can be secured for complex business operations.

Recently, researchers focusing their efforts in how to manage, handling and also processing the huge amount of data as known a big data deals with three concepts volume, Variety and velocity which requires a new mechanisms to manage, processing, storing, analyzing and securing the big data. as managing and processing of big data have many problems and required more efforts to handle these requirements when deal with big data, security is one of the challenges that arise when systems try to handle the concept of big data. More researches required to overcome the security of big data instead of current security algorithms and methods.

Cloud environment is widely used in industry and research aspects; therefore security is animportant aspect for organizations running on these cloud environments. Using proposed approaches, cloud environments can be secured for complex business operations.

REFERENCES

- "Security-Enhanced Linux."*Security-Enhanced Linux*. N.p. Web. 13 Dec 2013.
- A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.
- N, Gonzalez, Miers C, Redigolo F, Carvalho T, Simplicio M, de Sousa G.T, and Pourzandi M. "A Quantitative Analysis of Current Security Concerns and Solutions for Cloud Computing.". Athens: 2011., pp 231 – 238, Nov. 29 2011- Dec. 1 2011
- N, Gonzalez, Miers C, Redigolo F, Carvalho T, Simplicio M, de Sousa G.T, and Pourzandi M. "A Quantitative Analysis of Current Security Concerns and Solutions for Processing at cloud.". Athens: 2011., pp 231 – 238, Nov. 29 2011- Dec. 1 2011
- P.R , Anisha, Kishor Kumar Reddy C, Srinivasulu Reddy K, and Surender Reddy S. "Third Party Data Protection Applied To Cloud and Xacml Implementation in the Hadoop Environment With Sparql."2012. 39-46, Jul – Aug. 2012.
- Venkata Narasimha Inukollu, Sailaja Arsi ,and Srinivasa Rao Ravuri, SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUD COMPUTING, nternational Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014 .
- Wie, Jiang , Ravi V.T, And Agrawal G. "A Map-Reduce System With An Alternate Api For Multi-CoreEnvironments.". Melbourne, Vic: 2010, Pp. 84-93, 17-20 May. 2010.International Journal Of Network Security & Its Applications (Ijnsa), Vol.6, No.3, May 2014
- Zhao, Yaxiong , and Jie Wu. "Dache: A data aware caching for bigdata applications using the MapReduce framework." *INFOCOM, 2013 Proceedings IEEE,* Turin, Apr 14-19, 2013, pp. 35 - 39.