SECURITY FOR CLOUD COMPUTING USING MAP REDUCING FRAMEWORK WITH KEY GENERATION TECHNIQUE FOR A HADOOP ENVIRONMENT

U. Prathibha [1], Dr. E.J. Thomson Fredrick [2]

ABSTRACT

Cloud Computing is an emerging technology in today's world. Cloud computing is an efficient technology for storage and retrieval of data in the fastest mode. The main issue in cloud computing is data security. This paper proposes the bigdata technology with the Hadoop Distributed Framework for storing huge data in cloud in ahighly efficient manner. The distributed Environment demands secure computing to access personal or sensitive data,using Hadoop Distributed File System(HDFS).The quick,automatic fault detection and recovery mechanisms are available in HDFS.In order to avoid third party issues and produce unrelated data to the selective block in the cluster,the study proposes a Hadoop Distributed File System using, key generation techniques for Data Replication to avoid the issues in the Data replica (i.e., copies of distributed files) and corruption data.

Keywords: Cloud Security, Bigdata, Hadoop, HDFS, Map Reduce, Key generation

1. INTRODUCTION

Cloud computing is a technology which completely depends on Internet where data is stored and maintained by cloud service providers such as Google, Amazon, Salesforce.som Microsoft etc. Cloud computing creates a lot of security issues and a threat consists of data leakage, insecure interface, resource sharing, and interruptions.The security of cloud computing must be rectified by the users in order to make the cloud environment trustworthy. In cloud environment,alarge amount of data, big data, need to be stored and handled.Big data involvesstorage, retrieval and modification of huge amount of data. Hadoop Software platform is a very user-friendly software platformfor the purpose. Hadoop framework includes MapReduce - offline computing engine and Hadoop Distributed File System.Because map reduce donot depend upon user authentication, an unwanted user can adapt the job priorities of Hadoop and make the job to be completed faster or worse, create security issues in other jobs or kill the other jobs.Replication is defined similar blocks of data on different nodes and replication factor is represented by the number of replicas in every single block of data in a cluster.After the completion of replica, the data node sends a block report to a name node.HDFS can detect corruption in data replicacaused by security issues andthe name node will reschedule the re-replication work to restore the desired number of replicas by copying from another node with a known good replica.The failure process is undetected for long time until other replica fails. This paper proposes the key generation techniques to avoid these issues.

[1]Assistant Professor, Department of CS,CA & IT Karpagam Academy of Higher Education, Coimbatore, prathibha.prathi4@gmail.com

[2]Associate Professor, Department of CS,CA & IT, Karpagam Academy of Higher Education, Coimbatore, thomson500@gmail.com

2. LITERATURE REVIEW

This Paper analyzed issues on cloud computing security with Hadoop environment.In [1], S. Chandra Mouliswaranet al, 2012 explains that Hadoop coordinates the work among the cluster of machine and it focuses on how the replicas are managed in HDFS for providing high availability of data under extreme computational requirement.In[2], Dai Yuefaet al, 2009 analyzed the basic issue of cloud computing in data security. The data security requirement of cloud computing is obtained and a mathematical data model is set up for cloud computing by analyzing HDFS Architecture.In[3], Dharmik H. Patel and S.N.Gujar, 2016 describe the system with authentication, auditing, authorization and encryption within the cluster. In [4], Gurpreet Kaur and Manpreet Kaur, 2015explain the centralizing data acquisition and consolidation in the cloud, andhow by using cloud based virtualization infrastructure to mine data sets efficiently.Big-data methods provide new insight into extent data sets and the various techniques. Big-data technologies have been implemented for manipulating, analyzing and visualizing the big data.

In [5], Harin C Naik and Divyesh Joshi 2016, describe the uses and evolution of Hadoop and thebasic concepts of distributed file systems to manage the nodes and implementation of map reduce for programming pattern.In [6],Iqbaldeep Kaur et al, 2016have done research on Hadoop for using large datasets in distributed processing and to solve the problem for big data analytics.In [7], Karthik D and et al, 2015 proposed the Hadoop to solve the current data security problem for cloud disk in distributed network, and it provided a particular encryption method to include different secret level of client data for avoiding the data leakage in the cloud storage. It combined with symmetric encryption algorithm and uniqueness of client data authentication of RSA and the performance of Hadoop and the distributed network cloud data security storage disk can provide and access the protected, efficient data.

In [8], Monjur Ahmed and Mohammad Ashraf Hossain, 2014 described that cloud computing could face a lot of security challenges and possibilities occursimultaneously. The little credibility offer to the cloud computing when the security is not consistent, robust and flexibility. It focused on the security issues occur in cloud computing and cloud infrastructure.

In[9],MuralikrishnanRamane, et al, 2014 implemented an adaptive data replication scheme that depended on access word count prediction using Langrange's interpolation. Itwas adapted to fit the criteria and proved that itprocessed on a rack aware cluster setup which remarkably reduced the task completion time.But, once the volume of the data being processed increased there was a considerable cutback in computational speeds due to update cost. Optimizing thethreshold level for abalance between the update cost and replication factor is identified and presented graphically. These types of techniques are presentedin this paper for data replication issues.In [10], NavjotSekhon and Richa Mahajan,2017attempted to increase the security services and storage services by using Hadoop Distributed File System. The multi location storage data and services create lot of security issues. For that issue, the solution will be provided on HDFS Master and Slave node Architectureto produce the highly reliable performance.

In [11] PedroCaldeiraNeves and Jorge Bernardino A.C, 2015described the current issues of bothcloud and big

data technologies to implement the Big Data Support Systems, which host themand process successfully using cloud computing.In[12], Pooja. D.Bardiyaetal,2014analyzed that HDFS Framework required data security in cloud computing and highlighted data security as the major issue in cloud computing.In [13], Rabi Prasad Padhyet al, 2011described the various security issues and attacks, which include leakage of data, resource sharing, insecure interface, inside attacks and data availability arising because of the limited control of data. This paper also demonstrates the various cloud models and the main security issues and risks that are currently present within the cloud computing sector and also analyzes the challenges and key research in cloud computing.

In [14], Sharifnawaj Y. Inamdaret al, 2016 demonstrated that Hadoopwasakey to execute the security of client information in Hadoop Systems.In [15], Xuebin Chen et al, 2015 describedhow to ensure the usage and mass storage of data without any lag problem by designing the application platform for convenience. By the adoption of master slave distribution,a storage system with better performance is designed through the introduction of cloud computing technology. In [16], VijyendraKarpatne and E.J.Thomson Fredrik,2017 proposed anauthentication process in which a third party authenticator validates the authentication process. The authentication process is supported by a key seed mechanism.In [17], M. Cimi Thomas and S.Sheeja, 2017 proposed Elliptic Curve Cryptography (ECC) and its application in the Secure Socket Layer (SSL) for web applications. ECC is an asymmetric key encryption method which offers high security with smaller key sizes and it is used in the security protocols without degrading the performance of the web servers.

3. BIGDATAAND HADOOP

3.1 BIG DATA

Big data technology is used to store a huge amount of data. It is an innovative technology for both academic and corporate sectors. Security and privacy issues are determined by the Volume, Variety, Velocity, Value and Veracity of big data

Volume -Volume is defined to store huge amount of data can access to perform operations.

Variety -Variety deals with the different types of data from various sources within big data frameworks

Velocity - Velocity refers to big data systems which can retrieve and store data independently at different rates in which flow of data may occur in or out of the system and provide an abstraction layer

Value - Value defines the exact value of data i.e., the efficient value of the data for given information. If they are not provided withan exact value, all the data are worthless.

Veracity-Veracity is considered the trustworthy data, addressing data confidentiality, data integrity, and data availability.

3.2 Hadoop

Hadoop is a software framework which stores a huge amount of data and processes it.

Scalable: It store and process huge data like petabytes reliably.

Economical: It distributes the data and processing across clusters of commonly available computers (in thousands).

Efficient: By distributing the data, it can process in parallel on the nodes where the data is located.

Reliable: It automatically maintains multiple copies of

data and automatically redeploys computing tasks based on failures.The data can be managed with Hadoop to distribute the data and duplicates chunk of each data file across several nodes.

3.2.1 Features of Hadoop

Hadoop is optimized to handle massive quantities of various types of data. Hadoop replicates data across multiple computers. It provides high throughput with low latency.It complements both OLTP and OLAP.It is not good for processing small files because it stores a huge amount of data and not for parallelized data.

3.3 HDFS Daemons

Daemons mean "Background process".

" Name node

" Data Node

" Secondary name node

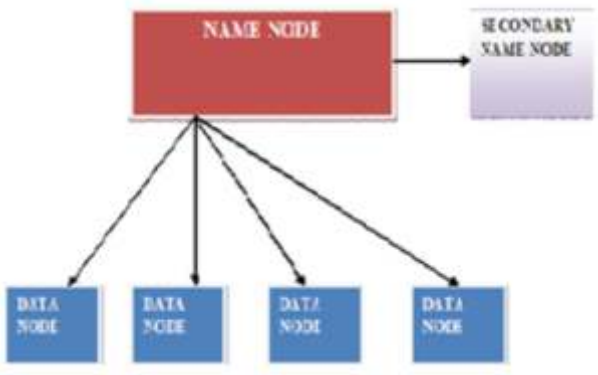

Fig 1: HDFS Daemons

3.3.1HDFS Daemons - Name Node(NN)

Name Node is the 'master' machine and controls all the Meta data for the cluster. For e.g.:Creation of blocks in the file and storage of nodes in a block. HDFS breaks large data into smaller pieces called Blocks. The size of the default block is 64MB.NN uses RACKID identity. Rack is a collection of data nodes within cluster. NN keeps track of the blocks of a file as it is placed on various Data nodes. NN manages file related operations such as read, write, create and delete. Its main job is managing the File System Namespace.

3.3.2 File System Namespace

File system namespace refersto a collection of files in cluster. It includes mapping of blocks to file, file prop655ertiesand is stored in a file called FSImage.HDFS supports a traditional hierarchical file organization. A user or an application can create directories and store files inside these directories. The file system namespace hierarchy is similar to most other existing file systems. One can create and remove files, move a file from one directory to another, or rename a file. The Name Node maintains the file system namespace. Any change to the file system namespace or its properties are recorded by the Name Node. An application can specify the number of replicas of a file that should be maintained by HDFS. The number of copies of a file is called the replication factor ofthat file. This information is stored by the Name Node. HDFS stores multiple data nodes per cluster. It stores each block of HDFS data in a separate file. It performs a Read/Write operation with Name and Data node communications.

3.4 MapReduce Programming

MapReduce Programming is a software frame work. Map Reduce Programming helps people to process massive amounts of data in parallel. It provides a, Key-value pair, Job Tracker (master) /Cluster, Task Tracker (slave)/Node.

Job Configuration: Application and Job parameters,Interaction between Job tracker and task tracker

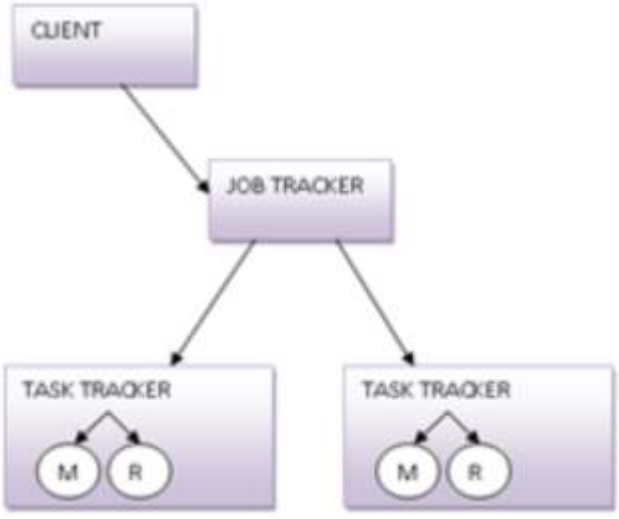

Fig 2 : Interaction between job tracker and task tracker

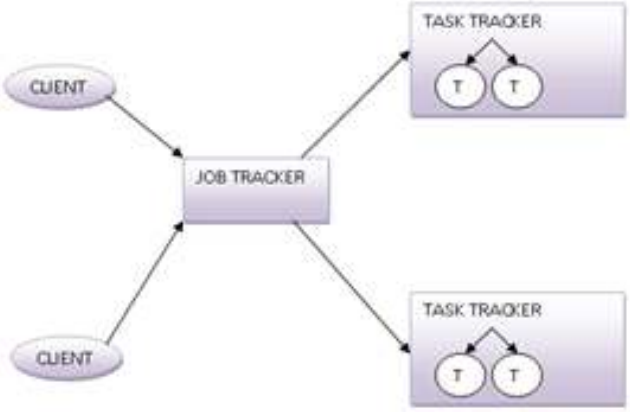

Fig 3 : Map Reduce Programming Architecture

Input: Text file

Driver class: Job configuration details

Mapper class: Overrides Map function based on the problem statement

Reducer class: Overrides Reduce function based on the problem statement.

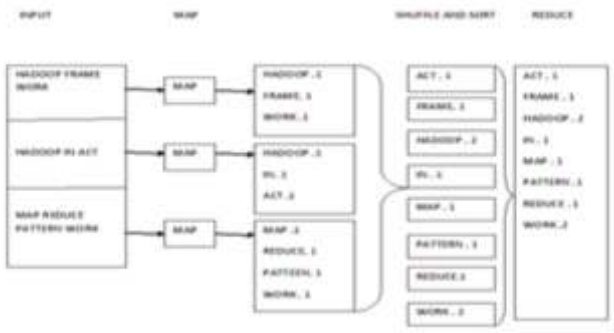

Fig 4: Map Reduce Work

The map task is done by Mapper class and the reduce task is done by reducer class. Input data set is split into multiple pieces of data. The framework creates several master and slave processes. There are several map tasks that work simultaneously. Map workers use partitioner function to divide the data into regions. Once the map is completed reducer work begins. Mappers class tokenizing the given input and sort it out. In next step reducing process reduces the matching pairs and produces the perfect output

4. ISSUES OF CLOUD SECURITY IN BIGDATA

In cloud computing, security is defined to be a crucial barrier in its path to success. In the cloud computing environment personal data security is a crucial concern. The security of data stored on cloud need to be increased. Most of the security issues in the cloud computing are listed and explained below.

Transparency of Location

Location of data storage is the main process in data security.Data protection issues and authentication problem may occur without the exact location of the data storage being known.

Denial of Service

It can be a potential or serious problem for cloud computing. In cloud computing infrastructure, it has

been the major common attack until now and there is no way to mitigate this type of problem.

Data Protection

The Cloud infrastructure shares the resources withmultiple users at any point of time. Any threat inside the cloud can interrupt the user data, which may cause some security issues like lack of data storage, data leakage etc.. Data protection in the cloud environment is more important.

Multi-Tenancy

The multiple cloud users share the resource in a single cloud. It createsa lot of security issues for isolation such as virtualization and resource management.

Network Security

With the creation of virtual servers (invisible networks) it is very difficult to track the network traffic and performance.When users access the cloud, the cloud environmentfollows some policies to access the data stored on cloud.These security policies must be followed by cloud to avoid any unauthorized access.

Data confidentiality issue

In cloud computing, users can store their data and information on remote servers owned by others or accessed through the internet. The data confidentiality issues are raised, when a government agency or any other entity shares the information stored on cloud.

Data Loss

If the cloud service provider is interrupted with any malicious attack, the cloud data can be lost. If the data is lost, there is no recovery plan for the lost information.

Data Privacy

Data privacy is the main security issuein Cloud computing. It is important for users to store their private or confidential data in the cloud. By involving a trusted third party, there is a chance of heterogeneity of users which affects security in the cloud.

Data Integrity

The essential feature of cloud storage is integrating monitoring.It can be defined as ensuring that the data is unaltered, correct and consistent. Data Integrity can be hampered at any level of storage.

Data Availability

Data availability is one of the prime concerns of the mission and safety critical organizations. When keeping data at remote systems owned by others, data owners may suffer from system failures of the service provider.

Insider Threat

An attack which happens inside the organizations may bereferred to as a threat. In cloud based services, all the employees have an individual authorization. Attacker will misuse theinformation such as customer accounts, financial forms and sensitive information.

Security solutions provided on the HDFS architecture has easy access and works with applications efficiently. It providesa parallel process to the node in a cluster and is fast and high reliable. In the cluster the replica chunks of data to node is highly reliable but not confidential. This leads to a poor maintenance of confidential data. To avoid these security issues, key generation techniquesis proposed in this paper.

5. KEY GENERATION TECHNIQUE TO OVERCOME CLOUD SECURITY ISSUES

In this proposed work, the cloud user implements HDFS framework. When the cloud user sends the file to the HDFS Cluster, HDFS breaks large data into smaller

pieces called Blocks.This study proposes to generate a key for Data replicas. So, this technique is used when the data is distributed. Data can be retrieved from the data replica files because each node can create three replica files for each node.

Data Replication: There is absolutely no need for a client application to track all blocks. It directs the client to the nearest replica to ensure high performance.

The number of copies of a file is called the replication factor of that file.It stores each block of HDFS data in a separate file.

As per the Hadoop Replica Placement Strategy, first replica is placed on the same node as the client. Then it places second replica on a node that is present on a different rack. It places the third replica on the same rack as second, but on a different node in the rack. Once replica locations have been set, a pipeline is built. This strategy provides good reliability

The data corruption on disk failure is secured against using replication. If a replica is corrupt due to any technical issues, then the other replicas can be used.

If the data node does not revert the heartbeats to the name node, then after a certain period of time, it declares the node as dead (i.e., not working node). At the same time it will re-replicate the given blocks that are on the failed data node using replicas stored on other nodes of the cluster.

Failure issues are reported to the name node, which organizes re-replication of the perfect replicas. Because HDFS is frequently used to store data that does not read very often, corrupt data is detected by undesirable read mode. If the failure process is undetected for a long time, other replicas also fail.The study proposes the key generation techniques to avoid the replica problem, and the process is illustrated and defined in Fig 5, given below.

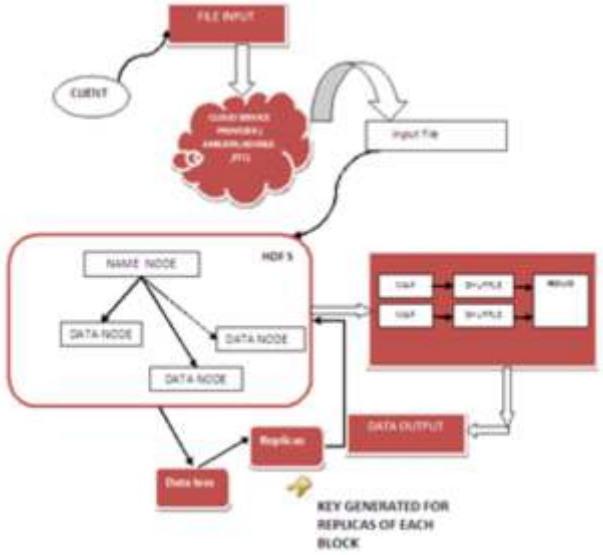

Fig 5: HDFS Secured data System and replication

5.1 KEY GENERATION FOR REPLICAS

In order to avoid the problem of identifying the copies of file, key generation techniques are used in a Hadoop framework. In the Data Pipeline, client application writes a block to the first Data Nodein the pipeline. Then this Data Node takes over and forwards the data to the next node in the pipeline. This process continues for all the data blocks, and subsequently all the replicas are written to the disk.

STEPS FOLLOWED IN KEY GENERATION TECHNIQUE

STEP 1: **Begin the process.**

STEP 2: Get an access token from cloud service provider (i.e., cloud server).

STEP 3: Client is redirected to the process by cloud service provider.

STEP 4: Validate the client token and the response is processed.

STEP 5: Client is given the input file to HDFS

Cluster with cloud service access i.e. cloud service provider associated with Hadoop environment.

STEP 6: First, Split the input file into number of blocks (i.e., each data node).

STEP 7: Each block is created with the three replication file copies of given input file as same file name is copied to different nodes.

STEP 8: Generatekey for each block susin grandomkey generator.

STEP 9: Read data from file and encrypt the data with the key, which is generated by key generator.

STEP 10: Append the key to the block files and load that encrypted data to HDFS.

STEP 11: Extract key from data and pass decrypted data to map reduce work processed by client.

STEP 12: Map reduce programming can be implemented for matching pairs and receiving perfect output. The process is explained in Fig 4.

STEP 13: Combine the output from all working n

odes and send it to client/user.

Corruption data is not possible in this technique because each block of file will encrypt with this key. When you retrieve the data from replicas, they will get decrypted and produce the output using the same secret key. It is authenticated by the secret key of the client (i.e.,) private key. Loss and corruption of Data can be reduced by this technique.

6. CONCLUSION

In the cloud environment, data security is the mostessential purpose for storing a huge amount of data. So, it is focused on Hadoop with cloud.Hadoop is a distributed framework and an easy access data with application and provides reliable replicate blocks of data to nodes in cluster. In order to avoid the occurrence of corruption in Data node duringreplication,a secret key is generated in each block for secured data. The loss and corruptionof data can be avoided using the proposed approach. In future, it will be necessary to make data security more flexible with efficient speed.

7. REFERENCES

[1] Chandra Mouliswaran and ShyamSathyan, "Study on Replica Management and High Availability inHadoop Distributed File System (Hdfs)" Journal of Science, Vol 2, Issue 2, 2012.

[2] Dai Yuefa, Wu Bo, GuYaqiang, Zhang Quan, Tang Chaojing "Data Security Model for Cloud Computing" Proceedings of the 2009 International Workshop on Information Security and Application (IWISA 2009), Qingdao, China, November 21-22, 2009.

[3] Dharmik H. Patel, Dr. S. N. Gujar "A Survey on Data Security System for Cloud Using Hadoop" International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 11, November 2016.

[4] Ms. GurpreetKaur, Ms. Manpreet Kaur, "Review Paper On Big Data Using Hadoop" International Journal of Computer Engineering & Technology (IJCET) Volume 6, Issue 12, Dec 2015.

[5] Harin C Naik, Divyesh Joshi "A Hadoop

Framework Require to Process Bigdata very easily and efficiently, 2016 IJSRSET | Volume 2 | Issue 2

[6] KarthikD,Manjunath T N,Srinivas K "A View on Data Security System for Cloud on Hadoop Framework", International Journal of Computer Applications (0975 - 8887), National Conference on Knowledge, Innovation in Technology and Engineering (NCKITE), 2015.

[7] Iqbaldeep Kaur, Navneet Kaur, Amandeep Ummat, Ja"spreet Kaur, Navjot Kaur "Research Paper on Big Data and Hadoop" IJCST Vol. 7, ISSue 4, oCT - DeC 2016

[8] Monjur Ahmed and Mohammad Ashraf Hossain "Cloud Computing and Security Issues In theCloud " International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.1, January 2014

[9] Murali Krishnan Ramane, Sharmila Krishnamoorthy and Sasikala Gowtham "An Experimental Evaluation of Performance of A Hadoop Cluster on Replica Management" International Journal on Computational Sciences & Applications (IJCSA) Vol.4, No.5, October 2014

[10] NavjotSekhon, Richa Mahajan "Data Security in Cloud Computing Using HDFS" International journal of Computer science Trends and Technology (IJCST) - Volume 5 Issue 2, March - April 2017.

[11] Pedro CaldeiraNeves A, B, Jorge Bernardino A, C "Big Data in the Cloud: A Survey ", Open Journal of Big Data (OJBD) Volume 1, Issue 2, 2015

[12] Miss. Pooja.D.Bardiya1, Miss.Rutuja. A.Gulhane2, Dr.Prof. P.P.Karde3 "Data Security using Hadoop on Cloud Computing IJCSMC, Vol. 3, Issue. 4, April 2014

[13] Rabi Prasad Padhy, Manas Ranjan Patra Suresh Chandra Satapathy "Cloud Computing: Security Issues and Research Challenges" IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS) Vol. 1, No. 2, December 2011

[14] Sharifnawaj Y.Inamdar, Ajit H. Jadhav, Rohit B. Desai, Pravin S. Shinde, Indrajeet M. Ghadage, Amit A. Gaikwad. "Data Security in Hadoop Distributed File System" Volume: 03 Issue: 04, Apr-2016

[15] Xuebin Chen1, Shi Wang1, Yanyan Dong1 and XuWang2"Big Data Storage Architecture Design in Cloud Computing "First National Conference, BDTA 2015, Harbin, China. December 2015.

[16] VijyendraKarpatne, E.J.Thomson Fredrik, "A Secured Data Transmission in cloud using Code Verification and Authentication Techniques", International Journal of Control Theory and Applications, Vol.10,2017.

[17] M Cimi Thomas, S.Sheeja, "Elliptic Curve Cryptography and its Application in the Secure Socket Layer/Transport Layer Security Protocol", International Journal of Control Theory and Applications, Vol.10, No. (29), pp.251-257, 2017.