iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** V **Month of publication:** May 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Secure Data Sharing in Big Data

J. V. N. S. Prasanth[1], Pattabiraman.V[2]
[1]Specialization in Big Data Analytics, [2]Associate Professor, VIT University, Chennai.

Abstract: ***The Storage of data is gradually increased in a huge amount. The data may contains some important or sensitive information store on a big data platform. The main use of this sharing is to share the data between the users in the secured way. which helps in reducing the cost for the enterprises and also helps in providing the value added services to the users. This paper proposes a framework for sharing the data in the secured way including storage ,usage, data delivery and destruction. We are using the ciphertext key technique in order to provide the data sharing process in secure and also safely. This framework provides security to the users information and share the data in a secure way. The owners will have the complete control on the data they have shared and also they will maintain the data that a particular user have been accessed. In this way this will protects the data securely and share the data safely.***
Keywords: Encryption, Decryption, Ciphertext, Partitioning.

I. INTRODUCTION

Generally Big data deals with the large amounts of data sets. Big data process the data to traditional data processing techniques, visualizing the data, data mining techniques, and machine learning algorithms and etc. With the huge development of the technology the rate of data that has been rapidly increased in an massive portion. The data has been of generated quickly everywhere and this data is of various forms like structured, semi structured, and unstructured data. By the process of collecting, sorting, analysing and mining these data an organization can obtain large amount of individual users that data can be used for various business if we can store the data on a big data platform. With the help of the big data as a platform. we can provide the data massively and also we can reduce the cost. Secure data sharing involves following factors. First There are security issues when the data is transferred from the user local space to big data platform .Second there may be some data computing and storage security problems on the big data platform. Third there are issues involving data destruction. With the help of the map reduce techniques the data retrieving process from the big data platform makes easy for the user in order to fetch the data accurately and also make the access process very faster. Existing technologies have partially solved the data sharing and privacy protection issues in various ways, but they have not consider the entire process in the full data security life cycle. In this paper it analyse security issues involving the data sharing life cycle and describe a system model created to ensure secure data sharing on the big data platform. With the help of ciphertext key encryption and decryption technique the data will be shared in a secured way on big data platform.
The Big data mainly deals with namely 5 V's they are:

A. Volume

The volume refers to the amount of the data produced. Just think about the data produced from emails, chat box, video clips and sensor data. The data is produced from last 10 years is 90%. The data is multiplying day by day. So the big data is came in to existence to deal with the big data.

B. Variety

Variety refers to different types of the data that are using. They are different types of data are producing now-a- days. There are both structured and unstructured data are present. Mostly 80% of the data is now unstructured data in the world right now. With the help of big data the variety of data is harness different types of data including messages, social media conversations, photos and the sensing data.

C. Velocity

Velocity dense to the speed of the data is produced now-a-days. Imagine how much the data is produced from the social media. The data producing from the social media is increasing rapidly. Big data technology allows to analyse the data that is producing by increasing order.

D. Value

Value determines the most important v for the big data. It is all well good having access to big data but unless we can turn it. So we can easily argue that value is most important feature of big data.

E. Veracity

Veracity refers to trustworthiness of the data. With many forms of the data present in these days there is no quality, accuracy and less controllable. But big data allows us to work with these data the volumes make up for the quality and accuracy.

II. OBJECTIVE

This project has been proposed the Secure Data Sharing in Big Data methodology that gives: information privacy and uprightness; get to control; information sharing without utilizing PC encryption; insider danger security; and forward and in reverse get to control. The Safe Information Partaking in Database system scrambles a record with a solitary encryption key. Two distinctive key shares for each of the clients are created, with the client just getting one share. The ownership of a solitary share of a key enables the approach to counter the insider dangers. The other key share is put away by a trusted outsider and furthermore with the assistance of the Hadoop dividing strategy. it will distinguish the information what the specific client has been getting to as often as possible it execute a working model of the technique and assess its execution.

III. EXISTING SYSTEM

This model is custom fitted to a stage and does not present any intermediate or representative server between the customer and the specialist organization. it relates all the more nearly to works utilizing encryption to ensure information oversaw by Untrusted clients. In such a case, a principle issue to address is that cryptographic systems can't be locally connected to standard DB. Of course, the quantity of exchanges every moment executed by DB is lower than those alluding to unique DB. It moves far from existing models that store only inhabitant information in the database, and spare meta data in the customer machine or split metadata between the database and a trusted intermediary .While considering situations where different customers can get to a similar database simultaneously.

IV. PROPOSED SYSTEM

This project has proposes the Secure Data Sharing in big data methodology that provides: Data privacy and integrity, get to control, information sharing without utilizing system concentrated with re-encryption; insider risk security; and forward and in reverse get to control. This technique encodes a document with a solitary encryption key. Two diverse key shares for each of the clients are created, with the client just getting one share. The ownership of a solitary share of a key enables the system to counter the insider dangers. The other key share is put away by a trusted outsider, which is known as the cryptographic server. it actualize a working model of the technique and assess its execution in view of the time devoured amid different operations. it performs few Hadoop strategies to investigate diverse clients information in the framework.

V. ADVANTAGES IN PROPOSED SYSTEM

To improve good Quality of Service (QoS). Circulating information among various suppliers and taking advantage of secure sharing. Every user having the own master key. Analysing what the user is accessing frequently.

VI. HADOOP

Hadoop is an Apache open source system written in java that permits appropriated handling of extensive datasets crosswise over groups of systems utilizing basic programming models. A Hadoop outline worked application works in a situation that gives disseminated capacity and calculation crosswise over bunches of systems. Hadoop is intended to scale up from single server to a thousands of machines, each covering neighbourhood computation and capacity.

Hadoop is mainly having four modules.

They are:

A. Hadoop Common

These are Java libraries and utilities required by using different Hadoop modules. These libraries offers file system and OS level abstractions and contains the necessary Java documents and scripts required to begin Hadoop.

B. Hadoop Yarn

This is a framework for process scheduling and cluster aid management.

C. Hadoop Distributed File System

A distributed report device that provides high-throughput get access to application information.

D. Hadoop Map Reduce

This is YARN-based totally system for parallel processing of Large Data sets.

VII. MAP REDUCE

Hadoop Map Reduce is a software program framework for without difficulty writing packages which manner huge quantities of facts in-parallel on huge clusters of commodity hardware in a reliable, fault-tolerant way. The term Map Reduce is mainly Depends upon two different tasks in which hadoop program performs.

A. The Map Task

This is the first task, which takes input facts and converts it into a set of facts, wherein character factors are brake down into tuples.

B. The Reduce Task

This task takes the output from a map task as enter and combines the ones information tuples right into a smaller set of tuples. The reduce mission is constantly completed after the map task.

VIII. HADOOP DISTRIBUTED FILE SYSTEM

Hadoop can work without delay with any mountable dispensed record machine together with Local FS, HFTP FS, S3 FS, and others, however the most common document machine utilized by Hadoop is the Hadoop Distributed File System (HDFS). HDFS utilizes an master/slave design where master comprises of a single NameNode that deals with the document framework metadata and at least one slave Data Nodes that store the actual information. A file in a HDFS namespace is part into a several blocks and those blocks are stored in an arrangement of Data Nodes. The NameNode decides the mapping of pieces to the Data Nodes. The Data Nodes deals with read and compose operation with the record framework. They also take care of block creation, deletion and replication based on instruction given by NameNode.

A client/application can present an occupation to the Hadoop (a hadoop work customer) for required process by determining the accompanying things: The area of the information and yield documents in the disseminated record framework. The java classes as container document containing the usage of guide and lessen capacities. The occupation arrangement by setting diverse parameters particular to the employment.

The Hadoop job client then submits the job (jar/executable) and design to the Job Tracker which then accepts the accountability of circulating the product/setup to the slaves, planning assignments and checking them, giving status and analytic data to the job client.

IX. HIVE

The main work of Hive Partition is also same as SQL Partition but the main difference between SQL partition and hive partition is SQL partition is only supported for single column in table but in Hive partition it supported for Multiple columns in a table .In Hive we can apply Hive Partition concept on Managed tables and External tables. If we not crated dynamic partition for hive, Hive also creates an automatic partition scheme when the table is created. A basic query in Hive reads the whole dataset regardless of the possibility that we have where condition channel. This turns into a bottleneck for running Map Reduce occupations over a vast table. We can defeat this issue by executing parcels in Hive. Hive makes it simple to execute parcels by utilizing the programmed segment conspire when the table is made. In Hive's usage of partitioning, information inside a table is part over various allotments. Each segment relates to a specific value(s) of segment column(s) and is put away as a sub-index inside the table's registry on HDFS. At the point when the table is questioned, where applicable, just the required partitions of the table are queried, along these lines decreasing the I/O and time required by the query.

X. INTRODUCTION ON SQOOP

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Hadoop have multiple features of data management. The major components of Hadoop is the ability to transfer data to their distributed file system, HDFS. One of the frameworks capable of this data transfer is Sqoop. Sqoop is a command line tool used for importing and exporting data between Hadoop and specified relational databases. Importing is the process of bringing data into Hadoop, exporting is the process of taking the data from Hadoop and putting it back into the system. Sqoop can manage both of these processes by using the Sqoop Import and Sqoop Export functions.

Fig 1. Sqoop Import.

Similarly Like Hadoop, Sqoop will also written in Java, Which provides an Application programming interface known as Java Database Connectivity (JDBC). This provides applications in order to provide access data stored in a RDBMS and inspect the nature of the data. If JDBC is native to a database platform, Sqoop can work directly with it. If not, however, it is possible to use Sqoop connectors to gain access to non-compliant external systems. Sqoop relies on the database to describe the schema of the data to be imported, and uses Map Reduce to import and export the data,which provides parallel operation.

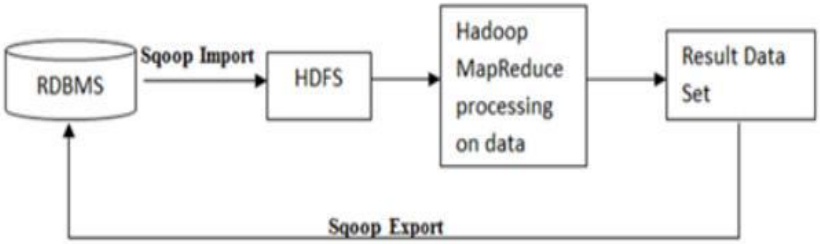


Fig 2. How to use Sqoop.

The data from the relational world will bring into the HDFS by using the Sqoop Import tool and then analyse the data with Map Reduce. with the help of resultant data set, which we can put back into RDBMS using the export process of Sqoop.

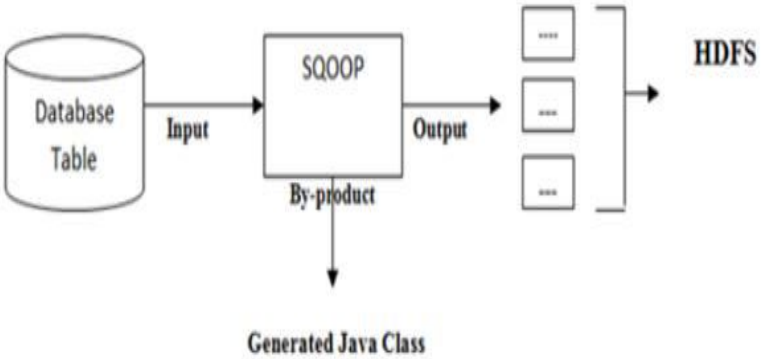


Fig 3. Importing Process.

The input to the import process is a database table. Sqoop initially reads this table row by row into HDFS with the output of the import process being a set of files on the Hadoop distributed file system. It have multiple files in the output and the import process is performed in parallel.

XI. JAVA FRAMEWORK

Java is a programming language initially created by James Gosling at Sun Microsystems and discharged in 1995 as a centre part of Sun Microsystems Java stage. The dialect determines quite a bit of its language structure from C and C++ yet has a more straightforward protest show and less low-level offices. Java applications are ordinarily gathered to bytecode that can keep running on any Java Virtual Machine (JVM) paying little heed to PC engineering. Java is broadly useful, simultaneous, class-based, and protest arranged, and is exceptionally intended to have as few execution conditions as could reasonably be expected. It is planned to give application designers ”a chance to compose once, run any place”.Java is considered by numerous as a stand out among the most compelling programming dialects of the twentieth century, and is broadly utilized from application programming to web applications The java structure is another stage free that rearranges application advancement Internet.

XII. SERVLETS

Prior in customer server figuring, every application had its own customer program and it filled in as a UI and should be introduced on every client’s systems. Most web applications utilize HTML/XHTML that are for the most part upheld by every one of the programs and website pages are shown to the customer as static archives. A website page can simply shows static substance and it likewise gives the client a chance to explore through the substance, however a web application gives a more intelligent experience.

XIII. BENEFITS OF JSP

One of the principle reasons why the Java Server Pages innovation has advanced into what it is today and it is still developing is the mind-boggling specialized need to improve application configuration by isolating element content from static format show information. Another advantage of using JSP is that it permits to all the more neatly separate the parts of web application/HTML creator from a product designer. The JSP innovation is honoured with various energizing advantages, which are chronicled as takes after: 1. The JSP innovation is stage autonomous, in its dynamic website pages, its web servers, and its fundamental server segments. That is, JSP pages perform splendidly with no bother on any stage, keep running on any web server, and web-empowered application server. The JSP pages can be gotten to from any web server. 2. The JSP innovation stresses the utilization of reusable parts. These parts can be consolidated or controlled towards growing more intentional segments and page plan. This unquestionably diminishes improvement time separated from the At advancement time, JSPs are altogether different from Servlets, nonetheless, they are pre compiled into Servlets at run time and executed by a JSP motor which is introduced on an Internet empowered application server, for example, BEA Web Rationale and IBM WebSphere.

`

XIV. JAVA SERVLETS

Java Servlet is a non-specific server expansion that implies a java class can be stacked progressively to extend the usefulness of a server. Servlets are utilized with web servers and keep running inside a Java Virtual Machine (JVM) on the server so these are protected and versatile. Dissimilar to applets they don’t require bolster for java in the web program. Dissimilar to CGI, servlets don’t utilize various procedures to deal with isolated demand. Web server cooperates with the customer through a web program. It conveys the website pages to the customer and to an application by utilizing the web program and he HTTP conventions separately. The characterize the web server as the bundle of substantial number of projects introduced on a PC associated with Web or Intranet for downloading the asked for documents utilizing Record Exchange Convention, serving email and building and distributing pages. A web server chips away at a customer server demonstrate.

XV. DATA ACCESS CONTROL ON STORAGE

In the process of data access control with the help of the ciphertext key encryption access control technique the process of key distribution process will be managed. However, when the access control strategy changes dynamically a data owner have the authority of key management. With the help of the owner key we can only able to access the data in order to get control on the data, By the help of multiple key encryption process during the data access the data transformation process can be done securely user can perform any task on the data available on the big data platform. More specifically,retrieval and comparison of the encrypted data produce correct results, but the data are not decrypted throughout the entire process. Each user’s private key is labelled with a set of key, and the data is encrypted with an attribute condition restricting the user to be able to decrypt the data only if the key is matches with the admin credentials. Distributed system with the information flow control with this we will able to track the flow of data with help of simple tracking rules. By modified CP-ABE algorithm is used to establish fine grained access control in which users are

revoked according to the theory of secret sharing.

XVI. TRUST AND SHARING PROCESS MANAGEMENT

Trust and sharing process management is the process of determining the trusting of the previous unknown user in context of his reliability in performing some action. for this the main research idea includes first to building a trusted platform based on few secure credentials, and then establishing the trust between platform through remote attention. Then, trust is extended to the network. It ensure s that the user's text data exist only in a private operating space of the virtual machines. Data in the memory and the user's key at a user specified time.

XVII. FRAMEWORK FOR DATA SHARING

Issuing and renting data on a semi trusted platform requires a data security mechanism. Developing secure channel for a full sensitive data life cycle is mainly depends upon the four aspects of safety problems: reliable submission,safe storing,riskless use, and secure destruction. A common and popular method of secure data data sharing platform is to providing the data submission with the help of multiple and secure data keys during the time of data submitting data on the platform. The complete data will be stored on the database the data will be retrieved from the database on the Hdfs with the help of the Sqoop tool. The data will be collected on the platform here the data that are accessed by the various number of users will be collected with the help of the partitioner technique. We should not trust the platform without proper information which may lead to loss of data.

XVIII. DABE SCHEME

The DABE scheme includes seven algorithms: Setup, Create User, Create Authority, Request AttributePK, Request AttributeSK, Encrypt and Decrypt. The description of these algorithms is as follows:

SETUP: The Setup algorithm takes as enter the safety parameter. And generate the public key PK which is utilized in all next algorithms and the secret master key MK.

CREATE USER: The CreateUser algorithm takes as input the public key PK, the master key MK, and a user name 'u'. It outputs a public person key PKu with a purpose to be utilized by attribute authorities to problem secret attribute keys for 'u', and a secret user key SKu, used for the decryption of ciphertexts.

CREATE AUTHORITY: The Create Authority algorithm is achieved through the attribute authority with identifier a as soon as at some stage in initialization. And produce the secret user key Ska.

REQUEST ATTRIBUTE PK: The Request Attribute PK set of rules is completed by using characteristic authorities on every occasion they obtain a request for a public attribute key. The set of rules assessments whether or not the authority is answerable for the attribute A. If this is the case, the set of rules outputs a public attribute key PKa, otherwise NULL.

REQUEST ATTRIBUTE SK: The Request Attribute SK set of rules is done with the aid of the attribute authority with identifier a while ever it receives a request for a secret characteristic key. The set of rules tests whether or not it has authority over characteristic A and whether the user u with public key PKu is eligible for A. If this is the case, Request Attribute SK outputs a Secret user key SKa, for user u. Otherwise, the algorithm outputs NULL.

ENCRYPT: The Encrypt algorithm takes as input the public key PK, a message M, an get entry to coverage A and the public keys PKA1 .....PKAN corresponding to all attributes occurring in the policy A. The set of rules encrypts M with A and outputs the ciphertext CT.

DECRYPT: The Decrypt set of rules receives as input a ciphertext CT produced by way of the Encrypt algorithm, an get admission to coverage A underneath which CT was encrypted, and a key ring SKu; SKA1,u..... SKAN,u for user u, which includes the name of the game user key Sku and mystery characteristic keys SKA1,u..... SKAN, u for attributes A1.....AN. The set of rules Decrypt decrypts the ciphertext CT and outputs the corresponding plaintext M if the attributes A1.....AN have been sufficient to satisfy A; in any other case it outputs NULL.

XIX. USER INTERFACE DESIGN

Interface design deals with the process of developing a method for modules in a system to connect and communicate. These modules can apply to hardware, software or the interface between a user and a machine.Application Clients need to see the application they have to login through the UI GUI is the media to associate Client and Media Database and login screen where client can enter his/her client name, password and secret key will check in database, if that will be a substantial username and password

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

then he/she can get to the database.

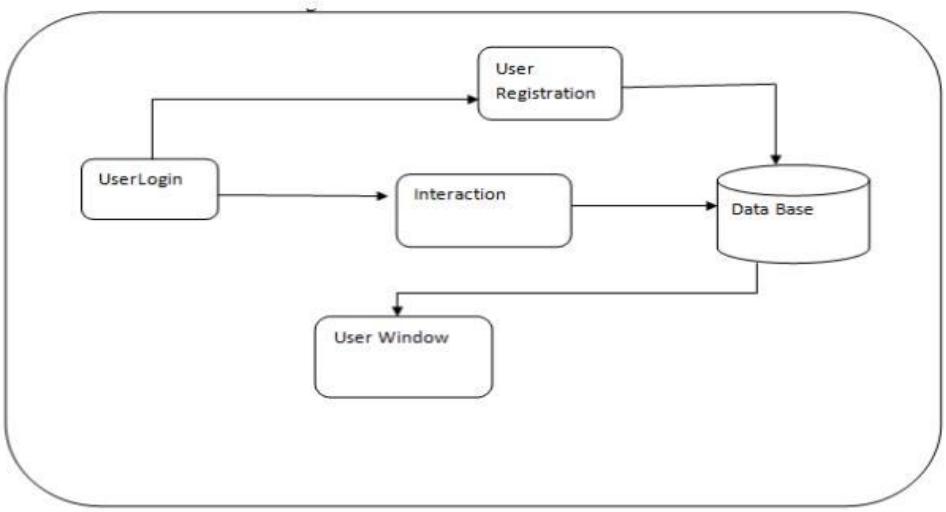


Fig 4.User Interface Design

XX. FILE UPLOAD

This module is used to help the user to upload their files in secured database. Before uploading the files in the cloud the data will be send to server. The uploaded data can be a pdf, text.

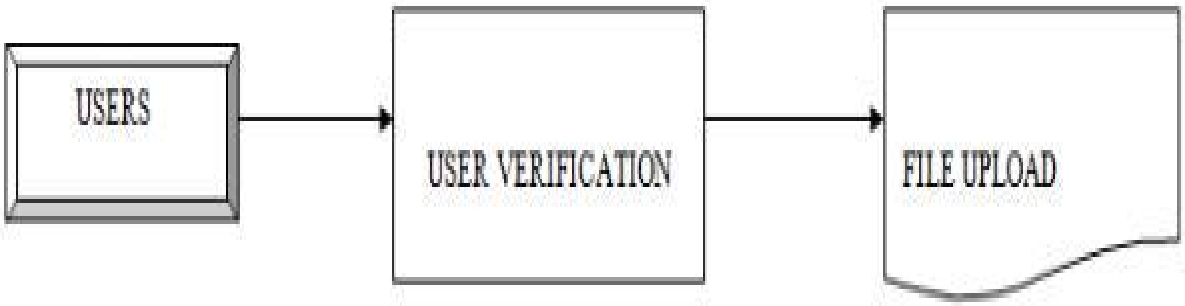


Fig 5.File Upload Design

XXI. CRYPTOGRAPHIC SERVER ACCESS

This module is used when the time of file uploading the file will be send to an cryptographic server. Here the cryptographic server is mentioned as the third party. So the third party provides the security here. Cryptographic Server used for key generation. Cryptography is the art of achieving security by encoding messages to make them non readable. Cryptography not only protects the information but also provides authentication to the user. Here the original information and encrypted information are referred as plaintext and cipher text respectively. The transformation of plaintext into unintelligible data known as cipher text is the process of encryption. Decryption is the reverse process of encryption i.e. conversion of cipher text into plain text. During communication, the sender performs the encryption with the help of a shared secret key and the receiver performs the decryption.

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

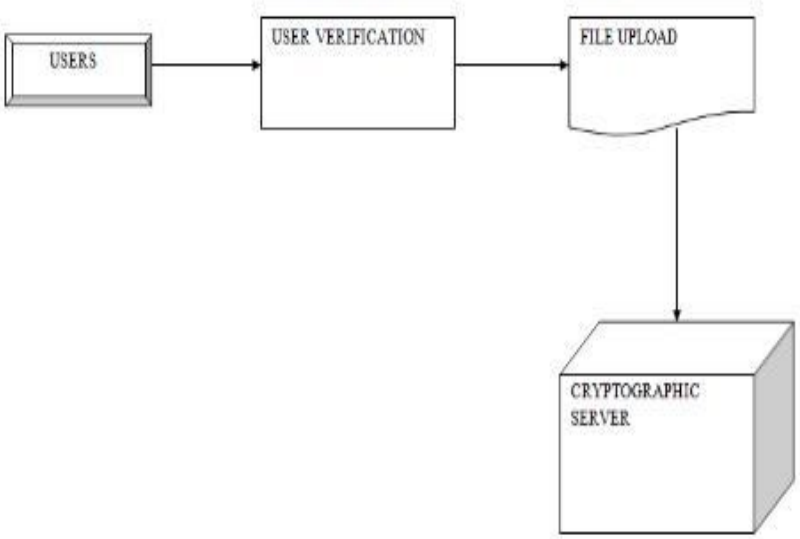


Fig 6. Cryptographic Server Access Design.

XXII. KEY GENERATION

In this module the Server generates two types of different secret key, one for file owner and other for users. Using the users key they can view the files or retrieve the files. But using the file owners key they done a modification like delete, edit etc.

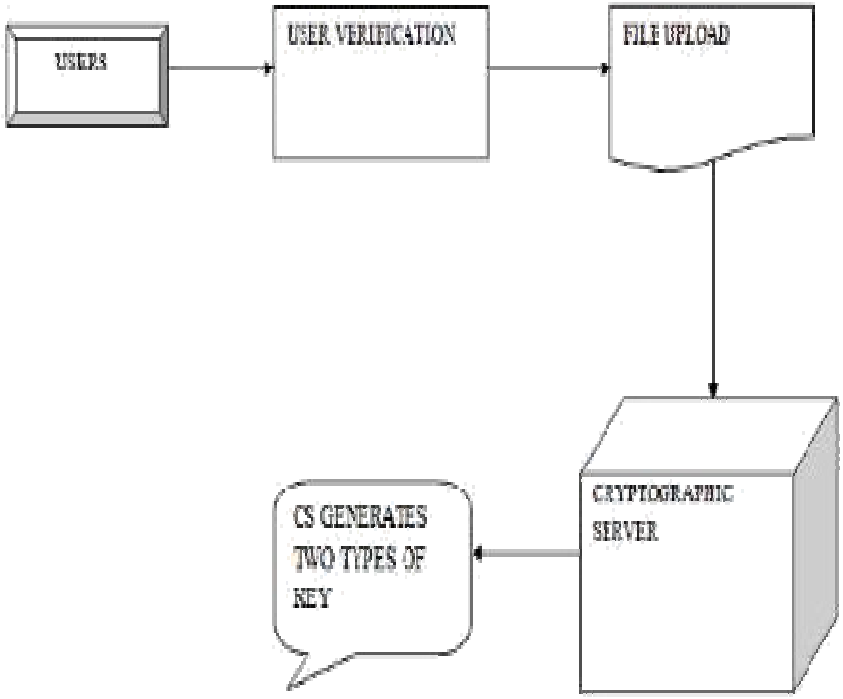


Fig 7. Key Generation Design.

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

XXIII. FILE ENCRYPTION AND DECRYPTION

In this method the uploaded data will be encrypted after the key generation and it will be stored in the database. If any user wants to access a particular data they need to provide authentication key. After verifying the key it will be encrypted. This module is used for security purpose. Here after verifying the required key. The data will be encrypted into readable format. This module is used to retrieving the data from the server. After finishing the key verification it will be decrypted into original data.

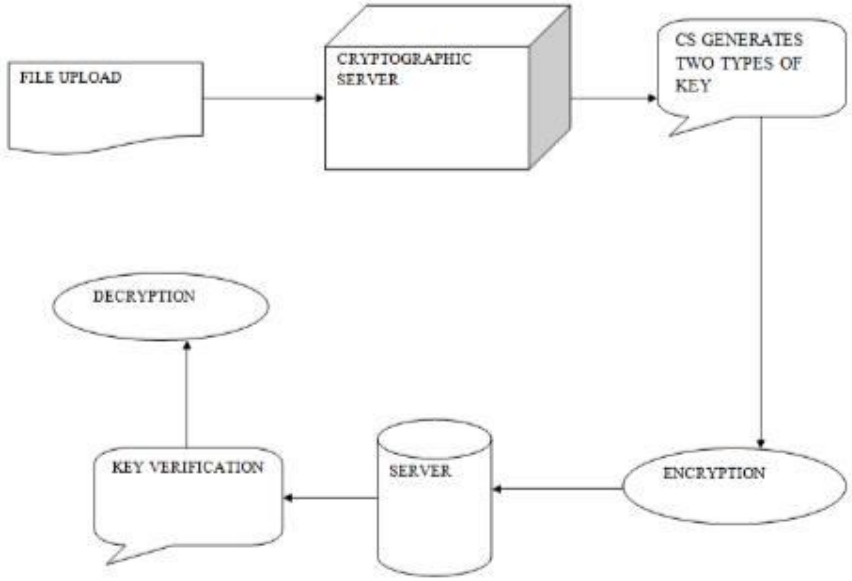


Fig 8. File Encryption and Decryption Design.

XXIV. HOW TO ANALYSE THE DATA USING HIVE

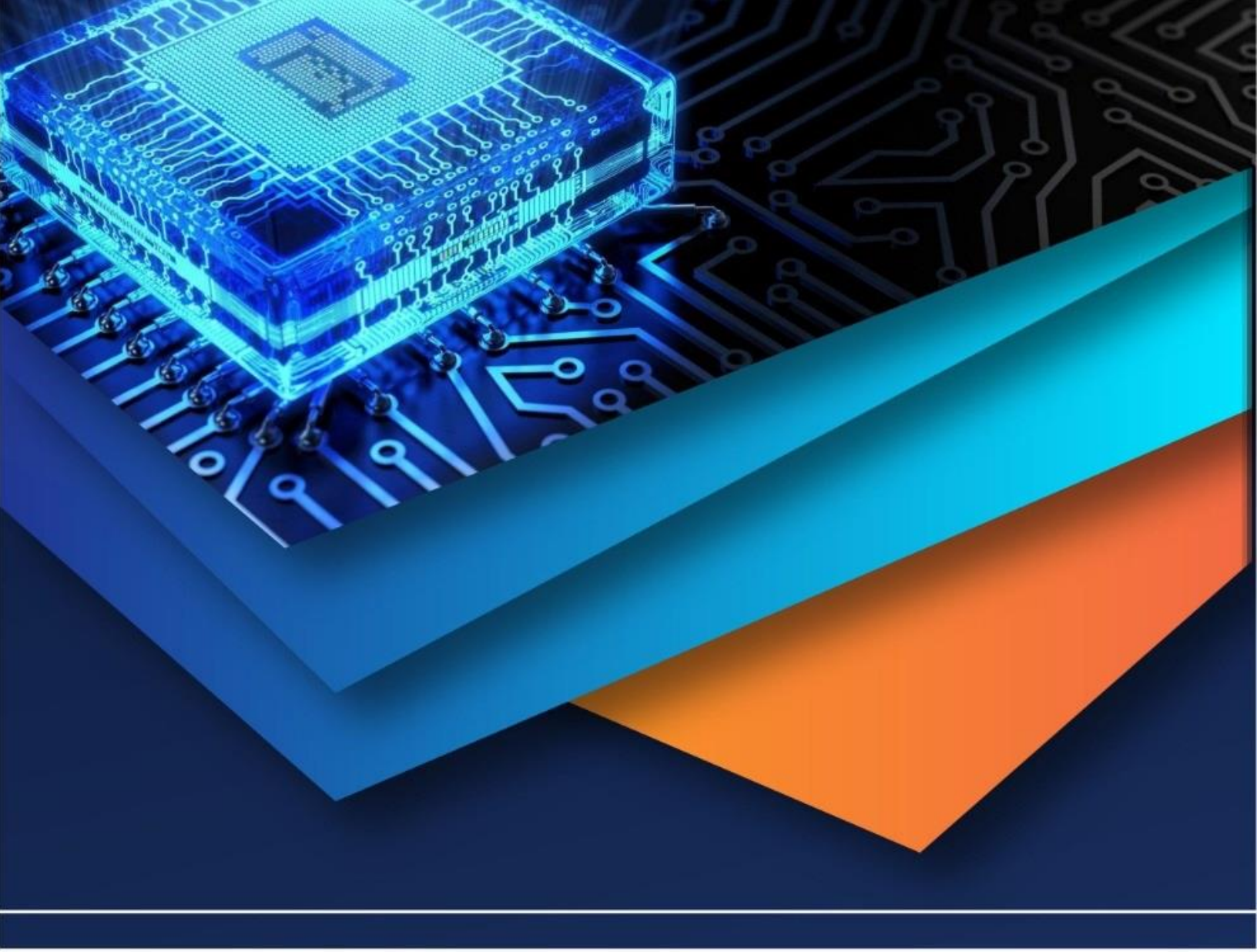
During this methodology initially we have to pass the particular file location of the respective file we have to pass the document we have to pass into the Hive storage database. In this phase Initially we have to Create Database into hive ecosystem and then we have to use that particular database. Here particular data that is present in the local memory have to be passed into the hive database using the load function with respect to local address into the user table. In order to analyse by using the user database present in the hive by selecting the uid, address, phono and age as the key attributes by specifying the age limits we can analyse.

XXV. CONCLUSION

In this work It takes the problem of Providing the data sharing process in the secured way such the data sharing process will be done with high security by using the cypertext key encryption technique which is helpful in the process of key encryption and decryption. With the help of encryption algorithms with respect to the specific decryption algorithm data will be decrypted. By using the developed web application using this web application data will be transferred to Hadoop HDFS using the Sqoop tool and then that particular data will be accessed by using the hive tool. With the help of hive based on the different user references data which is accessed by the specific user will be partition by using the Hadoop partition technique.

REFERENCES

[1] ”Scalable and Secure Sharing of Personal Health Records in Cloud Computing Using Attribute-Based Encryption”,Ming Li,Shucheng Yu,Yao Zheng, Kui Ren and Wenjing Lou,IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS.
[2] ”On Multi-Authority Cipertext-Policy Attribute-Based Encryption”,Sascha Muller,The preliminary version of this paper appeared in the proceedings of the International Conference on Information Security and Cryptology (ICISC) 2008.
[3] ”Distributing Data for Secure Database Services”,Dilys Thomas,In ACM Conference on Computer and Communications Security 2010
[4] ”Ciphertext-Policy Attribute-Based Encryption”,John Bethencourt,2007 IEEE Symposium on Security and Privacy.
[5] ”Personalization vs. Privacy in Big Data Analysis”,Benjamin Habegger,in 2008 IEEE Symposium on Security and Privacy,IEEE, May 2008.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY